

**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 la statistique inférentielle. 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 la statistique inférentielle. 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

Cette étude vise à évaluer la qualité de la relation et la quantité d’information reçue par le patient lors de son séjour à l’hôpital. L'objectif principal est donc de tester si les patients qui fréquentent les hôpitaux de la région parisienne sont globalement satisfait ou pas et le ou les facteurs contribuant le plus à leur satisfaction.

L’analyse a montré que les patients sont majoritairement des cadres d’entreprises et adultes d’âge moyen, 57 ans.

En ce qui concerne le résultat des tests statistiques, il ressort que :

* **Les patients masculins comme féminin recommandent les services** par lesquels ils sont passés à leur entourage. Ce qui signifie que **le personnel soignant ne fait pas de distinction entre les genres** dans le traitement des patients.
* **Les patients sont satisfaits de la qualité de la relation** avec le personnel soignant,
* La quantité d’information reçue du personnel soignant est la même quelque soit la profession du patient,
* La qualité de la relation avec le patient ne dépend pas du service fréquenté,
* La qualité de la relation ne dépend pas de l’âge du patient.

Ces résultats prouvent que les hôpitaux enquêtés sont globalement juste dans le traitement des patients.

Pour finir, nous félicitons le personnel soignant pour son professionnalisme et l’encourageons à continuer dans ce sens de respect de la dignité humaine.

Mots clés : Statistique inférentielle, test statistique, Satisfaction, évaluation, patients

# Abstract

This study aims to assess the quality of the relationship and the amount of information received by patients during their hospital stay. The main objective is to determine whether patients attending hospitals in the Parisian region are generally satisfied or not and to identify the factors contributing most to their satisfaction.

The univariate analysis revealed that the majority of patients are executives from businesses and middle-aged adults, with an average age of 57 years.

As for the results of the statistical tests, it is evident that:

- Both male and female patients recommend the services they have experienced to their acquaintances, indicating that healthcare personnel do not differentiate between genders in patient treatment.

- Patients are satisfied with the quality of the relationship with healthcare staff.

- The amount of information received from healthcare personnel is consistent regardless of the patient's profession.

- The quality of the relationship with the patient does not depend on the specific hospital department visited.

- The quality of the relationship is not influenced by the age of the patient.

These results demonstrate that the surveyed hospitals are generally fair in their treatment of patients. In conclusion, we commend the professionalism of healthcare personnel and their commitment to respecting human dignity.

Keywords: Inferential statistics; statistical test.

# Introduction Générale

Selon l'Organisation Mondiale de la Santé : « L'évaluation de la qualité des soins est une démarche qui permet de garantir à chaque patient des actes diagnostiques et thérapeutiques assurant le meilleur résultat en termes de santé, conformément à l'état actuel de la science médicale, au meilleur coût pour le meilleur résultat, au moindre risque iatrogène[[1]](#footnote-1) et pour sa plus grande satisfaction en termes de procédures, de résultats et de contacts humains à l'intérieur du système de soins »[[2]](#footnote-2).

La satisfaction du patient est donc une composante essentielle de l'évaluation de la qualité des soins. La mesure de cette satisfaction s'inscrit dans un contexte global du positionnement de l'usager dans l'organisation du système de santé et la montée des démarches qualités dans toutes les entreprises et services publics.

Dans cette optique, des données ont donc été collectées dans certains hôpitaux de la région parisienne auprès de 534 patients. L’objectif principal de cette collecte était de vérifier la satisfaction des patients au sujet de la qualité de la relation et de la quantité d’information reçues du personnel soignant.

Ainsi, au cœur de notre réflexion réside la question centrale de savoir si ***les usagers des hôpitaux de la région parisienne son globalement satisfait du personnel de santé***. Cette interrogation constitue le fil conducteur de notre étude.

La réponse à cette problématique nécessite de répondre aux questions suivantes :

* 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 ?
* Le score de relation est-il significativement corrélé à l’âge ?
* 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 ?

Pour mener à bien ce travail, nous poserons des hypothèses (Hypothèse nulle et Alternative) pour chaque question, comme l’impose la méthodologie de la statistique inférentielle.

Ainsi, afin de ressourdre efficacement cette étude, nous l’avons érigé essentiellement en deux parties, d’une part, prétraitement et analyse descriptive du jeu de donnée et d’autres parts, la statistique inférentielle proprement dite.

# 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 cas 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[[3]](#footnote-3)**.

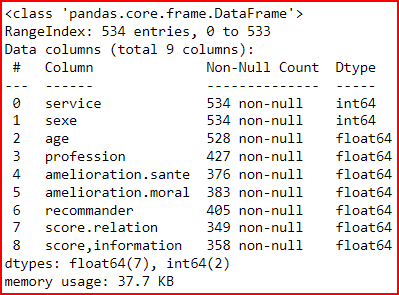
Elles portent sur l’évaluation de la qualité de relation et la quantité d’information reçue par le patient lors de son séjour à l’hôpital. Les patients ont été enquêtés dans plusieurs hôpitaux de la région parisienne sur 9 variables récapitulées ci-dessous.

