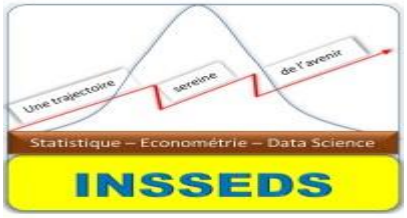
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

REPUBLIQUE DE CÔTE D'IVOIRE

Union – Discipline – Travail



INSTITUT SUPERIEUR DE STATISTIQUE, D'ECONOMETRIE ET DE DATASCIENCE

MASTER 2

STATISTIQUE ECONOMETRIE DATA SCIENCE

MINI PROJET SERIE TEMPORELLE

RAPPORT STATISTIQUE POUR CORPORACION FAVORITA

ANNEE UNIVERSITAIRE : 2024-2025

ETUDIANT :

N'DRI ONESIME

ENSEIGNANT :

AKPOSSO DIDIER MARTIAL

AVANT-PROPOS

L'association du théorique au pratique, des connaissances aux compétences et des savoir-faire aux savoirs est la principale tendance récente dans le secteur technique. Dans ce contexte, l'INSSEDS (Institut Supérieur de la Statistique, d'Econométrie et de la Data Science), dans sa formation en master professionnel en statistique, Économie et science des données, impose que les divers crédits soient validés en effectuant un mini-projet à la fin de chaque module. Le projet est donc structuré et supervisé de cette manière, visant principalement à faire de chaque élève un participant dynamique, Engagé et libre dans la vie active.

Ce document est un rapport de projet Muni axé sur la prévision des ventes. Il se divise principalement en trois parties : La section initiale traite de l'importation et organisation des ventes par mois, la seconde partie est associée à une analyse descriptive des ventes, la troisième partie est basée sur la prévision par la méthode HOT-WINNER.

En règle générale, toutes les analyses et conclusions présentées dans ce rapport relèvent de la responsabilité de l'auteur, qui ne sollicite ni autrui ni l'INSSEDS (Institut Supérieur de Statistique d'Econométrie et de Data Science).

Table des matières

INTRODUCTION GENERALE

Les épiceries traditionnelles doivent jongler délicatement entre les prévisions d'achat et de vente. Si vous prévoyez un peu trop, vous risquez de vous retrouver avec des stocks excédentaires de produits périssables. À l'inverse, si vous prévoyez un peu moins, les articles populaires disparaissent rapidement, laissant les clients frustrés et des ventes manquées. Ce dilemme devient encore plus complexe à mesure que les détaillants ajoutent de nouveaux points de vente avec des besoins uniques, introduisent de nouveaux produits et s'efforcent de suivre les goûts saisonniers en constante évolution et le marketing imprévisible des produits.

Corporación Favorita, un important détaillant en alimentation basé en Équateur, connaît bien ces défis. Exploitant des centaines de supermarchés avec plus de 200 000 produits différents dans ses

rayons, ce détaillant doit continuellement affiner ses prévisions de vente pour rester compétitif et répondre aux attentes de ses clients.

- **Contexte et justification**

Dans le secteur de la vente au détail, une gestion précise des stocks et des prévisions de vente est essentielle pour maintenir une rentabilité et une satisfaction clients optimaux. Les épiceries traditionnelles, comme Corporación Favorita en Équateur, doivent constamment équilibrer l'offre et la demande. Ces entreprises font face à des défis complexes, tels que la gestion des stocks de produits périssables, l'adaptation aux variations saisonnières des goûts des consommateurs et l'impact des promotions. Corporación Favorita, avec ses centaines de supermarchés et ses plus de 200 000 produits, doit affiner ses prévisions de vente pour répondre efficacement aux besoins de ses clients tout en minimisant les coûts. L'objectif de cette étude est de prédire les ventes des milliers de familles de produits vendues dans les magasins de l'épicerie « Favorita » situés en Équateur. Pour ce faire, nous utiliserons des données incluant les dates, les informations sur le magasin et le produit, les promotions en cours, ainsi que les chiffres de vente.

- **Principaux résultats attendus**

Amélioration des précisions de prévision.

Réduction des coûts de stockage :

Augmenter la satisfaction client en assurant la disponibilité constante des produits populaires.

Efficacité des campagnes promotionnelles.

Insights Stratégies sur les Tendances du Marché

- **Problematique**

Comment la corporation Favorita peut-elle optimiser ses prévisions de vente afin de réduire les coûts liés aux excédents de stock et aux ruptures de stock, tout en augmentant la satisfaction de ses clients

grâce à une gestion efficace des stocks, des promotions et des tendances saisonnières ?

- **Methodologie**

  – **Importation des informations et structuration mensuelle des ventes totales.**

  – **Étude détaillée des ventes**

  – **Estimation des 12 mois à venir en utilisant la méthode de Holt-Winter.**

- **Description du jeu de données**

Id : identifiant unique de chaque enregistrement dans le jeu de données.

Date : Date à laquelle les données de vente sont enregistrées.

Magasin_nbr : numéro d'identification unique du magasin où les ventes ont été effectuées.

Famille : Catégorie ou famille de produits vendus (par exemple, « AUTOMOBILE », « PUÉRICULTURE », etc.). Cette variable est également un facteur avec 33 niveaux représentant différentes catégories de produits.

Ventes : Quantité de ventes réalisées pour chaque enregistrement, représentée sous forme de nombre.

