

**I**nstitut **S**upérieur de **S**tatistique, d’**E**conométrie et de **D**ata **S**cience

*Ministère de l’Enseignement Supérieur et de la Recherche Scientifique*

**---------------------------------------------------------------------**

*Union – Discipline - Travail*

**REPUBLIQUE DE CÔTE D’IVOIRE**

**ENCADREUR PEDAGOGIQUE :**

**M. Didier Martial Akposso**

Directeur Fondateur de l’INSSEDS

Economiste-Statisticien

Data-Scientist

Rédigé par :

**SAVANE MORY**

Promotion 2023-2024

Ingénierie Statistique et Data Science

**Période de rédaction : Janvier 2024**

© Année académique : 2023-2024

**ETUDE DE LA SATISFACTION DES PATIENTS DES HOPITAUX PARISIENS**

**AVEC PYTHON**

**MINI-PROJET : STATISTIQUE INFERENTIELLE**

***Pour le diplôme d’Ingénieur Statisticien et Data***

**SUJET** :

# Avant-Propos

Ce travail s'inscrit dans le cadre de la formation au cycle d’ingénieur Statisticien et Data à l’Institut Supérieur de Statistique, d’Économétrie et de Data Science (INSSEDS). L'objectif de ce mini-projet est d'évaluer la compréhension du cours portant sur l'Analyse des séries temporelles et la prédiction à l’aide de la méthode Holt-Winters. Il est important de souligner que ce projet est strictement réalisé à des fins académiques et n'engage en aucun cas la responsabilité de l’école.

En tant qu'auteur, nous reconnaissons humblement que notre travail ne prétend pas avoir exploré tous les aspects nécessaires pour conférer à cette étude le caractère d'une recherche véritablement approfondie. Il s'agit plutôt d'une exploration limitée dans le cadre des enseignements dispensés au sein de notre formation.

Nous souhaitons ainsi offrir une contribution modeste à la compréhension des concepts étudiés tout en restant conscient de la complexité et de l'étendue du domaine de l'analyse des séries temporelles et de la prédiction. Ce travail ne prétend pas être exhaustif, mais plutôt constituer une étape préliminaire dans notre parcours académique.

Nous exprimons notre gratitude envers nos enseignants et toute l'équipe pédagogique de l'INSSEDS pour leur encadrement et leurs enseignements enrichissants qui ont contribué à la réalisation de ce projet.

Enfin, nous espérons que ce travail pourra susciter des réflexions et des discussions fructueuses au sein de la communauté académique et être une source d'apprentissage continu pour tous ceux qui s'intéressent à ce passionnant domaine.

# Sigles

ARMA : Autorégressif et Moyenne Mobile

ACF : Autocorrelation Function

PACF : Partial Autocorrelation Function

# Liste des illustrations

# Listes des figures

[Figure 1: Illustration de la conversion 9](file:///C:\Users\THINKPAD\OneDrive\Bureau\Miniprojets\MiniProjet2\MiniProjet2_SavanéMory_2023.docx#_Toc153491587)

[Figure 2: Visualisation des valeurs aberrantes 9](file:///C:\Users\THINKPAD\OneDrive\Bureau\Miniprojets\MiniProjet2\MiniProjet2_SavanéMory_2023.docx#_Toc153491588)

[Figure 3: Evolution et Répartition des ventes par Mois 10](file:///C:\Users\THINKPAD\OneDrive\Bureau\Miniprojets\MiniProjet2\MiniProjet2_SavanéMory_2023.docx#_Toc153491589)

[Figure 4: Histogramme et densité de la distribution des ventes 11](file:///C:\Users\THINKPAD\OneDrive\Bureau\Miniprojets\MiniProjet2\MiniProjet2_SavanéMory_2023.docx#_Toc153491590)

[Figure 5: Visualisation des différents types de modèle 14](file:///C:\Users\THINKPAD\OneDrive\Bureau\Miniprojets\MiniProjet2\MiniProjet2_SavanéMory_2023.docx#_Toc153491591)

[Figure 6: Décomposition de la série en Tendance, Saisonnalité et Résidu 15](file:///C:\Users\THINKPAD\OneDrive\Bureau\Miniprojets\MiniProjet2\MiniProjet2_SavanéMory_2023.docx#_Toc153491592)

[Figure 7: Autocorrélogrammes 15](file:///C:\Users\THINKPAD\OneDrive\Bureau\Miniprojets\MiniProjet2\MiniProjet2_SavanéMory_2023.docx#_Toc153491593)

[Figure 8: Résultat du modèle Holt-Winters 16](#_Toc153491594)

[Figure 9: Entrainement et test du modèle 17](file:///C:\Users\THINKPAD\OneDrive\Bureau\Miniprojets\MiniProjet2\MiniProjet2_SavanéMory_2023.docx#_Toc153491595)

[Figure 10: Graphiques des tests 18](#_Toc153491596)

[Figure 11: Prévision des ventes sur 12 mois 19](file:///C:\Users\THINKPAD\OneDrive\Bureau\Miniprojets\MiniProjet2\MiniProjet2_SavanéMory_2023.docx#_Toc153491597)

# Liste des tableaux

[Tableau 1: Dictionnaire des données 8](#_Toc153491686)

[Tableau 2: Résumé numérique des quantités vendues 11](#_Toc153491687)

[Tableau 3: Test de Stationnarité et normalité 16](#_Toc153491688)

[Tableau 4: Coefficients du modèle Holt-Winters 17](#_Toc153491689)

[Tableau 5: Résultat des tests 18](#_Toc153491690)

[Tableau 6: Prédiction des valeurs 19](#_Toc153491691)

# Sommaire

[Avant-Propos 1](#_Toc153576832)

[Sigles 2](#_Toc153576833)

[Liste des illustrations 3](#_Toc153576834)

[Listes des figures 3](#_Toc153576835)

[Liste des tableaux 3](#_Toc153576836)

[Sommaire 4](#_Toc153576837)

[Résumé 6](#_Toc153576838)

[Abstract 7](#_Toc153576839)

[Introduction Générale 9](#_Toc153576840)

[I. STATISTIQUE DESCRIPTIVE DES VENTES 11](#_Toc153576841)