Figure 1: Informations initiales du jeu de données

*Source : Réalisé par l’auteur avec Python*

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*

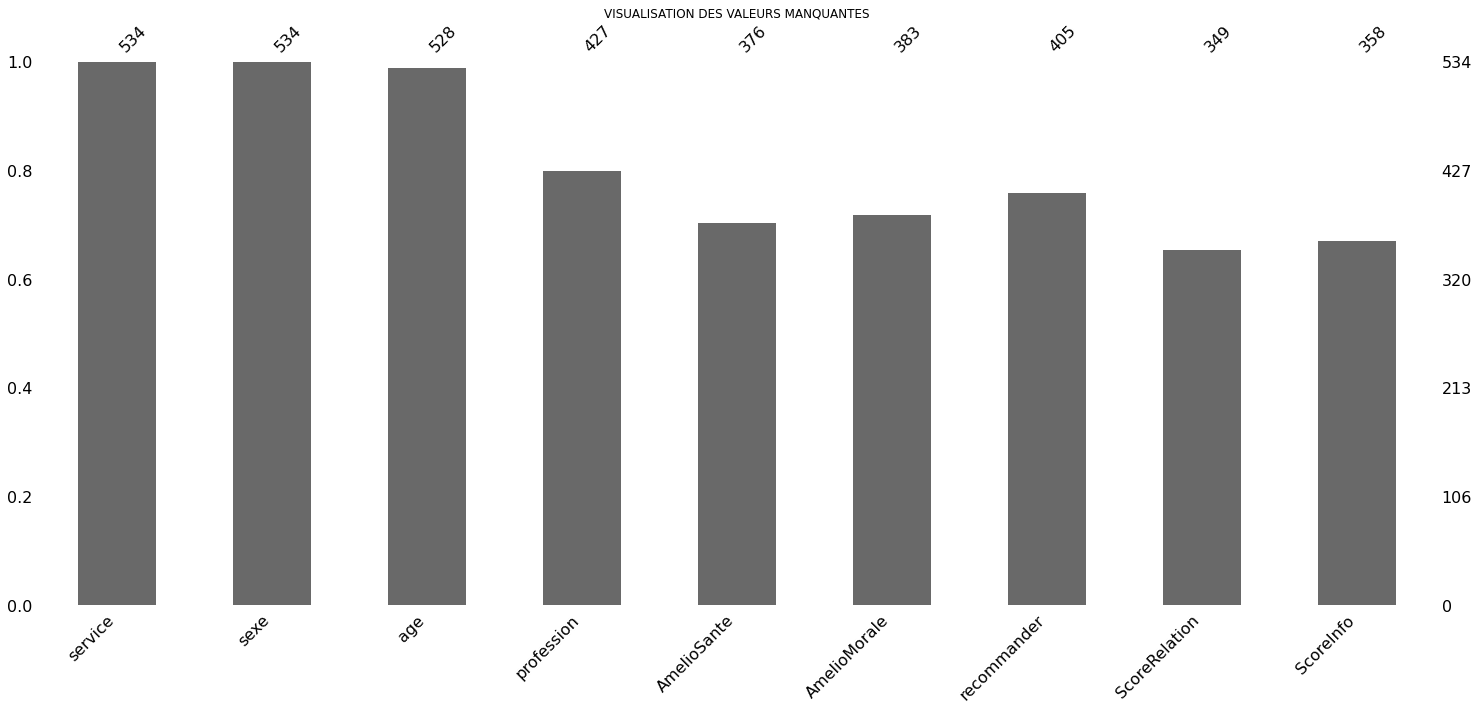
En résumé, la lecture des informations de bases de nos données nous montre des anomalies telles que : l’existence de valeurs manquantes et un mauvais typage de certaines variables. Dans la suite de notre analyse, nous allons traiter les valeurs manquantes, les valeurs aberrantes et recorder certaines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

Les tailles des six dernières variables montrent la présence de valeurs manquantes dans celles-ci. Le tableau ci-dessous récapitule le nombre de NaN par colonne.

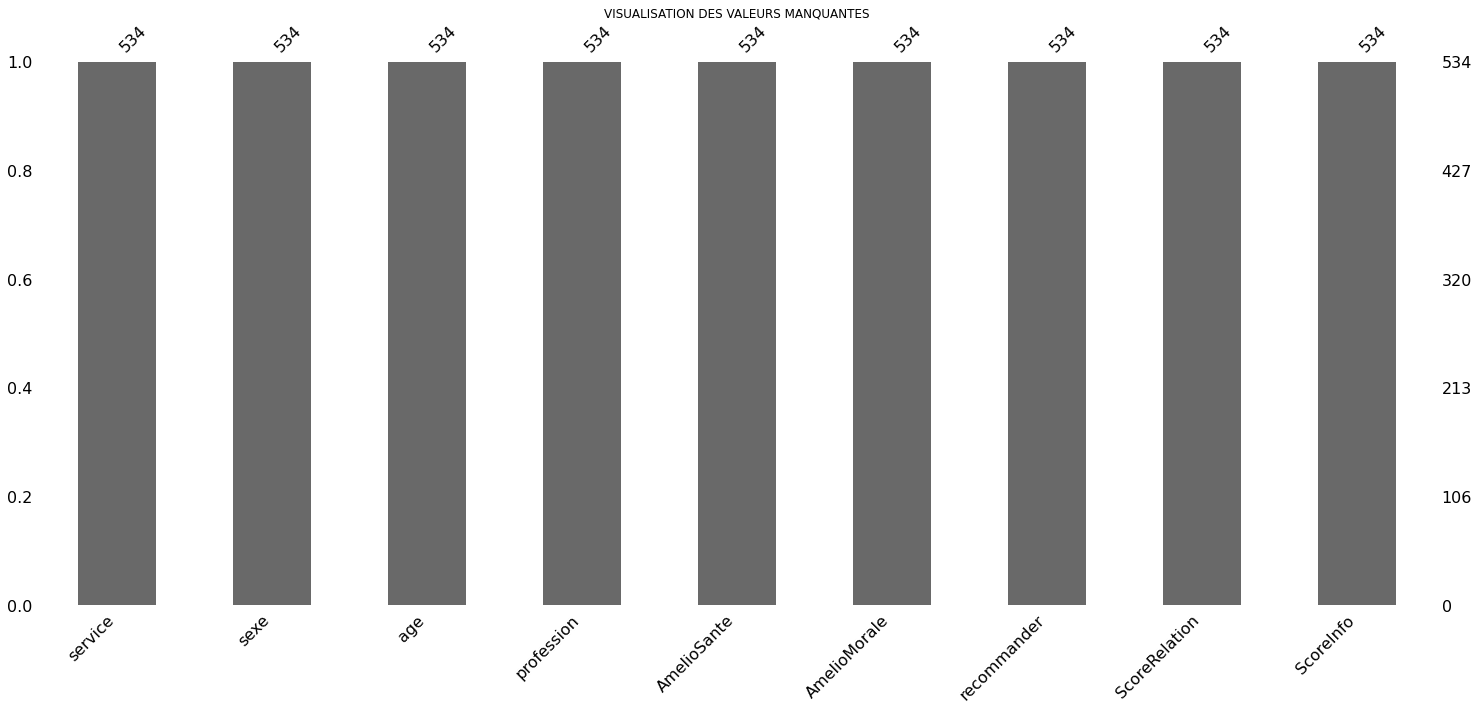
Tableau 2: Tableau des données manquantes par variables

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

*Source : Réalisé par l’auteur*

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.

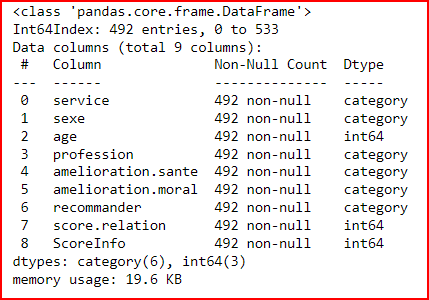
Les doublons peuvent provenir d’une collecte d’informations identiques provenant de différentes sources, d’erreurs humaines, ou encore de données ajoutées au lieu d’être mises à jour.

*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 pouvons constater 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.

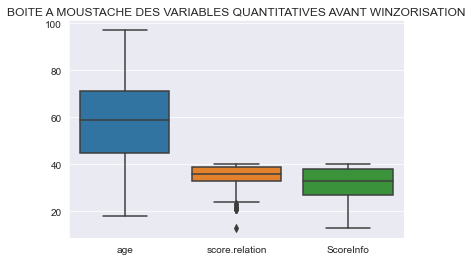
La détection des valeurs aberrantes ou des anomalies est l’un des problèmes fondamentaux de l’exploration des données. L’expansion émergente et continue des données nous fait repenser la façon dont nous abordons les anomalies et les cas d’utilisation qui peuvent être construits en examinant ces anomalies.