En promotion : Indicateur de promotion, prenant la valeur 0 ou 1 pour spécifier si le produit était en promotion lors de la vente (1 pour la promotion, 0 sinon).

La suite de notre analyse sera subdivisée en trois parties à savoir :

- **Importation et organisation des ventes totales par mois**
- **Analyses descriptives des ventes**
- **Prévision des ventes des 12 prochains mois**

1ere PARTIE : IMPORTATION ET ORGANISATION DES VENTES TOTALES PAR MOIS

a) Importation des données, visualisation et description des données

- **Exploration du jeu de données**

```
 'data.frame':    3000888 obs. of  5 variables:
  $ date       : Factor w/ 1684 levels "2013-01-01","2013-01-02",..: 1
1 1 1 1 1 1 1 1 1 ...
  $ store_nbr  : Factor w/ 54 levels "1","2","3","4",..: 1 1 1 1 1 1 1
1 1 1 ...
  $ family     : Factor w/ 33 levels "AUTOMOTIVE","BABY CARE",..: 1 2
3 4 5 6 7 8 9 10 ...
  $ sales      : num  0 0 0 0 0 0 0 0 0 0 ...
  $ onpromotion: Factor w/ 362 levels "0","1","2","3",..: 1 1 1 1 1 1
1 1 1 1 ...
```

A travers le dictionnaire du jeu de données on peut observer que ce jeu données comporte 3000888 observations et 5 variables dont 4 variables catégorielles et 1 variable quantitative qu'est les sales.

- **Visualisation des 5 premiers et deniers observations du jeu de données**

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

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

On peut voir qu'à la date du 2013-01-01 le magasin numéro 1 n'a obtenu aucune vente. Aussi, à la date du 2017-08-15 le magasin 9 a vendu 438 POUTRY, 155 PREPARED FOODS qui était en

promotion, 2420 PRODUCE, 121 SCHOOL AND OFFICE SIPPLIES et 16 SEAFOOD qui n'était pas en promotion.

b) Organisation des ventes par mois

```
A tibble: 56 × 2
   mois       ventes_totales
   <date>              <dbl>
 1 2013-01-01      10327625.
 2 2013-02-01       9658960.
 3 2013-03-01      11428497.
 4 2013-04-01      10993465.
 5 2013-05-01      11597704.
 6 2013-06-01      11689344.
 7 2013-07-01      11257401.
 8 2013-08-01      11737789.
 9 2013-09-01      11792933.
10 2013-10-01      11775620.
ℹ 46 more rows
```

Ce tableau comporte les ventes par mois de 2013 à 2017. Ainsi, après organisation des données des ventes par mois nous pouvons passer à la suite de l'analyse dans lequel nous convertisserons ce tableau en série temporelle et nous ferons une analyse descriptive des ventes.

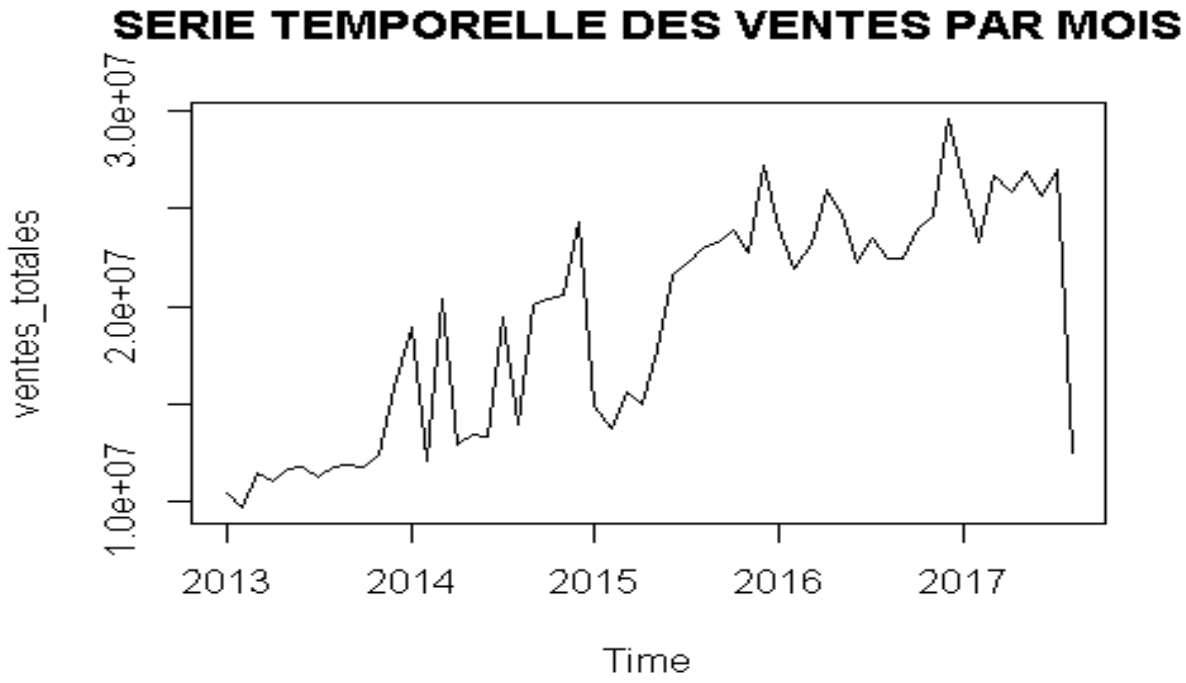
2ème PARTIE : ANALYSE DESCRIPTIVE DES VENTES

a) Conversion du tableau en série temporelle

```
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
2013 10327625  9658960 11428497 10993465 11597704 11689344 11257401 11737789
2014 18911641 12038353 20365584 12861251 13379785 13319958 19421891 13885176
2015 14896922 13742396 15598608 14955068 17730368 21615360 22209619 22963674
2016 23977805 21947409 23131781 25963025 24779432 22209219 23462672 22452414
2017 26328160 23250112 26704018 25895308 26911847 25682822 27011478 12433323
          Sep      Oct      Nov      Dec
2013 11792933 11775620 12356559 15803117
2014 20022416 20396101 20531635 24340454
2015 23240882 23878268 22804953 27243982
2016 22417448 24030390 24642640 29640288
2017
```