[Introduction 11](#_Toc153576842)

[1. Prétraitement des données 11](#_Toc153576843)

[a) Présentation des données 11](#_Toc153576844)

[b) Nettoyage des données 12](#_Toc153576845)

[2. Analyse descriptive de l’évolution des ventes entre 2013 et 2017 13](#_Toc153576846)

[a) Analyse des graphiques et des indicateurs 13](#_Toc153576847)

[b) Interprétation 15](#_Toc153576848)

[Conclusion 15](#_Toc153576849)

[II. ANALYSE TEMPORELLE ET PREVISION DES VENTES 16](#_Toc153576850)

[Introduction 16](#_Toc153576851)

[1. Série temporelle 16](#_Toc153576852)

[a) Définition des concepts 16](#_Toc153576853)

[b) Le modele de prévision de Holt-Winters 17](#_Toc153576854)

[2. Conditions préliminaires 18](#_Toc153576855)

[a) Décomposition de la série 18](#_Toc153576856)

[b) Vérification de la stationnarité 18](#_Toc153576857)

[3. Prévision des ventes d’articles 20](#_Toc153576858)

[a) Construction du modèle 20](#_Toc153576859)

[b) Validation du modèle 21](#_Toc153576860)

[c) Prédiction des ventes pour les 12 prochains mois. 23](#_Toc153576861)

[Conclusion 24](#_Toc153576862)

[Conclusion générale 25](#_Toc153576863)

[Annexe 27](#_Toc153576864)

[Bibliographie et Webographie 30](#_Toc153576865)

# Résumé

L'étude vise à élaborer une technique robuste de prévision des ventes pour les épiceries Favorita en Équateur en utilisant la méthode de Holt-Winters. Ancrée dans le contexte dynamique du commerce de détail, l'objectif principal est de déterminer si ce modèle peut efficacement saisir la tendance à la hausse des ventes et la saisonnalité, fournissant ainsi des prévisions fiables pour orienter la gestion des ventes et les décisions stratégiques de l'entreprise.

L'analyse des données historiques de 2013 à 2017 révèle une croissance constante des ventes, avec des performances marquées en décembre, atteignant environ 25 millions d'unités. La distribution des ventes, caractérisée par deux sommets successifs, indique une propension à vendre en grande quantité. La méthode de Holt-Winters est validée par des tests de résidus démontrant sa robustesse avec des résidus stationnaires, un bruit blanc, et une distribution gaussienne.

En guise de recommandation, nous conseillons l'adoption de la méthode de Holt-Winters pour la prévision des ventes. Les prévisions générées doivent être mises en œuvre pour optimiser la gestion des stocks, planifier les opérations, surtout pendant les périodes de pointe. Toutefois, il est crucial de reconnaître les limites de l'analyse, notamment la dépendance aux tendances passées. Des améliorations potentielles incluent l'intégration de variables externes et une évaluation continue du modèle pour une adaptation proactive aux changements. Cette étude contribue ainsi à la précision des prévisions dans le secteur du commerce de détail, offrant des perspectives stratégiques pour les épiceries Favorita en Équateur.

Mots clés : Séries temporelles ; Prédiction ; méthode de Holt-Winters

# Abstract

This study aims to develop a robust sales forecasting technique for Favorita supermarkets in Ecuador using the Holt-Winters method. Rooted in the dynamic context of the retail industry, the primary objective is to determine whether this model can effectively capture the upward trend in sales and seasonality, thus providing reliable forecasts to guide inventory management and strategic decision-making for the company.

Analysis of historical data from 2013 to 2017 reveals a consistent growth in sales, with notable performances in December, reaching around 25 million units. The sales distribution, characterized by two successive peaks, indicates a tendency to sell in large quantities. The Holt-Winters method is validated through residual tests, demonstrating its robustness with stationary residuals, white noise, and a Gaussian distribution.

In conclusion, we recommend the continued adoption of the Holt-Winters method for sales forecasting. The generated forecasts should be implemented to optimize inventory management and operational planning, especially during peak periods. However, it is crucial to acknowledge the limitations of the analysis, including reliance on past trends. Potential improvements involve incorporating external variables and ongoing model evaluation for proactive adaptation to changes. This study contributes to the accuracy of forecasts in the retail sector, offering strategic insights for Favorita supermarkets in Ecuador.

Key words: Temporal series; prediction; Holt-Winters method

# Introduction Générale

Dans le contexte dynamique du commerce de détail, où la compétitivité est cruciale, la capacité à anticiper les tendances de vente demeure un impératif pour les entreprises. Notre étude se positionne au cœur de cette nécessité en cherchant à estimer correctement les ventes et à gérer les stocks. Le problème clé est de prévoir les besoins en ventes et en stocks pour chaque emplacement afin d’éviter les surstocks et les sous-stocks, permettant ainsi à l’entreprise d’offrir le meilleur service client tout en réduisant les pertes et en garantissant la durabilité du magasin. Face à l'évolution constante du marché, l'importance stratégique d'outils de prévision précis et adaptables devient indéniable pour orienter les décisions opérationnelles et optimiser la gestion des stocks.

Au cœur de notre réflexion réside la question centrale de savoir si ***le modèle de Holt-Winters peut véritablement saisir la dynamique de la tendance à la hausse des ventes et la saisonnalité propre aux épiceries Favorita en Équateur***. Cette interrogation constitue le fil conducteur de notre étude, visant à évaluer la pertinence de ce modèle dans la fourniture de prévisions fiables, cruciales pour une gestion proactive des stocks et des décisions stratégiques éclairées.

Notre analyse a donc pour objectif principal d’élaborer une technique de prédiction robuste pour les épiceries Favorita en Équateur.