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)

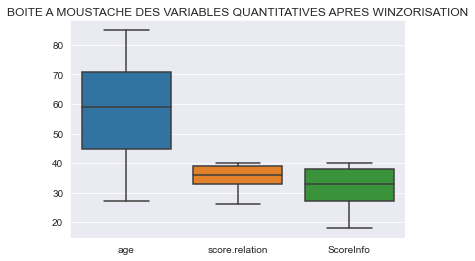
Figure 5: Boîte à moustache avant Winzorisation

*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[[4]](#footnote-4) consiste à remplacer les valeurs extrêmes par des valeurs moins extrêmes afin de réduire leur impact sur les analyses statistiques.

La figure ci-après 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 été découvertes au niveau de la variable **score.relation** et ont été ramenées à la borne inférieur.

Ce travail préalable a 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

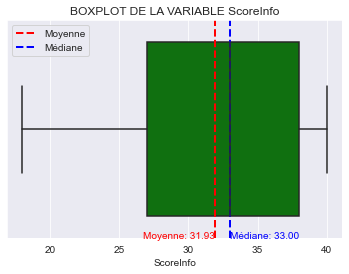
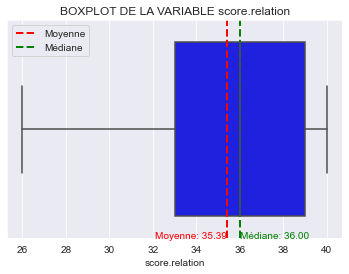
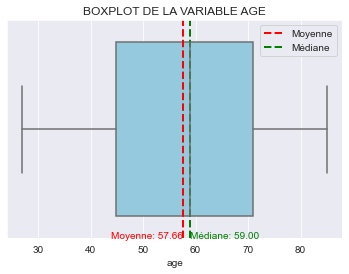
Selon Bernard PY, dans son livre *Statistique descriptive, nouvelle méthode pour bien comprendre et réussir (éditions Economica)* : « La statistique [descriptive] est un ensemble de méthodes permettant de décrire et d'analyser, de façon quantifiée, des phénomènes repérés par des éléments nombreux, de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

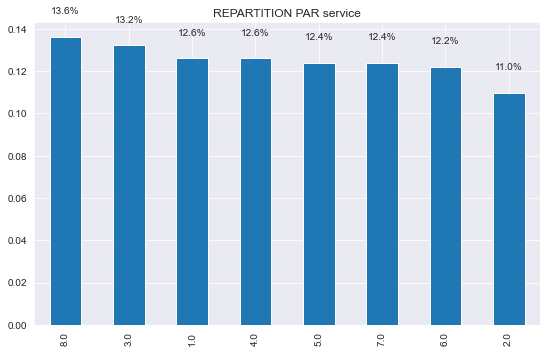
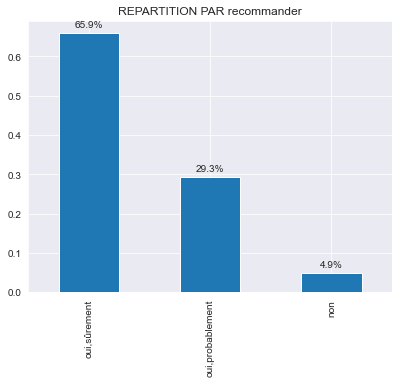
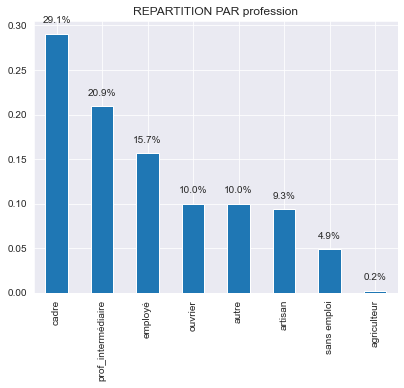
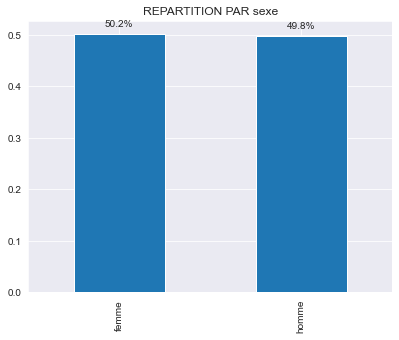
Le tableau ci-dessus vient conforter l’analyse graphique. Il nous permet d’affirmer **qu’en moyenne les patients sont âgés d’environ 58 ans avec un écart type de 17 ans**, cependant, l’âge minimal et modal est de 27 ans.

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*

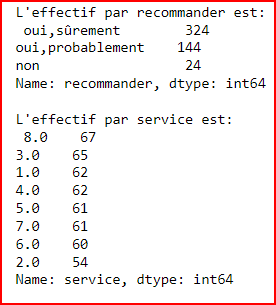
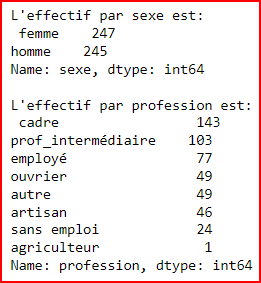
Les graphiques ci-dessus met en évidence que 50,2% des enquêtés sont de sexe masculin. Pratiquement **66% des patients recommandent les services** par lesquels ils sont passés durant le séjour à l’hôpital, seulement 5% environ ne recommandent pas.

*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

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 |

# Introduction

L'étude présentée s'inscrit dans le cadre du test de comparaison de deux populations indépendantes, une méthodologie statistique puissante visant à évaluer les différences significatives entre deux groupes distincts au sein d'une population. Notre analyse se focalise sur la relation entre le personnel soignant des hôpitaux de la région parisienne et les patients, en explorant les nuances de cette interaction en fonction du genre. Le choix de ce test vise à fournir des insights précieux sur la dynamique des relations entre le personnel soignant et les patients masculins et féminins, tout en permettant une évaluation approfondie des facteurs contribuant à la satisfaction globale des patients.

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

Cette question nous emmène à vérifier si les avis des patients sont différents selon qu’ils soient de sexe masculin ou de sexe féminin. En d’autres termes, les avis de recommandation dépendent-t-ils du genre de la personne qui recommande ?

Il s’agira donc de tester s’il y a égalité significative de proportion entre deux groupes, homme et femme qui recommandent les services par lesquels ils sont passés.

* **Formulation des hypothèses**

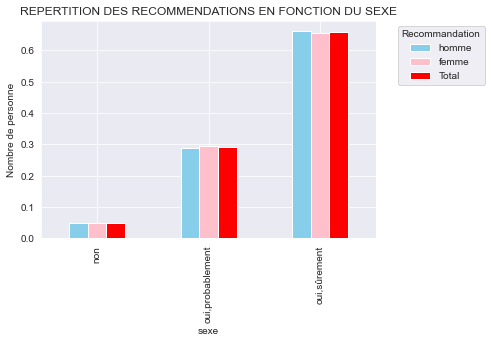
Les hypothèses statistiques sont formulées comme suit :

: la proportion de femmes ayant recommandée le service par lequel elles sont passées est la même que celle des hommes.