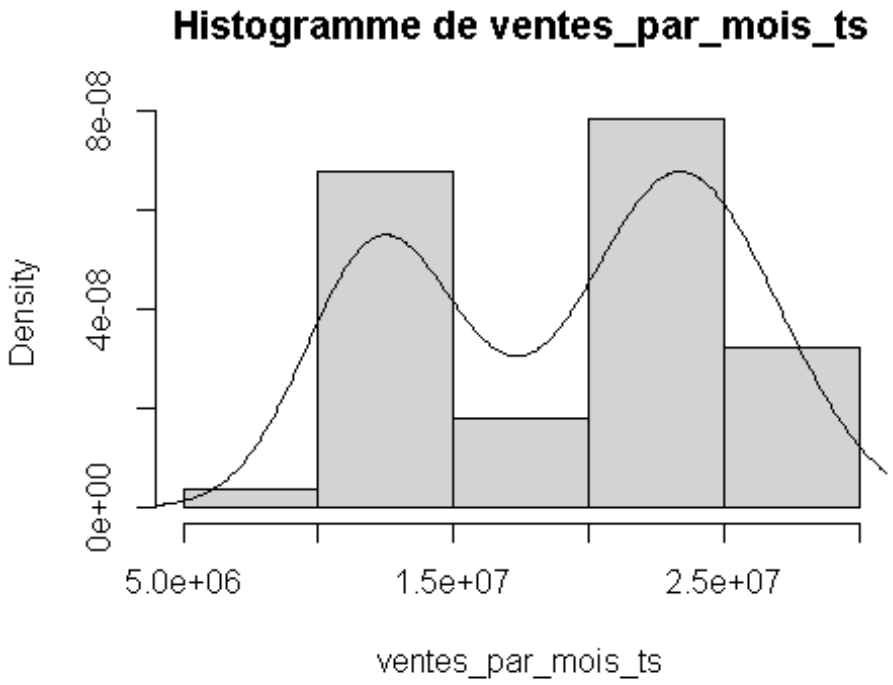
b) Graphique

- **Serie temporelle**



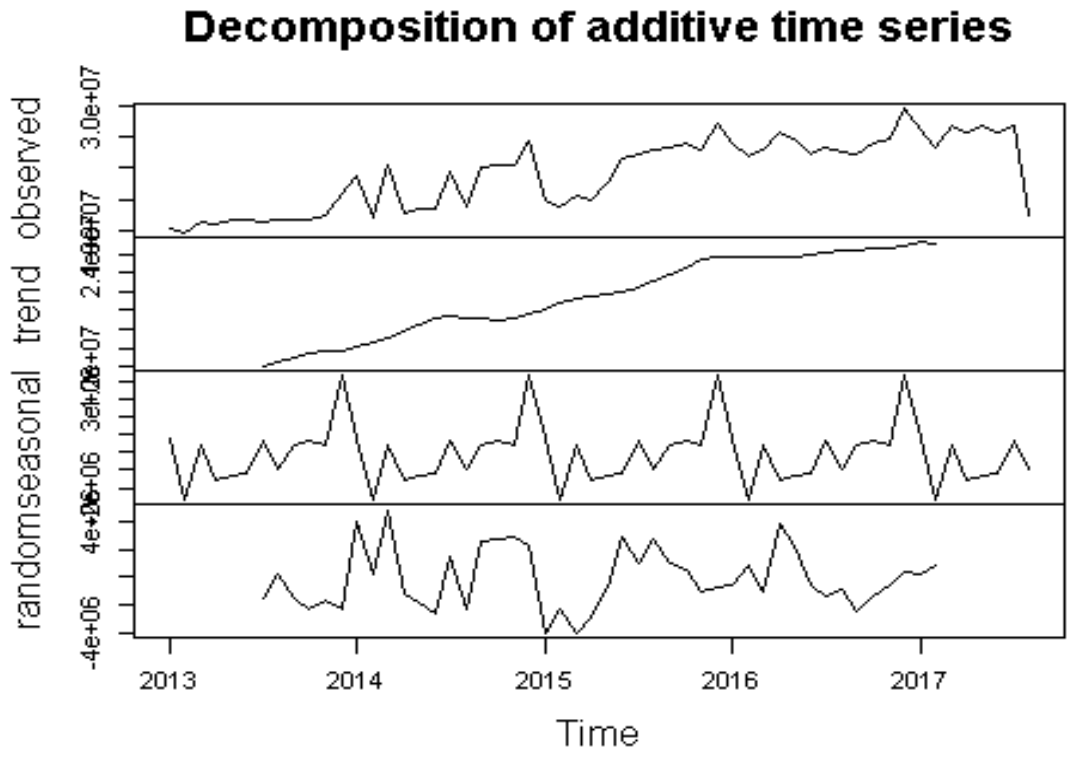
En comparant les ventes de chaque année, il est possible d'identifier les années particulièrement performantes ou moins performantes. Par exemple, si l'on constate une forte augmentation en 2016 par rapport à 2015, cela pourrait indiquer un changement positif dans la stratégie de vente ou des facteurs externes favorables. Notons qu'en joignant les minima et les maximas on peut dire que la série est additive.