L’atteinte de cet objectif passera par la résolution des objectifs spécifiques ci-dessous :

* Déterminer la présence de tendance
* Déterminer la présence de saisonnalité
* Eliminer les effets de tendance et de saisonnalité
* Vérifier les résidus

Pour mener à bien ce travail, nous posons les hypothèses suivantes :

1. L’évolution des ventes a une tendance haussière et une saisonnalité
2. Le modèle de Holt-Winters est adapté pour la prévision des ventes

Pour atteindre ces objectifs, nous avons adopté une méthodologie rigoureuse. L'analyse repose sur les données historiques des ventes de 2013 à 2017 des épiceries Favorita en Équateur. Le choix délibéré de la méthode de Holt-Winters, en raison de sa capacité à modéliser conjointement la tendance et la saisonnalité, structure notre approche. Les tests de résidus ont été appliqués comme un indicateur critique de la qualité du modèle. Notre démarche intègre également une réflexion approfondie sur les limites inhérentes à l'analyse, offrant ainsi des perspectives pour une amélioration continue des résultats.

Ainsi, afin de ressourdre efficacement cette étude, nous l’avons érigé essentiellement en deux parties, d’une part, l’analyse descriptive des ventes et d’autres parts, la prédiction des ventes sur 12 mois.

PARTIE I : PRETRAITEMENT ET STATISTIQUE DESCRIPTIVE

# PRETRAITEMENT DU JEU DE DONNEES

## Introduction

Le nettoyage des données ou data cleansing est une étape essentielle de l’analyse statistique et la Data Science. Il s’agit de corriger ou supprimer des enregistrements inexacts dans des jeux de données afin de pouvoir les exploiter par la suite.

C’est un processus qui vise, de manière générale, à améliorer **la qualité des données**. En pratique, il consiste à **importer les données** dans un logiciel statistique, dans notre Jupyter Notebook du logiciel Python. Ensuite, faire une **première visualisation** pour avoir une idée succincte de la structure de nos données et enfin, identifier et corriger les valeurs manquantes, valeurs aberrantes, valeurs extrêmes afin de qu’elles puissent être plus cohérentes et sans erreurs.

## Présentation des données

Les données utilisées dans ce rapport sont issues de la plateforme dédiée à la Data Science, **Kaggle[[1]](#footnote-1)**.

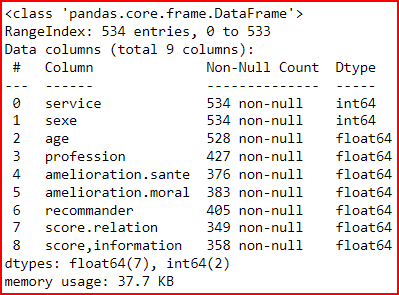
Elles portent sur l’évaluation de la qualité de relation et la quantité d’information reçue par le patient lors de son séjour à l’hôpital. Les patients ont été enquêtés dans plusieurs hôpitaux de la région parisienne sur 9 variables récapitulées ci-dessous.

Figure 1: Informations initiales du jeu de données

Le dictionnaire des données se présente comme suit :

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Nature/Type | Description | Modalités |
| service | category (Qualitative) | Numéro des services ayant accueilli le patient | De 1 à 8 |
| sexe | category (Qualitative) | Le genre du patient | homme et femme |
| age | int64 (Quantitative) | Age en année | Age compris entre 27 et 85 |
| profession | category (Qualitative) | Le travail exercé par les patients enquetés | 'agriculteur', 'artisan', 'cadre', 'prof\_intermédiaire', 'employé', 'ouvrier', 'sans emploi', 'autre' |
| amelioration.sante | category (Qualitative) | Les quantités d’articles vendues | Numérique |
| amelioration.moral | category (Qualitative) | Nombre de produits en promotion | Numérique |
| recommander | category (Qualitative) | Niveau de recommandation à l’entourage | 0 : Non ; 1 : oui, probablement ; 2 : oui, sûrement |
| score.relation | int64 (Quantitative) | Notation de la qualité de la relation avec le personnel soignant pendant le séjour | De 10 à 40 |
| ScoreInfo | int64 (Quantitative) | Notation de la qualité de l’information recue pendant le séjour | De 10 à 40 |

Tableau 1: Dictionnaire des données

*Source : Réalisé par l’auteur*

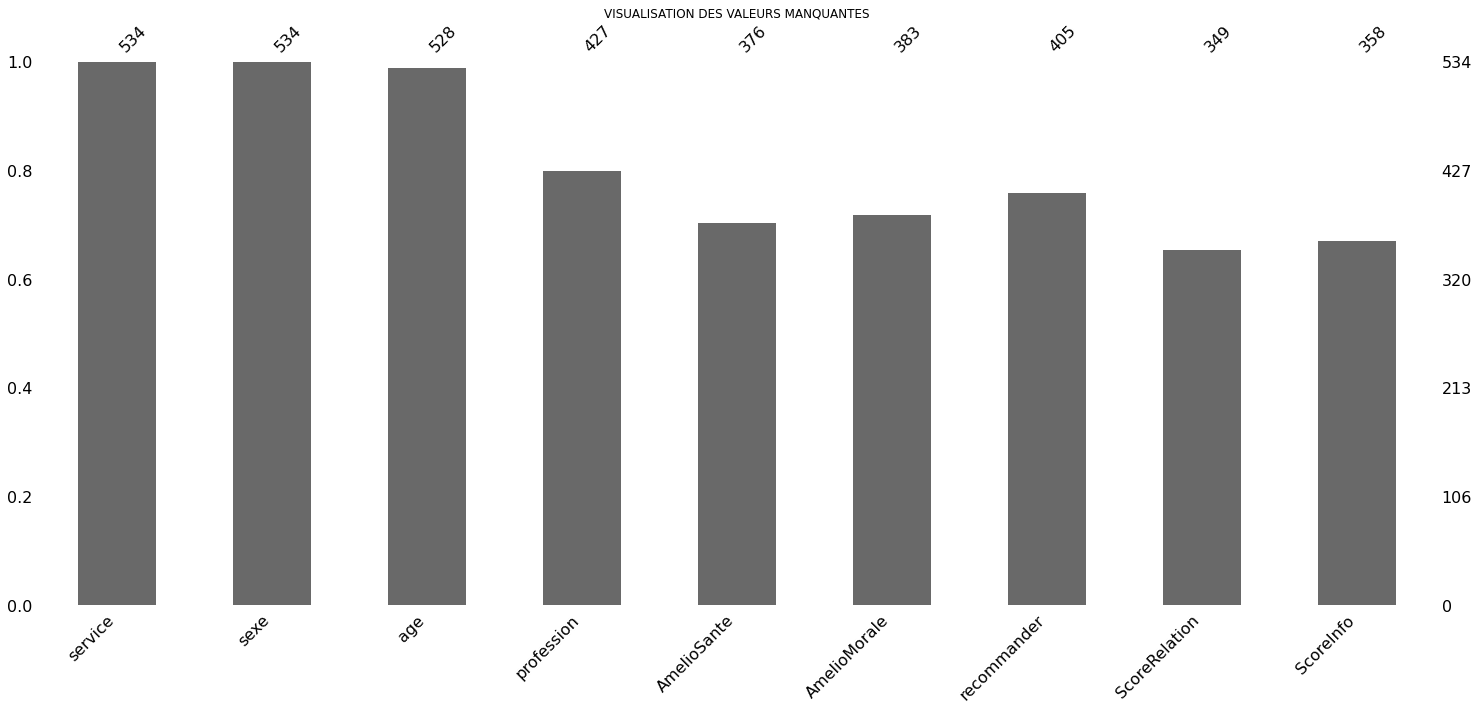
En résumé, la lecture des informations de bases de nos données nous montre des anomalies telles que : l’existence de valeurs manquantes et un mauvais typage de certaines variables. Dans la suite de notre analyse, nous allons traiter les valeurs manquantes, les valeurs aberrantes et recorder certaines variables si nécessaire.