: la proportion de femme ayant recommandée le service par lequel elles sont passées n’est pas la même que celle des hommes.

La vérification de ces hypothèses se fait sous le logiciel Python.

* **Vérification des hypothèses**

Cette étape consiste à calculer des indicateurs et faire des graphiques pour les deux sous-groupes afin de faire le test statistique qui permet de trancher.

*Source : Réalisé par l’auteur sous Python*

Figure 10: Diagramme groupé de la repartition des recommandations par sexe

Ce graphique nous montre clairement que les proportions de recommandation chez les hommes et chez les femmes ne varient pas grandement. Les bars des diagrammes ont pratiquement les mêmes tailles quelques soient les genres.

Tableau 4: Tableau de contingence des effectifs empiriques des hommes et des femmes

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Homme | Femme | Total | Homme | Femme | Total |
| Effectif empirique | | | | **Effectif théorique** | | |
| non | 12 | 12 | **24** | 11.95 | 12.04 | **24** |
| oui, probablement | 71 | 73 | **144** | 71.70 | 72.29 | **144** |
| oui, sûrement | 162 | 162 | **324** | 161.34 | 162.65 | **324** |
| Total | **245** | **247** | **492** | **245** | **247** | **492** |

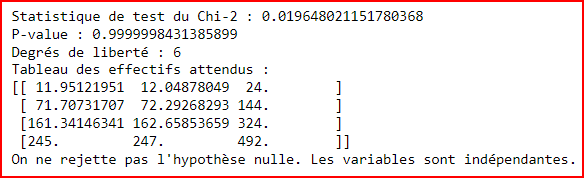
*Source : Réalisé par l’auteur avec Python*

Les tableaux ci-dessus peuvent nous permettre d’affirmer que toutes les valeurs théoriques sont supérieures à 5 et qu’il n’y a pas grande différence entre effectif observé dans l’échantillon et l’effectif calculé du test de Khy-deux. Nous soupçonnons donc une indépendance des réponses des femmes de celles des hommes.

* **Décision**

Suite à l’implémentation du test de Khy-deux, il ressort que les deux populations sont statistiquement indépendantes.

Figure 11: Résultat du test de Khy-deux



*Source : Réalisé par l’auteur sous Python*

* **Conclusion**

Le test de Khy-deux nous permet de conclure que nous acceptons l’hypothèse nulle qui stipule que **la proportion de femme qui recommande le service où elle a séjourné n’est pas statistiquement différente de celle des hommes**. En d’autres termes, les recommandations ne varient pas en fonction du genre des patients.

Cela voudrait dire que les patients apprécient les services du personnel soignant de ces différents hôpitaux sont globalement corrects.

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

La question suggère que nous souhaitons déterminer s'il existe une différence significative dans les scores de relation entre deux groupes, à savoir les hommes et les femmes. En d'autres termes, le score relatif à la qualité de la relation avec le personnel soignant lors du séjour est-il significativement différent d'un sexe à l'autre ?

Cela peut être interprété comme une comparaison des moyennes des scores de relation entre ces deux groupes.

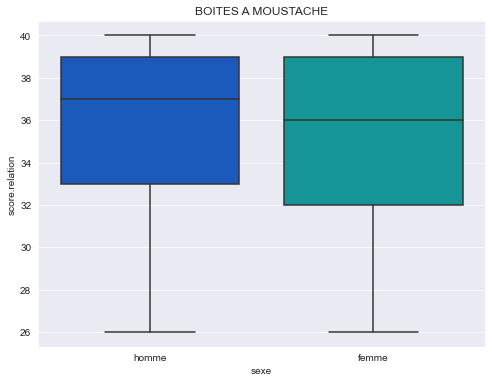
* **Formulation des hypothèses**

Les hypothèses statistiques sont formulées comme suit :

: la moyenne du score relation chez les hommes est statistiquement égale à celle des femmes.

: la moyenne du score relation chez les hommes est statistiquement différente de celle des femmes.

La vérification de ces hypothèses se fait sous le logiciel Python.

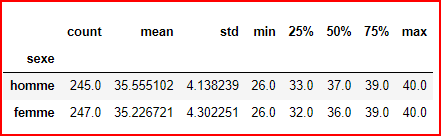
* **Vérification des hypothèses**
* **Etape 1 :** analyse graphique des sous-populations

*Source : Réalisé par l’auteur sous Python*

Figure 12: Boite à Moustache des sous-groupes

Les formes des boîtes à moustache des scores relations chez les hommes et chez les femmes ne semblent pas varier grandement même si la moyenne chez les femmes est un peu plus basse que celle des hommes.

La suite de notre analyse nous permettra de clarifier ce fait.

* **Etape 2 :** Statistiques de base des sous populations

*Source : Réalisé par l’auteur sous Python*

Figure 13: Résumé statistiques du score relation par genre

Ce tableau présente une grande similitude entre les deux sous populations, les caractéristiques de tendance centrale et les effectifs sont très voisines.

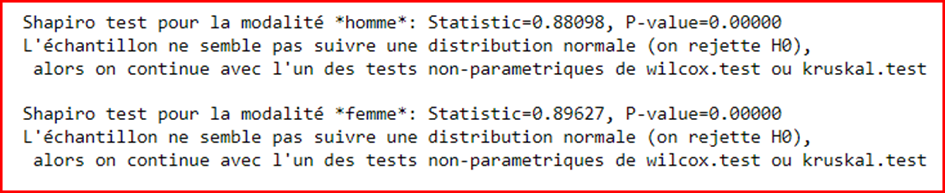
* **Etape 3 :** Test de normalité dans les sous populations

*Source : Réalisé par l’auteur sous Python*

Figure 14: Histogramme du score relation par genre

Les formes des histogrammes ci-dessus ne sont pas semblables à la forme typique en cloche de la distribution gaussienne. Elles sont plus étalées à droite, par conséquent, nos deux sous populations ne semblent pas suivre la loi normale.

Nous utiliserons le test de Shapiro-Wilk pour vérifier cette assertion.



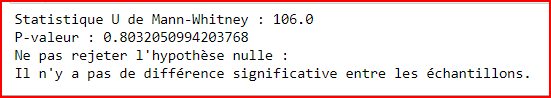
*Source : Réalisé par l’auteur sous Python*

Figure 15: Résultats du test de Shapiro

Le test de Shapiro Wilk confirme **la non-normalité des distributions des deux modalités** : homme et femme.

Le test d’égalité des proportions se fera donc grâce au test non-paramétrique des rangs de Wilcoxon ou test U de Mann-Whitney.

* **Décision**

L’implémentation du test de Wilcoxon nous donne le résultat ci-dessous :

*Source : Réalisé par l’auteur sous Python*

Figure 16: Résultat du test de Wilcoxon

Nous constatons clairement que le test des rangs de Wilcoxon autorise à accepter l’hypothèse nulle avec une P-value de 0.80 qui est largement supérieur à 5%.