- **Histogramme**



Les ventes par mois les plus représentées sont comprises entre 20000000 et 25000000.

c) Tendance et composante saisonnière



Composant observé : il s'agit des valeurs réelles mesurées au fil du temps, qui présentent une certaine variabilité. On peut observer des pics et des creux, indiquant des fluctuations dans les données.

Tendance : la tendance montre la direction générale des données sur la période de 2013 à 2017. On peut voir si les valeurs augmentent, diminuent ou restent stables. Dans ce cas, la tendance semble montrer une augmentation générale, bien que des variations soient présentes.

Composant saisonnier : ce composant représente les variations périodiques qui se répètent à intervalles réguliers. Les fluctuations saisonnières peuvent être dues à des facteurs saisonniers, comme des changements climatiques ou des événements spécifiques à certaines périodes de l'année.

Composant aléatoire : ce composant capture les variations imprévisibles qui ne peuvent pas être expliquées par les autres composants. Il représente le bruit dans les données et peut être influencé par des événements exceptionnels ou des erreurs de mesure.

d) Indice statistique

- **Indice de tendance centrale**

MINIMUM : 9658960 Ce qui signifie que la plus petite vente par mois des magasins de l'épicerie "Favorita" est de 9658960.

MAXIMUM : 29640288 Ce qui signifie que la plus grande vente par mois des magasins de l'épicerie "Favorita" est de 29640288.

MOYENNE : 19172231 En moyenne les ventes par mois des magasins de l'épicerie "Favorita" est de 19172231.

25% (13205281) : 25 % des épiceries « Favorita » génèrent moins de 1 320 528 ventes mensuelles.

50% (20463868) : 50% des épiceries « Favorita » génèrent moins de 20463868 ventes mensuelles.

75% (23903152) : 75% des épiceries « Favorita » génèrent moins de 23903152 ventes mensuelles.

- **Indice de dispersion**

VARIANCE: 3.372074e+13

ECART-TYPE : 5806956 La plupart des ventes par mois des épiceries « Favorita » est compris entre 13365275 et 24979187.

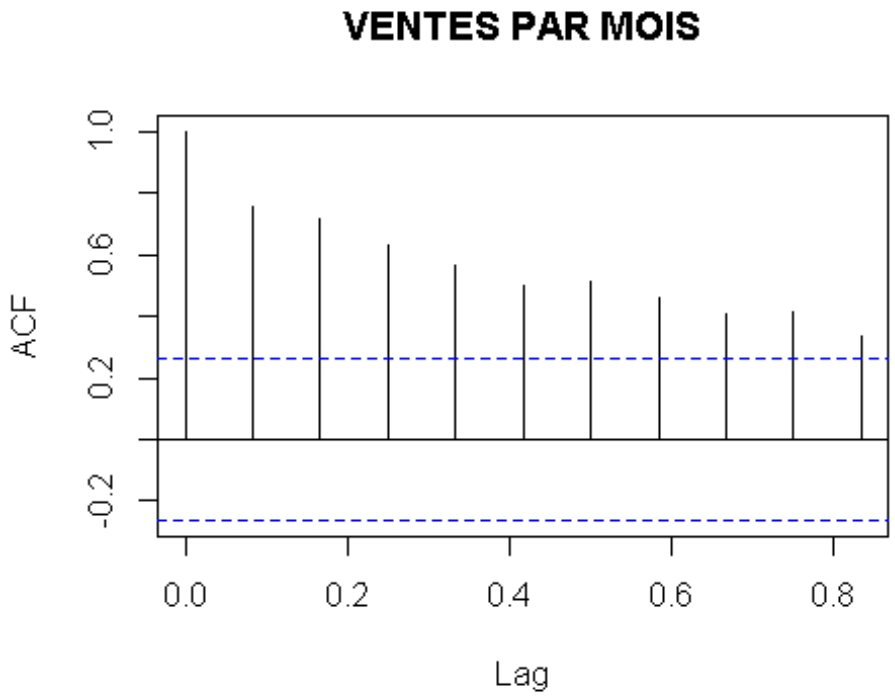
- **Indice de forme**

Skewness : -0.1546965 distribution étalée à gauche. Cela peut indiquer que la majorité des valeurs sont concentrées à droite de la moyenne et que quelques valeurs plus petites tirent la moyenne vers la gauche.

Kurtosis : 1.573615 distribution platikurtique. Une kurtosis de 1.573615 indique que ta distribution de ventes par mois est platikurtique, ce qui signifie qu'elle est plus aplatie que la courbe normale et présente moins de valeurs extrêmes.