## Traitement des valeurs manquantes et doublons

* *Traitement des valeurs manquantes*

Nous allons visualiser les valeurs manquantes (NaN) grâce au diagramme des données manquantes avant et après traitement.

Figure 2: Visualisation des données manquantes avant traitement



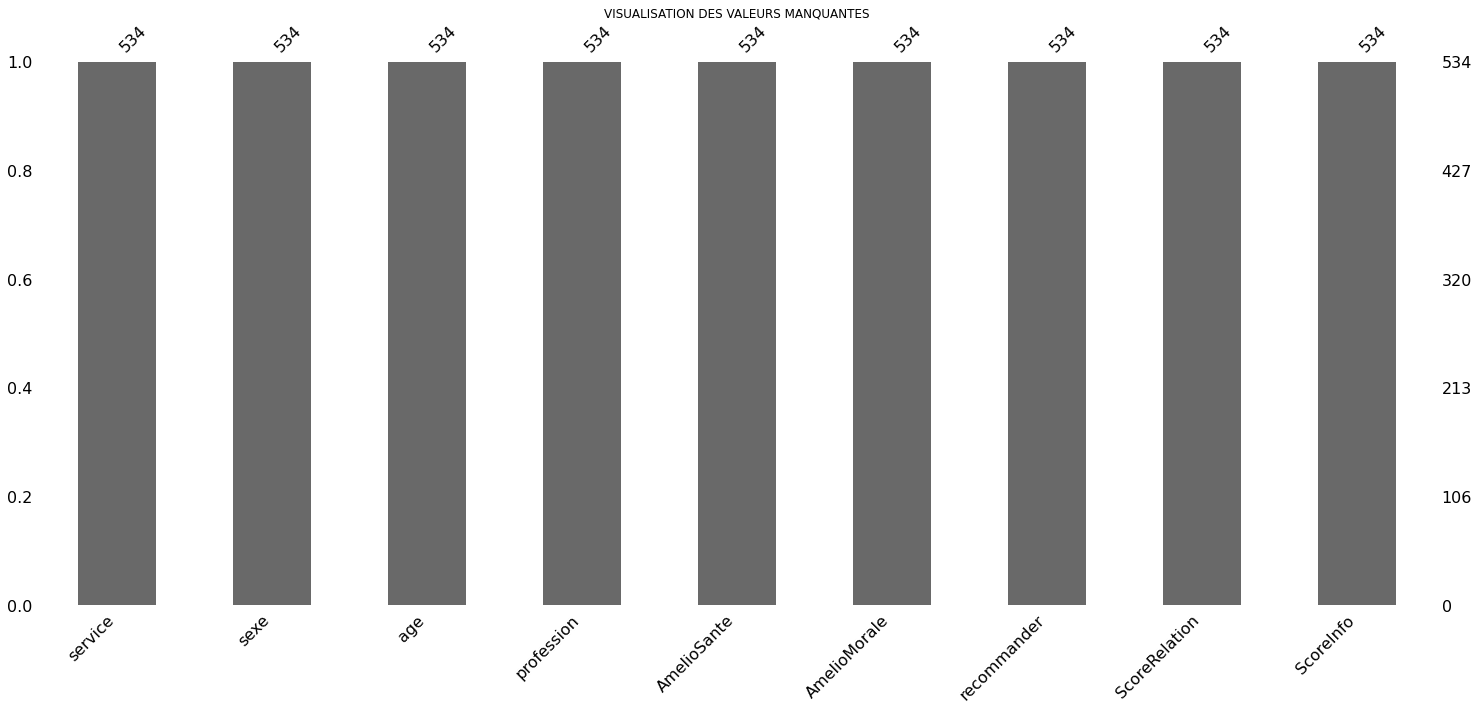
*Source : Réalisé par l’auteur avec Python*

Tableau 2: Tableau des données manquantes par variables

|  |  |
| --- | --- |
| **Variables** | **Nombre de NaN** |
| service | 0 |
| sexe | 0 |
| age | 6 |
| profession | 107 |
| AmelioSante | 158 |
| AmelioMorale | 151 |
| recommander | 129 |
| ScoreRelation | 185 |
| ScoreInfo | 176 |

Nous pouvons pertinemment constater l’existence de données manquantes dans toutes les variables exceptées les colonnes **sexe** et **service.** Nous allons donc procéder par imputation avec la méthode des K-NN (K plus proches voisins).

Figure 3: Résultat après imputation



*Source : Réalisé par l’auteur avec Python*

Le diagramme ci-dessus nous permet de constater qu’effectivement l’imputation a fonctionné et que nos données sont complètes.

Nous pouvons donc passer à la vérification des doublons.

* *Traitement des doublons*

Selon la plateforme de formation en ligne DataScientest, la présence de données redondantes est un problème très fréquent dans la plupart des organisations.