* **Conclusion du test**

À un niveau de signification de 5%, les résultats suggèrent qu'il n'y a pas de différence significative entre les hommes et les femmes en ce qui concerne le score moyen de relation.

En d'autres termes, sur la base de l'échantillon de données que nous avons, il n'y a pas suffisamment de preuves pour affirmer que la moyenne du score de relation avec le personnel soignant est significativement différente entre les hommes et les femmes.

Par conséquent, nous pouvons affirmer qu’**en moyenne la relation entre le personnel soignant et les patients est le même quel que soit le sexe**.

# Conclusion

En somme, l’attitude du personnel soignant au regard des patients, qu’ils soient de sexe masculin ou féminin, est très acceptable sur la base de notre échantillon.

Ces résultats prouvent que le personnel soignant joue bien son rôle en ce qui concerne la qualité de la relation avec les patients. Les patients étant satisfaits, recommandent alors les services par lesquels ils sont passés.

# TEST DE COMPARAISON DE PLUSIEURS POPULATIONS INDEPENDANTES

## Introduction

L'analyse de notre étude se déploie également dans le cadre du test de comparaison de plusieurs populations indépendantes, une méthodologie statistique éclairante visant à évaluer les distinctions significatives entre divers groupes au sein d'une population donnée. Notre examen s'est concentré sur le personnel soignant des hôpitaux de la région parisienne, en évaluant son attitude envers les patients en fonction de critères tels que l'âge, la profession, et le genre. Ce test de comparaison de plusieurs populations indépendantes offre une perspective riche et nuancée sur les variations potentielles dans l'interaction entre le personnel soignant et les patients, permettant ainsi une compréhension approfondie des dynamiques interpersonnelles au sein du contexte hospitalier.

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

La présente étude vise à déterminer si la qualité de l'information perçue varie en fonction de la profession des individus. L'objectif est d'analyser les différences éventuelles dans la perception de la qualité de l'information entre divers groupes professionnels.

Nous examinerons si les patients de domaines différents ont des évaluations similaires de la qualité de l'information reçue ou s'il existe des disparités significatives. Cette analyse contribuera à éclairer la relation potentielle entre la profession des individus et leur perception de la qualité de l'information reçue du personnel soignant.

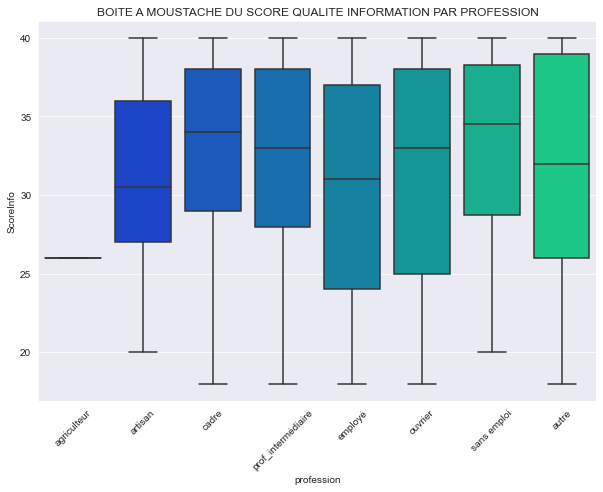
* **Formulation des hypothèses**

Les hypothèses statistiques sont formulées comme suit :

: la qualité de l’information reçue est la même selon les professions

: la qualité de l’information reçue est significativement differente selon les professions

La vérification de ces hypothèses se fait sous le logiciel Python.

* **Vérification des hypothèses**
* **Etape 1 :** analyse graphique des sous-populations

*Source : Réalisé par l’auteur sous Python*

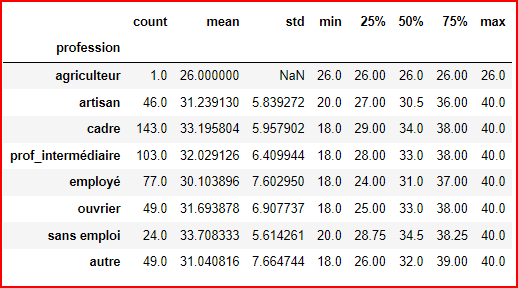
Figure 17: Boîte à Moustache du ScoreInfo par profession

Les formes des boîtes à moustache des scores informations (notation de la qualité de l’information) par profession ne semblent pas varier grandement même si les moyennes représentées sur chaque boîte varient selon les différentes professions.

La suite de notre analyse nous permettra de clarifier ce fait.

* **Etape 2 :** Statistiques de base des sous populations

Figure 18: Résumé statistiques du ScoreInfo par profession



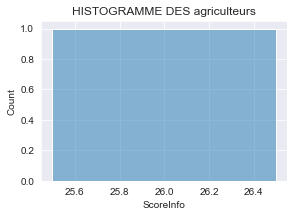
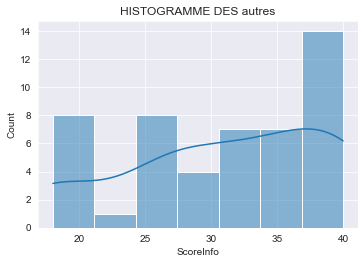
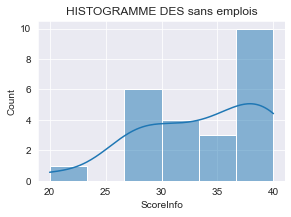
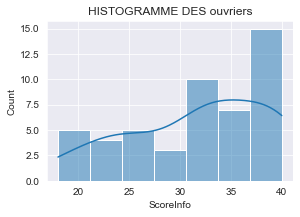
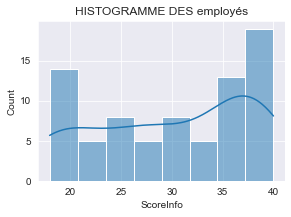
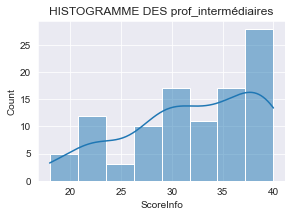
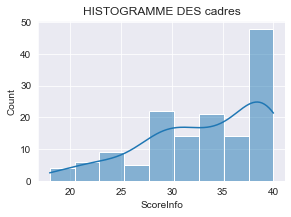
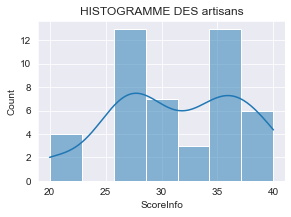
*Source : Réalisé par l’auteur sous Python*

Ce tableau présente une grande similitude entre ces sous populations, les caractéristiques de tendance centrale et les effectifs sont très voisines.