- **Indice de dependance**
- **autocorrelation simple**

0.0000	0.0833	0.1667	0.2500	0.3333	0.4167	0.5000	0.5833	0.6667	0.7500	0.8333
1.000	0.759	0.714	0.634	0.566	0.503	0.514	0.460	0.409	0.413	0.337



Le graphique ainsi obtenu est un corrélogramme. On peut constater une forte autocorrélation

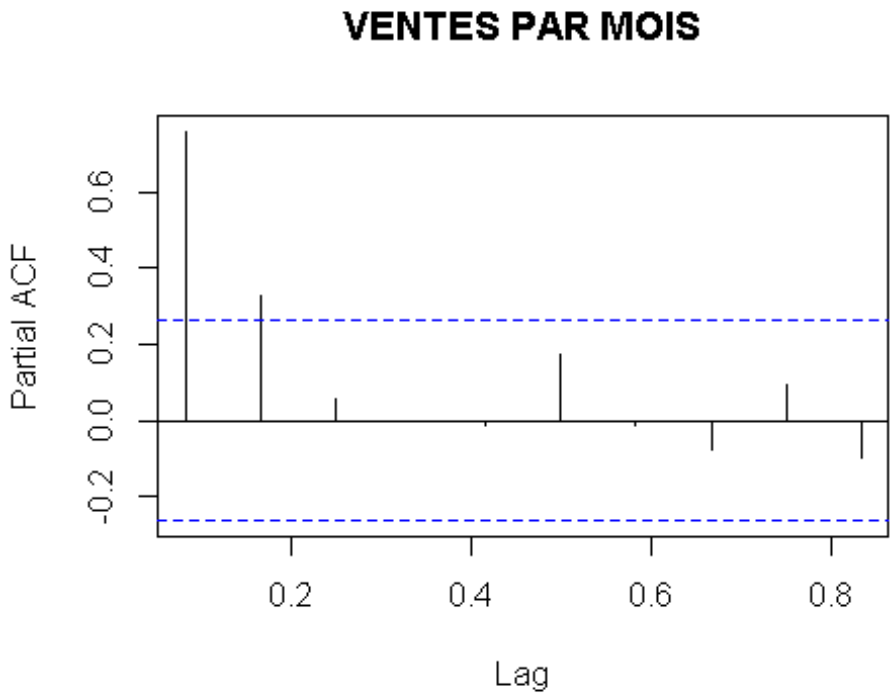
- d'ordre 1 (0.759)

- d'ordre 2 (0.714)

- d'ordre 3 (0.634)

– **Autocorrélation partielle**

L'autocorrélation partielle (PACF) permet de quantifier la dépendance linéaire entre deux réalisations successives mais conditionnellement aux réalisations intermédiaires.

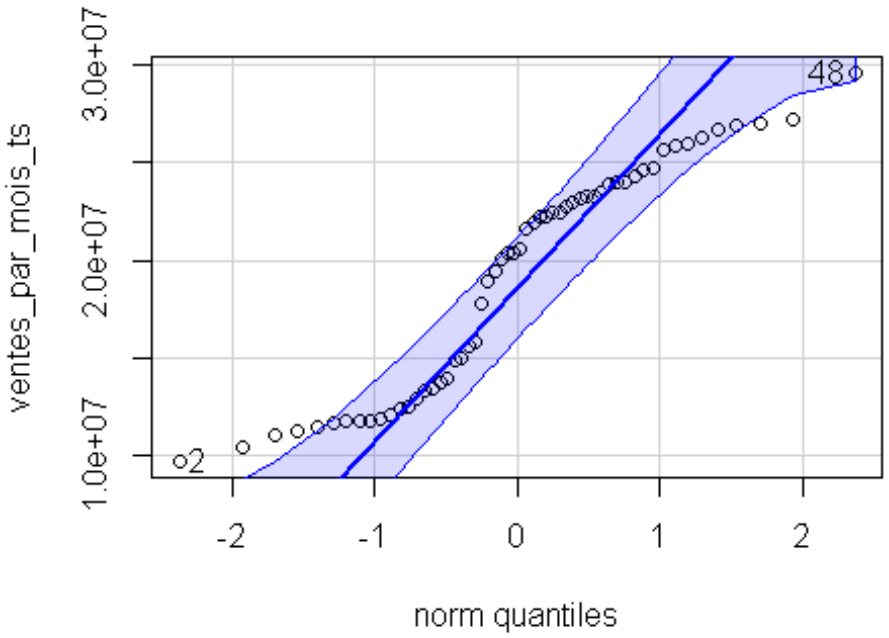



```
0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333
0.759  0.326  0.057  -0.003 -0.012 0.172 -0.012  -0.079 0.092 -0.098
```

Les autocorrélations observées aux décalages 0.1667 et 0.2500 étaient un effet résiduel de l'autocorrélation pour un décalage de 0.0833.

e) <u>Test de normalité</u>

- **Graphique**



- **<u>Test</u>**

H0 : la distribution suit une loi normale

H1 : la distribution ne suit pas une loi normale

P-value = 0.0005107 < 0,05, on rejette *H*0 : Donc la distribution ne suit pas une loi normale.