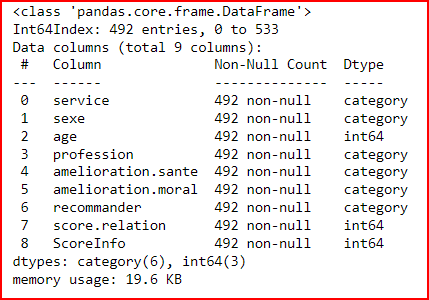
Les doublons peuvent provenir d’une collecte d’informations identiques provenant de différentes sources, d’erreurs humaines, ou encore de données ajoutées au lieu d’être mises à jour.

*Les doublons peuvent fausser tout type d’analyse de données, voire entraîner indirectement de mauvaises prises de décision.*

Les données redondantes peuvent également coûter très cher à l’entreprise si elles sont nombreuses. Heureusement, il existe sur Python, par exemple, des fonctions (comme *drop\_duplicates* de Pandas) permettant de se débarrasser de doublons très facilement. (DataScientest 2023)

Nous constatons la présence de 42 doublons dans les données soumises à notre analyse. La fonction *drop\_duplicates* de python nous permet de nous en débarrasser.

En voici le résultat :



*Source : Réalisé par l’auteur avec Python*

Figure 4: Traitement des doublons

Nous constatons que l’effectif des données passe de 534 à 492 suite à la suppression des individus en double.

## Traitement des valeurs aberrantes

En statistiques, une valeur aberrante est une valeur qui diffère grandement de la distribution d’une variable. Il s’agit d’une observation anormale, qui s’écarte de données par ailleurs bien structurées.

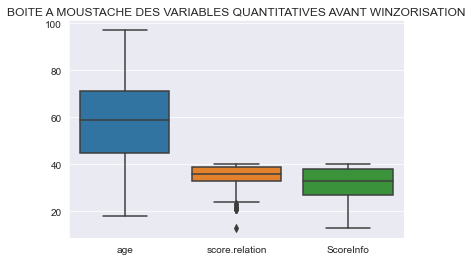
La détection des valeurs aberrantes ou des anomalies est l’un des problèmes fondamentaux de l’exploration des données. L’expansion émergente et continue des données nous fait repenser la façon dont nous abordons les anomalies et les cas d’utilisation qui peuvent être construits en examinant ces anomalies.

En Data Science, les valeurs aberrantes peuvent affecter certains paramètres statistiques, comme la moyenne. Si les Outliers ne sont pas détectés, cela peut fausser notre compréhension d’un jeu de données et nous conduire à émettre des hypothèses erronées sur ce dernier.

Une autre raison pour laquelle il est important de s’intéresser aux valeurs aberrantes est qu’une majorité des algorithmes de Machine Learning sont très sensibles aux données sur lesquelles ils sont entraînés, ainsi qu’à leurs distributions.

Avoir des Outliers dans le jeu d’entraînement d’un modèle de Machine Learning peut rendre la phase d’entraînement plus longue et potentiellement biaisée.

Par conséquent, le modèle de prédiction produit sera moins performant ou moins précis. (DataScientest 2023)

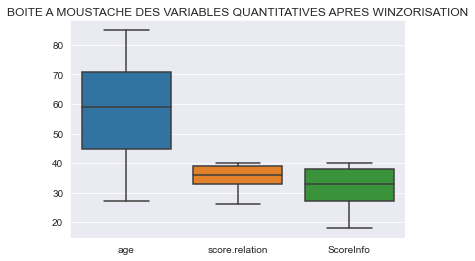
Figure 5: Boîte à moustache avant Winzorisation

*Source : Réalisé par l’auteur avec Python*

Les boîtes à moustache ci-dessous permettent de constater que la variable **score.relation** est la seule qui présente des valeurs aberrantes au niveau de la moustache inferieur.

La technique de Winzorisation[[2]](#footnote-2) consiste à remplacer les valeurs extrêmes par des valeurs moins extrêmes afin de réduire leur impact sur les analyses statistiques.

La figure ci-dessous nous montre qu’effectivement nos données sont maintenant exemptes de valeurs aberrantes.

Figure 6: : Boîte à moustache après Winzorisation

*Source : Réalisé par l’auteur avec Python*

# Conclusion

En conclusion, Cette première phase de prétraitement a permis de traiter les valeurs manquantes, qui étaient relativement élevées dans notre jeu de données, ainsi que les doublons qui étaient au nombre de 42. De plus, les valeurs extrêmes et aberrantes ont découvertes au niveau de la variable **score.relation** et ont été ramenées à la borne inférieur.

Ce travail préalable à permis de rendre nos données beaucoup plus aptes à l’analyse. La seconde phase consistera à faire l’analyse descriptive univarié de nos variables d’intérêt.

# STATISTIQUE DESCRIPTIVE UNIVARIÉE

# Introduction

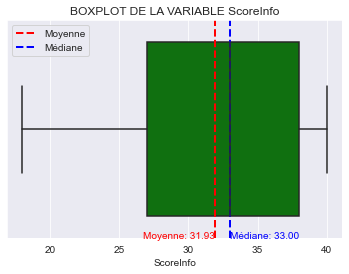
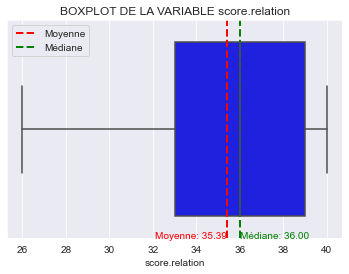
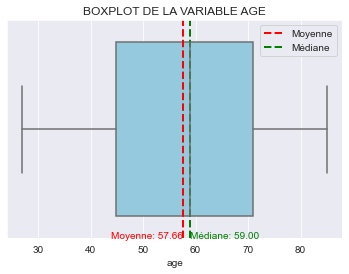
Selon Bernard PY, dans son livre *Statistique descriptive, nouvelle méthode pour bien comprendre et réussir (éditions Economica)* : « La statistique [descriptive] est un ensemble de méthodes permettant de décrire et d'analyser, de façon quantifiée, des phénomènes repérés par des éléments nombreux, de même nature, susceptibles d'être dénombrés et classés. » (MAZEROLLE 2006).