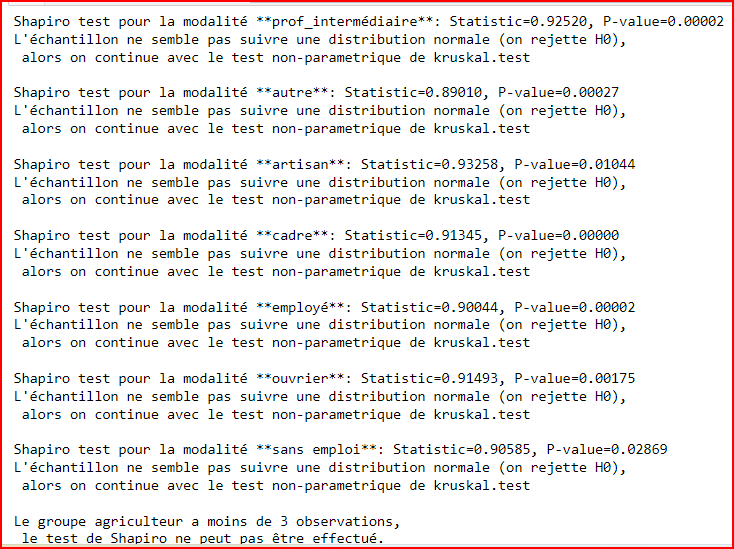
* **Etape 3 :** Test de normalité dans les sous populations

Les formes des histogrammes ci-après ne sont pas semblables à la forme typique en cloche de la distribution gaussienne. Elles sont de formes assez irrégulières, par conséquent, nos sous populations ne semblent pas suivre la loi normale. Nous utiliserons le test de Shapiro-Wilk pour vérifier cette assertion.

Figure 19: Histogramme du score relation par profession



*Source : Réalisé par l’auteur sous Python*



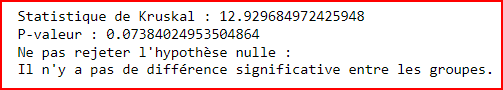
*Source : Réalisé par l’auteur sous Python*

Figure 20: Résultats du test de Shapiro

Le résultat du test de Shapiro Wilk ci-dessus confirme **la non-normalité des distributions des six modalités des professions**.

Le test d’égalité des proportions se fera donc grâce au test non-paramétrique de Kruskal Wallis.

* **Décision**

L’implémentation du test de Kruskal Wallis nous donne le résultat ci-dessous :

*Source : Réalisé par l’auteur sous Python*

Figure 21: Résultat du test de Kruskal Walli

Nous constatons clairement que le test de Kruskal-Wallis nous autorise à accepter l’hypothèse nulle avec une P-value de 0.07 qui est largement supérieur à 5%.

* **Conclusion**

La p-value résultante du test de Kruskal-Wallis est supérieure au niveau de signification choisi (0.07 > 0.05 dans notre cas), cela suggère que nous ne pouvons pas rejeter l'hypothèse nulle.

Ce qui signifie que, sur la base du test global, il n'y a pas suffisamment de preuves pour conclure qu'il y a une différence significative entre les groupes.

Par conséquent, **la qualité de l’information reçue est la même selon les professions**. En d’autres termes, le personnel soignant ne fait pas de distinction selon la profession quant à l’information donnée aux patients.

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

Cette étude cherche à déterminer si la qualité des relations avec le personnel soignant varie en fonction du service dans lequel le patient a été accueilli.

Il s’agira donc, de tester les éventuelles différences dans la qualité des relations entre les patients et le personnel soignant au sein de différents services.

*Nous chercherons à établir si le service d'accueil du patient a une incidence significative sur la perception de la qualité des relations avec le personnel soignant*. Cette analyse contribuera à approfondir notre compréhension des facteurs influençant la qualité des interactions entre les patients et le personnel soignant.

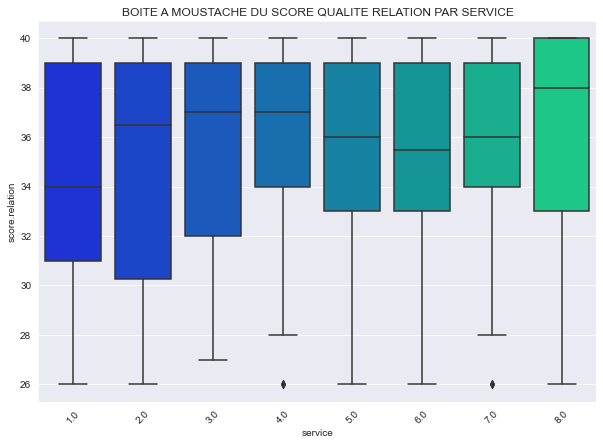
* **Formulation des hypothèses**

Les hypothèses statistiques sont formulées comme suit :

: la qualité de la relation avec les patients est la même quel que soit le service.

: la qualité de la relation avec les patients est significativement différente selon le service.

La vérification de ces hypothèses se fait sous le logiciel Python.

* **Vérification des hypothèses**
* **Etape 1 :** analyse graphique des sous-populations

*Source : Réalisé par l’auteur sous Python*

Figure 22: Boîte moustache du score relation par service

Les formes des boîtes à moustache des scores relations (notation de la qualité de la relation avec le personnel soignant) par service ne semblent pas varier grandement même si les moyennes représentées sur chaque boîte varient selon les différents services.

La suite de notre analyse nous permettra de mieux comprendre cet état de fait.

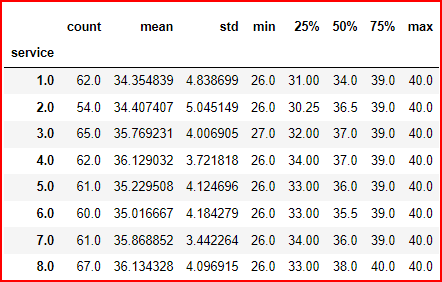
* **Etape 2 :** Statistiques de base des sous-populations

Figure 23: Résumé statistiques du score.relation par service

*Source : Réalisé par l’auteur sous Python*

Ce tableau présente une grande similitude entre ces sous populations, les caractéristiques de tendance centrale comme la moyenne du score qui tourne autour de 34/40 avec un écart type de score de 4 et les effectifs sont très voisines, en moyenne 60 individus par service.

* **Etape 3 :** Test de normalité dans les sous populations

Les formes des histogrammes ci-après ne sont pas semblables à la forme typique en cloche de la distribution gaussienne. Elles sont de formes assez irrégulières, par conséquent, nos sous populations ne semblent pas suivre la loi normale. Cette observation est confirmée par le test de Shapiro Wilk.