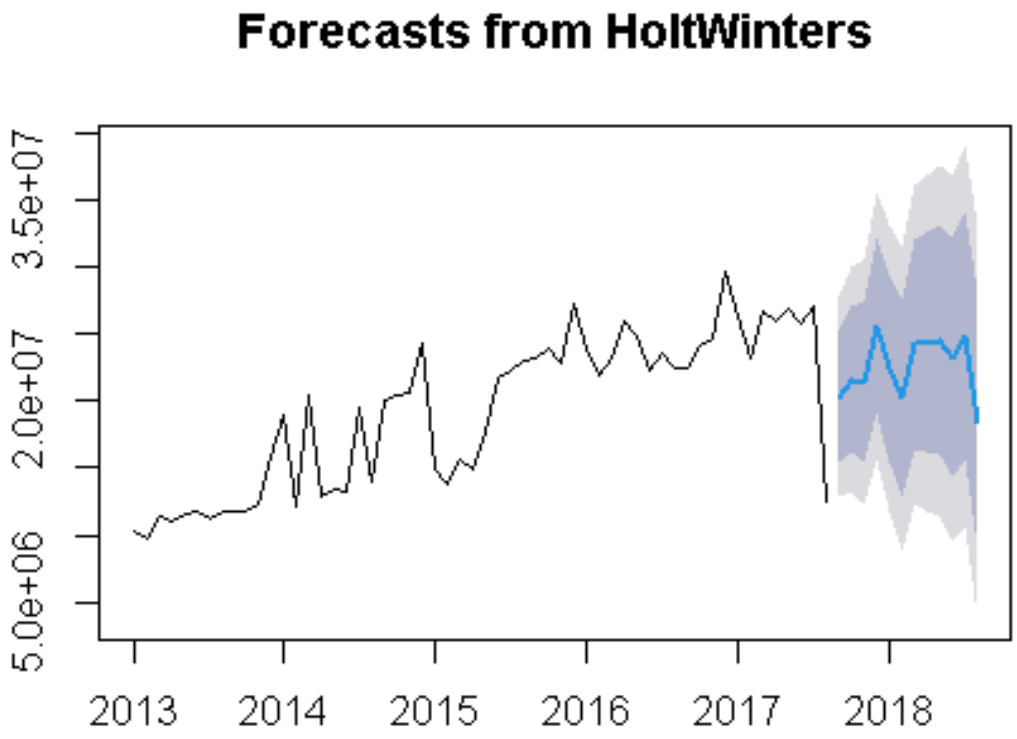
3ème PARTIE : PREVISION DES VENTES DES 12 PROCHAINS MOIS

- **<u>Méthode Hot-winter</u>**

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Sep 2017	20320486	15454257	25186715	12878231	27762741
Oct 2017	21567336	16100814	27033858	13207012	29927660
Nov 2017	21419414	15412290	27426538	12232311	30606518

Dec 2017	25653572	19150633	32156511	15708185	35598959
Jan 2018	22443581	15480040	29407121	11793765	33093397
Feb 2018	20209519	12814008	27605029	8899062	31519975
Mar 2018	24212750	16409145	32016354	12278166	36147333
Apr 2018	24323908	16132514	32515301	11796252	36851563
May 2018	24522327	15960692	33083963	11428436	37616219
Jun 2018	23227290	14310773	32143808	9590654	36863927
Jul 2018	24875385	15617580	34133190	10716794	39033976
Aug 2018	18335847	8748896	27922798	3673870	32997824

Le point forecast pour septembre 2017 est de 20320486 et que l'intervalle de confiance à 95 % est compris entre 12878231 et 27762741, cela signifie que la vraie valeur à 95 % de chances de se situer entre 12878231 et 27762741 unités, avec 20320486 comme estimation centrale.