Cette partie nous permettra de mieux cerner les spécificités de nos données grâce aux différents indicateurs statistiques et aux graphiques.

## Analyse univariée des variables quantitatives

Les variables quantitatives sont au nombre de trois dans notre jeu de données, ce sont les variables : **sexe**, **ScoreInfo** et **score.relation**. Visualisons tout d’abord quelques graphiques et appuyons cela par des indicateurs statistiques.

Figure 7: Boxplots des variables quantitatives



*Source : Réalisé par l’auteur avec Python*

La visualisation des graphiques ci-dessus met en exergue les boîtes à moustache des trois variables avec les principales caractéristiques qui sont les moyennes et les médianes.

Nous constatons que pour ces trois variables, les médianes sont relativement supérieures aux moyennes. Ce qui indique que les moyennes sont beaucoup influencées pour la queue inférieure de la distribution, c’est-à-dire par les petites valeurs qui sont dans la distribution.

Le tableau ci-dessous nous permettra de clarifier cette situation.

|  |  |  |  |
| --- | --- | --- | --- |
| Indicateurs | age | score.relation | ScoreInfo |
| INDICATEUR DE TENDANCE CENTRALE ET DE POSITION | | | |
| Minimum | 27 | 26 | 18 |
| Maximum | 85 | 40 | 40 |
| Moyenne | 57.66 | 35.39 | 31.93 |
| Ecart-type | 16.99 | 4.22 | 6.64 |
| Mode | 27 | 40 | 40 |
| INDICATEUR DE FORME | | | |
| Skewness | -0.15 | -0.77 | -0.54 |
| Kurtosis | -0.99 | -0.46 | -0.77 |

Tableau 3: Résumé statistiques

*Source : Réalisé par l’auteur avec Python*

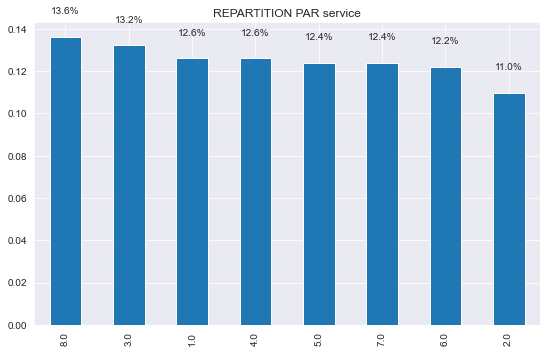
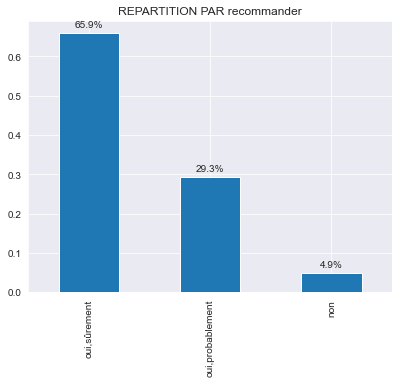
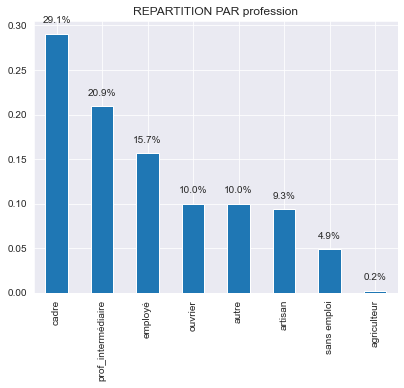
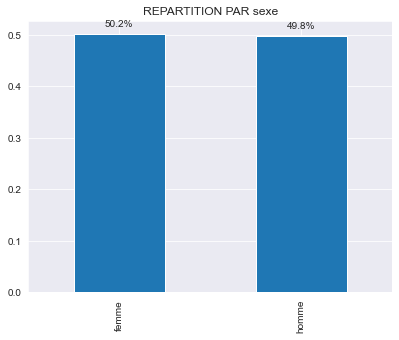
Le tableau ci-dessus vient conforter l’analyse graphique. Il nous permet d’affirmer **qu’en moyenne les patients sont âgés d’environ 58 ans avec un écart type de 17 ans**, cependant, l’âge minimal et modal est de 27 ans.

L’analyse des indicateurs de tendance centrale des deux autres variables montre que **les scores moyens** attribués par les patients concernant la qualité de la relation avec le personnel soignant et la qualité de l’information reçue sont respectivement **d’environ 35 et de 32** sur une échelle de 10 à 40 ce qui montre qu’ils sont **globalement satisfaits.**

Les valeurs des indicateurs de forme montrent que les distributions sont pour la plupart étalées à gauche et platikurtique.

## Analyse univariée des variables qualitatives

Figure 8: Diagramme en barre des variables qualitatives



*Source : Réalisé par l’auteur sous Python*

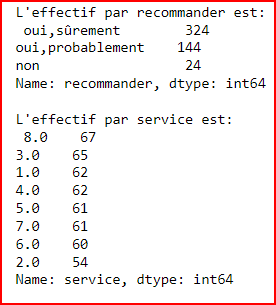
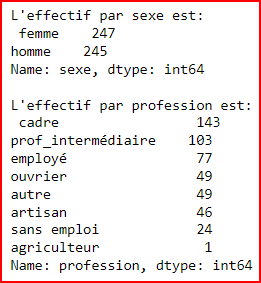
Les graphiques ci-dessus met en évidence que 50,2% des enquêtés sont de sexe masculin. Pratiquement **66% des patients recommandent les services** par lesquels ils sont passés durant le séjour à l’hôpital, seulement 5% environ ne recommandent pas.

*Source : Réalisé par l’auteur avec Python*

De plus, nous constatons que la majorité des patients sont des cadres (29%) suivi des professions intermédiaires (20%) et en dernière position les agriculteurs (0,2%).

*Source : Réalisé par l’auteur sous Python*

Figure 9: Tableau des effectifs des variables qualitatives



## Conclusion

La statistique descriptive a permis de mieux comprendre les spécificités des nos différentes variables d’intérêt. Elle nous a permis de savoir que les patients sont majoritairement des cadres adultes qui ont globalement été satisfait des services des hôpitaux où ils ont séjourné.