*Source : Réalisé par l’auteur sous Python*

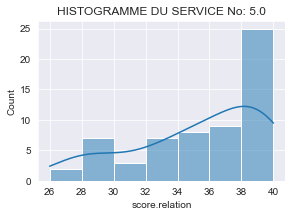
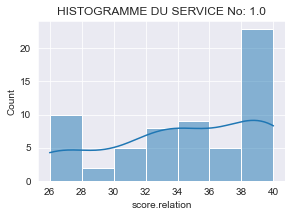
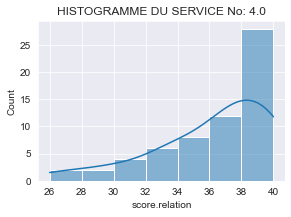
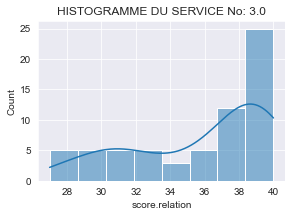


Figure 24: Histogrammes du ScoreInfo par service frequenté

Figure 25: Résultats du test de Shapiro

Le résultat du test de Normalité de Shapiro confirme l’observation obtenue lors de l’analyse des histogrammes, ces variables ne suivent pas loi normale. Nous allons donc faire le test de Kruskal-Wallis.

* **Décision**

L’implémentation du test de Kruskal Wallis nous donne le résultat ci-dessous :

Figure 26: Résultat du test de Kruskal Wallis

*Source : Réalisé par l’auteur sous Python*

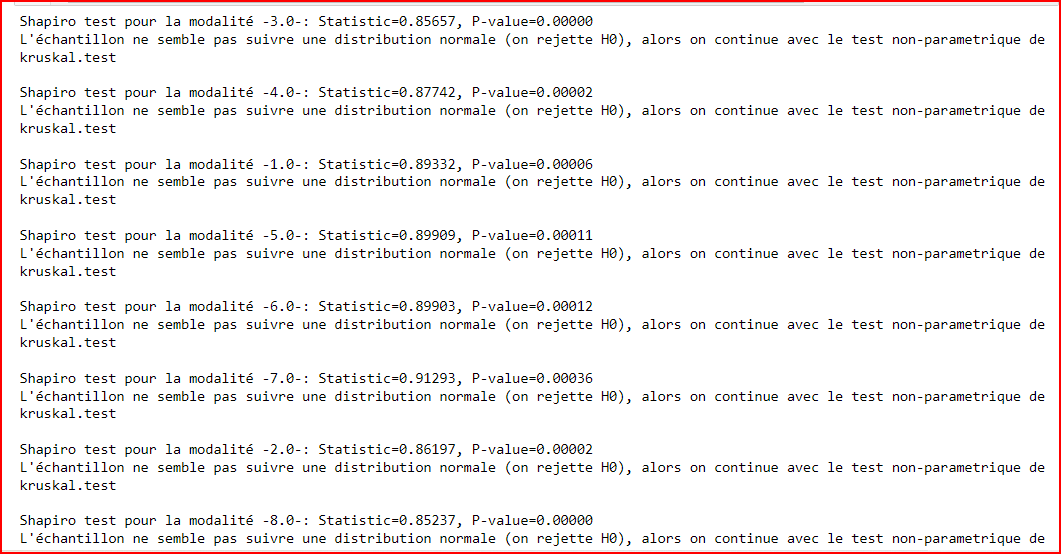


Figure 27: Résultats des tests de Shapiro Wilk

*Source : Réalisé par l’auteur sous Python*

* **Décision**

*Source : Réalisé par l’auteur sous Python*

Nous constatons clairement que le test de Kruskal-Wallis nous autorise à accepter l’hypothèse nulle avec une P-value de 0.23 qui est largement supérieur à 5%.

* **Conclusion**

Pour conclure, nous acceptons l’hypothèse nulle qui stipule **qu’il n'y a pas de différence significative entre la qualité des relations avec le personnel soignant en fonction du service ayant accueilli le patient**.

Ce signifierait que les différents services entretiennent de bonnes relations avec les patients qu’ils accueillent.

# Conclusion

Pour conclure, **la qualité de l’information reçue est la même selon les professions**. En d’autres termes, le personnel soignant ne fait pas de distinction selon la profession quant à l’information donnée aux patients.

De plus, il n'y a pas de différence significative entre la qualité des relations avec le personnel soignant en fonction du service ayant accueilli le patient.

# TEST DE LIAISON OU D’INDEPENDANCE :

# Le score de relation est-il significativement corrélé à l’âge ?

## Introduction

Cette dernière question vise à déterminer l'existence d'une corrélation significative entre le score de relation et l'âge des patients.

En d’autres termes, nous testerons si les notes attribuées par les patients sont corrélées à leur âge ?

Pour ce faire, nous allons utiliser le coefficient de corrélation de Pearson ou de Spearman pour évaluer la force et la direction de la relation entre ces deux variables.

* **Formulation des hypothèses**

Les hypothèses statistiques sont formulées comme suit :

: le score de la relation attribué par les patients n’est pas significativement corrélé à leur âge

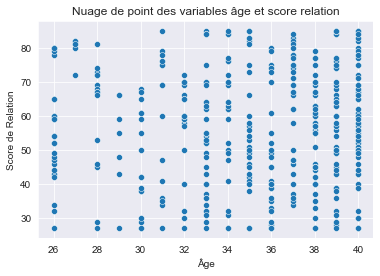
: le score de la relation attribué par les patients est significativement corrélé à leur âge

La vérification de ces hypothèses se fait sous le logiciel Python.

* **Vérification des hypothèses**
* **Etape 1 :** analyse graphique des sous-populations

Figure 28: Nuage de point de la variable age et du score.relation

*Source : Réalisé par l’auteur sous Python*

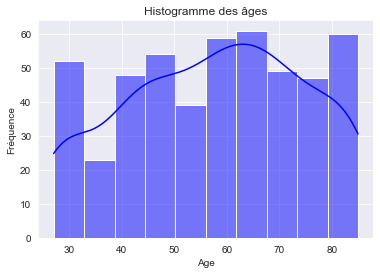
****

La répartition des points sur le graphique ci-dessus ne montre pas une corrélation explicite entre les deux variables.

La suite de notre analyse nous permettra de mieux comprendre cet état de fait.

* **Etape 2 :** Test de normalité dans les sous populations

La forme de l’histogramme n’est pas semblable à la forme typique en cloche de la distribution gaussienne. Elle a une forme assez irrégulière, par conséquent, la variable *age* ne semble pas suivre la loi normale.