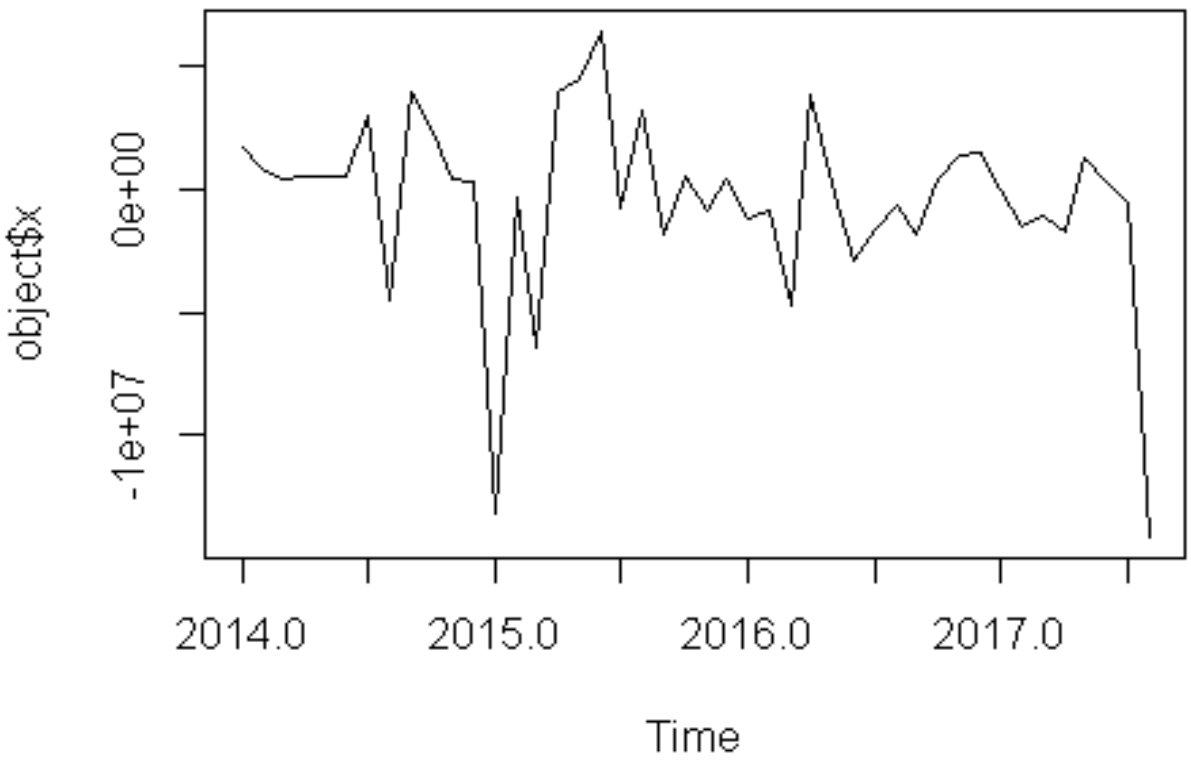


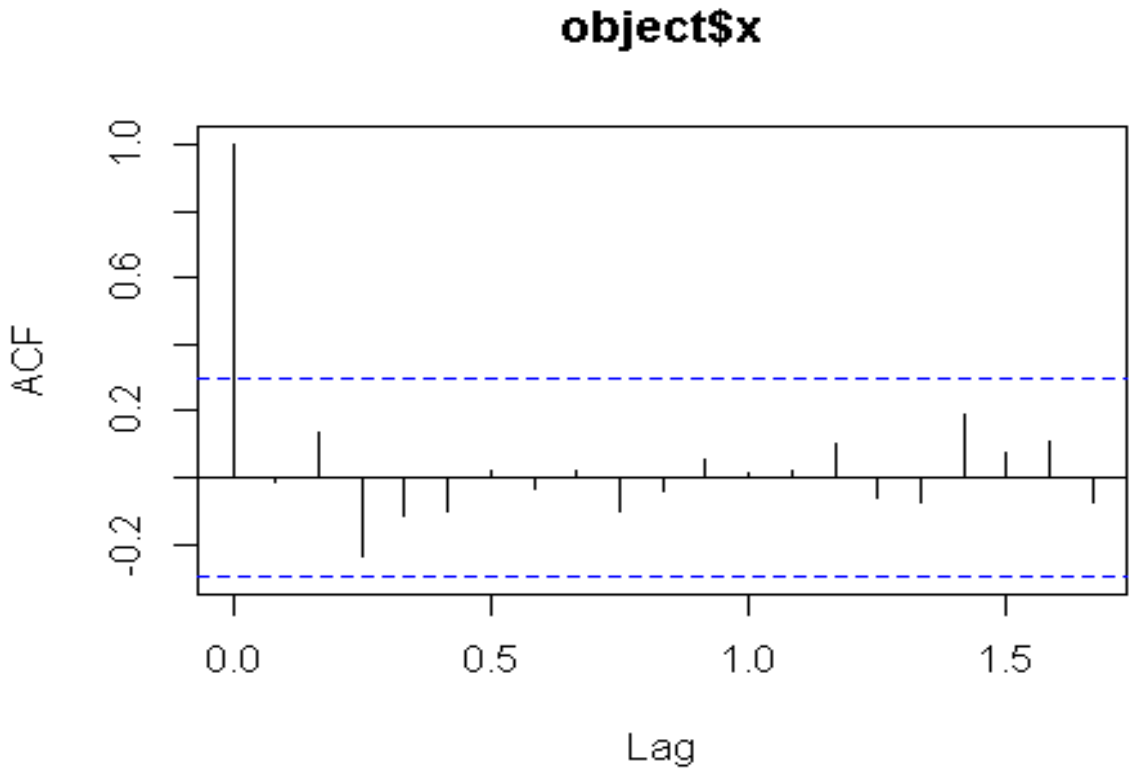
- **<u>Validation du modèle de prévision</u>**

– **Recuperation des résidus**

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

- **Graphique des résidus**





- **TEST**
  - **Box-Ljung test**

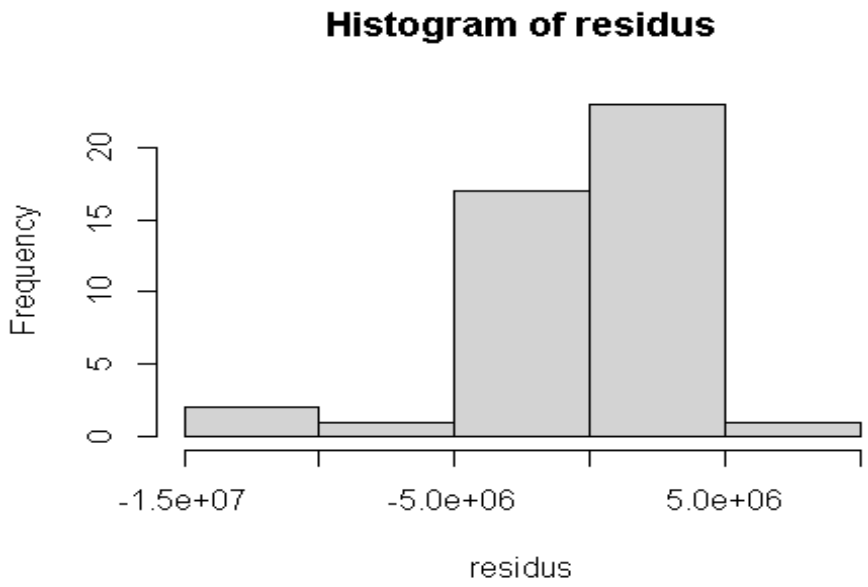
H0 : la série est un bruit blanc

H1 : la série n'est pas un bruit blanc

Conclusion : p-value= 0.9199 > 0.05, on ne peut rejeter H0, la série est un bruit blanc. Ici, la statistique de test de Ljung-Box est de 11.888, et la valeur p est de 0.9199. Pour être sûr que le modèle prédictif ne peut pas être amélioré, il est également judicieux de vérifier si les erreurs de prévision sont normalement réparties de moyenne zéro et de variance constante.