# Conclusion

Au terme de cette première partie, il ressort que les patients ont un profil de cadre d’entreprise de 57 ans en moyenne. Ils ont globalement apprécié la relation avec le personnel soignant en leur attribuant un score de 35/40 en moyenne. De plus, ces patients ont jugé avoir reçu le maximum d’information durant leur passage à l’hôpital.

La partie suivante nous permettra de tester la pertinence des résultats de ces analyses simplement descriptive.

PARTIE II : STATISTIQUE INFERENTIELLE

# TEST DE COMPARAISON DE DEUX POPULATIONS INDEPENDANTES

## La proportion de sujets recommandant le service dans lequel ils sont passés est-il le même quel que soit le genre ?

## La moyenne du score de relation est-il significativement différent chez les hommes et chez les femmes ?

# TEST DE COMPARAISON DE PLUSIEURS POPULATIONS INDEPENDANTES

## Peut-on affirmer que la qualité de l’information reçue est la même quelle que soit la profession ?

## La qualité des relations avec le personnel soignant est-il fonction du service ayant accueilli le patient ?

# TEST DE LIAISON OU D’INDEPENDANCE : Le score de relation est-il significativement corrélé à l’âge ?

# Conclusion générale

Notre étude visait à élaborer une technique robuste de prévision des ventes des épiceries Favorita en Équateur grâce à la méthode de Holt-Winters. La préoccupation majeure était de déterminer si ce modèle pouvait efficacement capturer la tendance à la hausse des ventes et la saisonnalité observée avec pour but ultime de fournir des prévisions fiables afin d'éclairer la gestion des stocks et de soutenir les décisions stratégiques de l'entreprise.

L’analyse des données historiques des ventes de 2013 à 2017 a révélé **une croissance constante des quantités d'articles vendues au fil des années, avec des performances particulièrement fortes en décembre**, où les ventes atteignent environ 25 millions d'unités. **La distribution des ventes présente une forme irrégulière**, indiquant une tendance à vendre en grande quantité. **La méthode de Holt-Winters a été confirmée comme un choix judicieux**, avec des tests de résidus démontrant la robustesse du modèle, ses résidus étant stationnaires, présentant un bruit blanc et suivant une distribution gaussienne.

Sur la base de ces résultats, nous recommandons **l'adoption continue de la méthode de Holt-Winters pour la prévision des ventes**. Il est conseillé de mettre en œuvre les prévisions générées pour optimiser la gestion des stocks, planifier les opérations et améliorer la performance commerciale, en particulier pendant les périodes de pointe comme le mois de décembre. De plus, il serait bénéfique de surveiller de près les facteurs externes susceptibles d'influencer les ventes, tels que les tendances du marché et les changements saisonniers.

Cependant, il est important de noter que toute analyse comporte des limites. Dans notre cas, les résultats sont basés sur des données historiques et supposent une continuité des tendances passées, ce qui peut ne pas toujours être le cas. Les facteurs externes imprévus peuvent également influencer les ventes de manière significative. De plus, la méthode de Holt-Winters suppose une stabilité des paramètres, ce qui peut nécessiter une réévaluation régulière.

Alors, pour améliorer les résultats, des analyses plus approfondies pourraient être envisagées, telles que l'intégration de variables externes (comme les promotions ou les événements spéciaux) dans le modèle. Une évaluation continue de la performance du modèle, avec des ajustements si nécessaire, serait également recommandée. Des analyses de sensibilité pour évaluer l'impact de différentes hypothèses sur les résultats pourraient fournir une compréhension plus approfondie des prévisions générées par le modèle. Enfin, des comparaisons avec d'autres méthodes de prévision pourraient être effectuées pour évaluer la pertinence continue de la méthode de Holt-Winters dans un contexte évolutif.

# Annexe

|  |
| --- |
| **Importation des packages** |
| import numpy as np  import pandas as pd  import matplotlib.pyplot as plt  import seaborn as sns  from statsmodels.tsa.holtwinters import ExponentialSmoothing as HWES  import datetime  from sklearn.metrics import mean\_squared\_error  from statsmodels.tsa.seasonal import seasonal\_decompose, STL  from statsmodels.graphics.tsaplots import plot\_acf, plot\_pacf  from statsmodels.tsa.stattools import adfuller  from scipy.stats import shapiro  from scipy.stats.mstats import normaltest  import statsmodels.api as sm  import statsmodels.tsa.api as smt |
| **Importation et Organisation des données par mois** |
| # Importation des donnees  favorita = pd.read\_csv("epicerie.csv", index\_col='date', parse\_dates=True)  df = favorita.copy()  # Organisation par mois  df = df.resample('M').sum()  # Comptage des NaN  df.isna().sum()  # Détecter les doublons  len(df[df.duplicated()])  # Analyse des ventes  sales = df['sales']  sales.describe() |
| **Prétraitement des données** |
| # Valeurs Aberrantes  sns.set(rc={'figure.figsize':(14,10)})  sns.boxplot(x=df.index.month\_name(), y='sales', data=df, palette='muted')  plt.ylabel('Ventes')  plt.xlabel('Mois')  plt.title("BOITES A MOUSTACHE DES QUANTITES VENDUES PAR MOIS")  plt.savefig("BOITES A MOUSTACHE.png", dpi=300)  # Détecter les doublons  len(df[df.duplicated()]) |
| **Visualisation** |
| sns.set\_style('dark')  plt.figure(figsize=(10, 7))  sns.distplot(a=sales, bins=10, color='blue',hist\_kws={"edgecolor": 'white'})  plt.xlabel('Ventes')  plt.title("REPARTITION DES QUANTITES VENDUES")  # Representation des ventes  sns.set\_style("darkgrid")  plt.figure(figsize=(15, 5))  plt.subplot(111)  plt.plot(sales, marker = 'o')  plt.title("EVOLUTION DES VENTES PAR MOIS")  plt.xlabel("Dates")  plt.show()  composants1 = seasonal\_decompose(sales, model='additive')  composants1.plot();  plot\_acf(df.sales);  plot\_pacf(df.sales); |
| **Prevision: Construction du modele** |
| df\_train = sales.iloc[:-12]  df\_test = sales.iloc[-12:]  model = HWES(df\_train, seasonal\_periods=12, trend='add', seasonal='add')  fitted = model.fit()  print(fitted.summary())  sales\_forecast = fitted.forecast(steps=12)  print(df\_test, sales\_forecast)  fig = plt.figure()  fig.suptitle('EVOLUTION DES VENTES DANS LES EPICERIES FAVORITA 2013-2017 ')  past, = plt.plot(df\_train.index, df\_train, 'b.-', label='Ventes Hystoriques')  future, = plt.plot(df\_test.index, df\_test, 'r.-', label='Ventes Actuelles')  predicted\_future, = plt.plot(df\_test.index, sales\_forecast, 'g.-', label='Prévision des ventes')  plt.legend(handles=[past, future, predicted\_future])  plt.show() |
| **Test sur les residus** |
| from statsmodels.stats.diagnostic import acorr\_ljungbox  res=fitted.resid  # Test de Bruit-blanc  acorr\_ljungbox(res, return\_df=True)  # Tes de normalite  shapiro(res)  # stationarite  adf\_testRes = adfuller(res)  print(f'ADF Statistic: {adf\_testRes[0]}')  print(f'p-value: {adf\_testRes[1]}')  def ts\_plot(y, lags=None, title='', fig\_size=(14,12)):  """  Calcul de l'acf, pacf, de l'histogramme et du QQ-plot d'une série temp  """  # on transforme en Series si l'argument y n'en est pas une  if not isinstance(y, pd.Series):  y = pd.Series(y)    # initialisation de la figure et des axes  fig = plt.figure(figsize=fig\_size)  ts\_ax = fig.add\_subplot(311)  acf\_ax = fig.add\_subplot(323)  pacf\_ax = fig.add\_subplot(324)  qq\_ax = fig.add\_subplot(325)  hist\_ax = fig.add\_subplot(326)    # la serie temporelle  y.plot(ax=ts\_ax)  ts\_ax.set\_title(title);    # ACF et PACF  smt.graphics.plot\_acf(y, lags=lags, ax=acf\_ax, alpha=0.05)  smt.graphics.plot\_pacf(y, lags=lags, ax=pacf\_ax, alpha=0.05, method='ywm')    # QQ-plot  sm.qqplot(y, line='s', ax=qq\_ax)  qq\_ax.set\_title('QQ-Plot')    # histogramme  y.plot(ax=hist\_ax, kind='hist', bins=25);  hist\_ax.set\_title('Histogramme');  plt.tight\_layout();  plt.show() |
| **Modèle final** |
| modelFin = HWES(sales, seasonal\_periods=12, trend='add', seasonal='add')  fittedFin = modelFin.fit()  prevision = fittedFin.forecast(12)  sales.plot(legend=True, label='sales', figsize=(10,6))  prevision.plot(legend=True, label='prediction') |

# Bibliographie et Webographie

Akposso, Didier Martial. «PREDICTION DES SERIES TEMPORELLES.» *LISSAGE EXPONENTIEL SIMPLE-DOUBLE-HOLT WINTERS AVEC R.* 2022.

Köseoğlu, Buse. *Medium.* 15 07 2023. https://buse-koseoglu13.medium.com/guide-to-time-series-analysis-with-python-2-moving-average-process-784328325e5f (accès le 12 12, 2023).

Louis. *STATOSCOP.* 04 Mai 2021. https://blog.statoscop.fr/timeseries-1.html (accès le 12 12, 2023).

MAZEROLLE, Fabrice. *Statistique Descriptive.* Marseille: Gualino Editeur, 2006.

**Table des matières**

[Avant-Propos 1](#_Toc153576462)

[Sigles 2](#_Toc153576463)

[Liste des illustrations 3](#_Toc153576464)

[Listes des figures 3](#_Toc153576465)

[Liste des tableaux 3](#_Toc153576466)

[Sommaire 4](#_Toc153576467)

[Résumé 5](#_Toc153576468)

[Abstract 6](#_Toc153576469)

[Introduction Générale 9](#_Toc153576470)

[I. STATISTIQUE DESCRIPTIVE DES VENTES 11](#_Toc153576471)

[Introduction 11](#_Toc153576472)

[1. Prétraitement des données 11](#_Toc153576473)

[a) Présentation des données 11](#_Toc153576474)

[b) Nettoyage des données 12](#_Toc153576475)

[2. Analyse descriptive de l’évolution des ventes entre 2013 et 2017 13](#_Toc153576476)

[a) Analyse des graphiques et des indicateurs 13](#_Toc153576477)

[b) Interprétation 15](#_Toc153576478)

[Conclusion 15](#_Toc153576479)

[II. ANALYSE TEMPORELLE ET PREVISION DES VENTES 16](#_Toc153576480)

[Introduction 16](#_Toc153576481)

[1. Série temporelle 16](#_Toc153576482)

[a) Définition des concepts 16](#_Toc153576483)

[b) Le modele de prévision de Holt-Winters 17](#_Toc153576484)

[2. Conditions préliminaires 18](#_Toc153576485)

[a) Décomposition de la série 18](#_Toc153576486)

[b) Vérification de la stationnarité 18](#_Toc153576487)

[3. Prévision des ventes d’articles 20](#_Toc153576488)

[a) Construction du modèle 20](#_Toc153576489)

[b) Validation du modèle 21](#_Toc153576490)

[c) Prédiction des ventes pour les 12 prochains mois. 23](#_Toc153576491)

[Conclusion 24](#_Toc153576492)

[Conclusion générale 25](#_Toc153576493)

[Annexe 27](#_Toc153576494)

[Bibliographie et Webographie 30](#_Toc153576495)

1. https://www.kaggle.com/datasets/samawel97/satisfaction-hospital/code [↑](#footnote-ref-1)
2. Une technique de transformation des données utilisée en statistiques pour atténuer l'impact des valeurs extrêmes, des valeurs aberrantes ou des points de données influents dans un ensemble de donnée. [↑](#footnote-ref-2)