- **Shapiro-Wilk normalité test**

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



On peut aussi faire un test de Shapiro Wilk

H0 : la série suit une loi normale

H1 : la série ne suit pas une loi normale

Conclusion : la p-value < 0.05, on rejette H0, la série ne suit pas une loi normale

- **Moyenne des résidus**

La série de ventes mensuelles présentant une moyenne des résidus de -424 918,6 indique que le modèle néglige constamment les ventes effectives. Cela pourrait donc influencer considérablement la planification et les choix fondés sur ces prévisions.

CONCLUSION GENERALE

Pour conclure, l'emploi de méthodes de prévision pour anticiper les ventes de chaque catégorie d'articles dans chaque boutique de la marque « Favorita » en Équateur présenterait un potentiel considérable pour optimiser la gestion des stocks et la planification des opérations. En créant des schémas exacts de prévision des ventes, les commerçants peuvent anticiper plus efficacement la demande, prévenir les déficits de marchandises, diminuer les dépenses associées au stockage excessif et optimiser les capacités d'inventaire pour augmenter le chiffre d'affaires tout en réduisant les pertes. Ces estimations peuvent également aider à prendre des décisions liées aux prix, aux offres promotionnelles et aux stratégies marketing. Elles fournissent en effet des informations détaillées sur la manière d'augmenter les ventes et de fidéliser les clients. En fin de compte, l'utilisation de ce modèle de prévision des ventes contribue à améliorer l'efficacité opérationnelle et la rentabilité de la chaîne de supermarchés « Favorita », renforçant ainsi sa position dans le domaine.

ANNEXES

SOURCE DE CODE

- **Exploration du jeu de données**

```
knitr::opts_chunk$set(echo = TRUE)

store <-
read.csv("C:/Users/HP/Downloads/INSSEDS/MINIPROJETINSSEDS_serie_temp/store.csv", stringsAsFactors=TRUE,
row.names = 1)
library(dplyr)
store$store_nbr = as.factor(store$store_nbr)
store$onpromotion = as.factor(store$onpromotion)
str(store)
```

- **Visualisation des 5 premiers et dernières observations du jeu de données**

```
head (store, 5)
```

```
tail (store, 5)
```

- **Organisation des ventes par mois**

```
library(dplyr)
 library(lubridate)
 store <- store %>%
   mutate (date = as.Date(date, format = "%Y-%m-%d"))
 ventes_par_mois <- store %>%
   mutate(mois = floor_date(date, "month")) %>%
   group_by(mois) %>%
   summarise(ventes_totales = sum(sales, na.rm = TRUE))
 print(ventes_par_mois)
```

- **Conversion du tableau en série temporelle**

```
library(dplyr)
 library(lubridate)
```

```
 store <- store %>%
   mutate(date = as.Date(date, format = "%Y-%m-%d"))
 ventes_par_mois <- store %>%
   mutate(mois = floor_date(date, "month")) %>%
   group_by(mois) %>%
   summarise(ventes_totales = sum(sales, na.rm = TRUE))
 ventes_par_mois$mois=NULL
 print(ventes_par_mois)

library(tseries)
ventes_par_mois_ts = ts(ventes_par_mois, frequency =
12, start = c(2013, 1))
print(ventes_par_mois_ts)
```

- **Graphiques**
- **Série temporelle**

```
plot(ventes_par_mois_ts, col="black", main="SERIE
TEMPORELLE DES VENTES PAR MOIS")
```

- **Histogramme**

```
hist(ventes_par_mois_ts, main = "Histogramme de
ventes_par_mois_ts", prob=TRUE)
lines(density(ventes_par_mois_ts,na.rm = FALSE))
```

- **Tendance et composante saisonnière**

```
decomposition_add=decompose(ventes_par_mois_ts, type =
"add")
plot(decomposition_add)
```

- **Indice statistique**

```
library(onesime)

onesime_qt_resume(ventes_par_mois_ts)
```

- **Autocorrélation simple**

```
acf(ventes_par_mois_ts,lag.max=10,plot = FALSE,
main="VENTES PAR MOIS")
```

```
acf(ventes_par_mois_ts,lag.max=10,plot = TRUE,
main="VENTES PAR MOIS")
```

Autocorrélation partielle

```
pacf(ventes_par_mois_ts,lag.max=10,plot = FALSE,
main="VENTES PAR MOIS")

pacf(ventes_par_mois_ts,lag.max=10,plot = TRUE,
main="VENTES PAR MOIS")
```

Test de normalité

Graphique

```
library(car)
qqPlot(ventes_par_mois_ts)
```

Test

```
shapiro.test(ventes_par_mois_ts)
```

Méthode Hot-winter

```
library(tseries)
library(forecast)
data(ventes_par_mois_ts)

xlisse <- HoltWinters(ventes_par_mois_ts)
forecast(xlisse,12)

plot(forecast(xlisse,12))

plot(xlisse)
```

Récupération des résidus

```
residus <- residuals(xlisse)
residus
```

Graphique des résidus

```
plot(residus)

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

Box-Ljung test

```
Box.test(residus, lag=20, type="Ljung-Box")
```

Shapiro-Wilk normality test

```
hist(residus)
```

```
shapiro.test(residus)
```

Moyenne des résidus

```
mean(residus)
```