*Source : Réalisé par l’auteur sous Python*

Figure 29: Histogramme de la variable âge

Les QQplot des deux variables nous montrent aussi que les variables ne suivent pas une loi normale. En voici les figures :

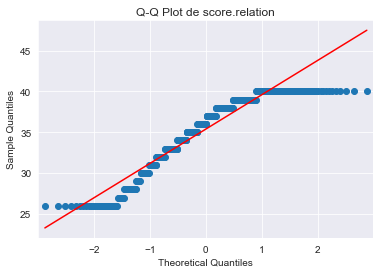
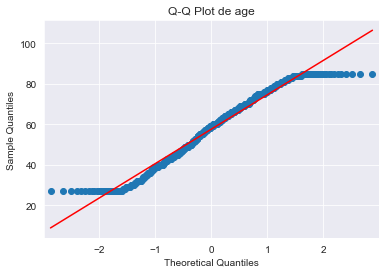


Figure 30: qq-plot des variables age et score.relation

*Source : Réalisé par l’auteur sous Python*

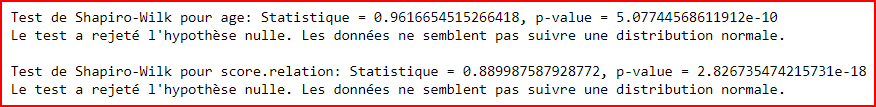
****Les de Shapiro Wilk nous permet de conclure à la non-normalité des variables **age** et **score.relation.**

Figure 31: résultats du test de Shapiro wilk sur les variables age et score.relation

*Source : Réalisé par l’auteur sous Python*

* **Décision**

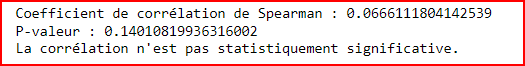
Les coefficients de corrélation de Pearson et de Spearman donnent respectivement la valeur 0,05 et 0,067 cette valeur est très faible. Cela traduit un manque de corrélation entre les deux variables.

Figure 32: Resultat du test de Spearman

*Source : Réalisé par l’auteur sous Python*

Le test sur ce coefficient nous montre que l’âge est très faiblement corrélé aux sore.relation attribué par les patients.

## Conclusion

Au terme de cette partie, nous retenons que nous acceptons l’hypothèse nulle de la non corrélation significative des âges et la valeur du score de la relation avec le personnel soignant.

# Conclusion générale

Notre étude visait à déterminer si les patients qui fréquentent les hôpitaux de la région parisienne sont satisfaits de la qualité de la relation et des informations reçues du personnel médical.

L’analyse descriptive des données recueillies auprès de 534 patients des hôpitaux de cette région a révélé que ceux-cisont des cadres d’entreprise de 57 ans en moyenne. Ces derniers ont apprécié la relation avec le personnel soignant en attribuant un score moyen de 35/40. De plus, les patients ont jugé avoir reçu le maximum d’information durant leur passage à l’hôpital.

Les résultats obtenus de cette première analyse seront testés grâce à des tests statistiques adaptés à cet effet.

La seconde partie portant sur la statistique inférentielle des questions issues de l’objectif principal montre que **le personnel soignant des hôpitaux de la région parisienne ne fait de discrimination selon le genre, l’âge et la profession des patients**. Les usagers sont globalement satisfaits, car plus de 66% des enquêtés *recommandent sûrement* les services par lesquels ils sont passés.

Les résultats sont très satisfaisants, ils montrent que le personnel de ces hôpitaux joue correctement leurs rôles. Cependant, nous ne pouvons nous empêcher de faire des recommandations afin d’améliorer davantage ces services.

Alors, sur la base de ces résultats, nous recommandons :

* *Renforcement de la Communication avec les Patients* : améliorer la communication entre le personnel soignant et les patients contribue à une meilleure compréhension des soins et à une relation de confiance, ce qui peut avoir un impact significatif sur la satisfaction des patients.
* *Formation Continue pour le Personnel Soignant :* la formation continue garantit que le personnel dispose des compétences nécessaires pour offrir des soins de haute qualité. Cela inclut des aspects tels que la communication, la sensibilité culturelle et la gestion du stress.
* *Évaluation Continue de la Satisfaction des Patients* : la rétroaction régulière des patients permet de comprendre les besoins changeants et d'apporter des ajustements continus pour répondre aux attentes des patients et maintenir un haut niveau de satisfaction.
* *Analyse Approfondie des Recommandations* : comprendre en détail les raisons derrière les recommandations des patients permet d'identifier les forces et les faiblesses spécifiques du système de soins, orientant ainsi les efforts d'amélioration.
* *Sensibilisation à la Non-Discrimination* : bien que les résultats indiquent une absence de discrimination, maintenir une sensibilisation constante et des formations régulières peut prévenir d'éventuelles situations de discrimination, garantissant un traitement équitable pour tous les patients.

Ces recommandations visent à améliorer la qualité des soins, à renforcer la confiance des patients et à favoriser un environnement de soins de santé positif dans les hôpitaux de la région parisienne.

# Annexe

|  |
| --- |
| **Importation des packages** |
| #chargement des packages necessaire  import matplotlib.pyplot as plt  import pandas as pd  import missingno as msno |
|  |
| # Importation des donnees  hopital = pd.read\_csv("satisfaction\_hopital.csv", sep= ";") |
| **Prétraitement des données** |
| #transformation de la variable 'score,information' en 'ScoreInfo'  df.rename(columns={'score,information': 'ScoreInfo'}, inplace=True)  import seaborn as sns  import matplotlib.pyplot as plt  # Création de la heatmap pour visualiser les valeurs manquantes  sns.heatmap(df.isnull(), cbar=False, cmap='YlGnBu')  plt.title('Visualisation des Valeurs Manquantes dans le Jeu de Données')  plt.show()  # Utilisation de la fonction pour déterminer K  variables\_a\_imputer = ['age', 'profession', 'amelioration.sante', 'amelioration.moral', 'recommander', 'score.relation', 'ScoreInfo']  trouver\_k\_optimal(df, variables\_a\_imputer)  import pandas as pd  from sklearn.impute import KNNImputer  # Liste des variables à imputer  variables\_a\_imputer = ['service', 'sexe', 'age', 'profession', 'amelioration.sante', 'amelioration.moral', 'recommander', 'score.relation', 'ScoreInfo']  # Imputation des valeurs manquantes avec K-plus proche voisin  imputer = KNNImputer(n\_neighbors=1) # Vous pouvez ajuster le nombre de voisins selon vos besoins  df\_imputed = pd.DataFrame(imputer.fit\_transform(df[variables\_a\_imputer]), columns=variables\_a\_imputer)  # df\_imputed contient maintenant les valeurs imputées  #Suppression des doublons  df\_imputed=df\_imputed.drop\_duplicates()  #visualisation et traitement des valeurs aberrantes  import seaborn as sns  ## Afficher la boîte à moustaches  sns.set\_style('darkgrid')  sns.boxplot(data= df\_imputed)  plt.title('BOITE A MOUSTACHE DES VARIABLES QUANTITATIVES AVANT WINZORISATION')  plt.show()  #transformations des variable  df\_propre['sexe'] = df\_propre['sexe'].replace({0.0: 'homme', 1.0: 'femme'})  df\_propre['recommander'] = df\_propre['recommander'].replace({0.0: 'non', 1.0: 'oui,probablement', 2.0: 'oui,sûrement'})  df\_propre['AmelioSante'] = df\_propre['amelioration.sante'].replace({0.0: 'aggrravée', 1.0: 'aucune amélioration', 2.0: 'peu améliorée', 3.0: 'nettement améliorée'})  df\_propre['AmelioMorale'] = df\_propre['amelioration.moral'].replace({0.0: 'aggrravé', 1.0: 'aucune amélioration', 2.0: 'peu amélioré', 3.0: 'nettement amélioré'})  df\_propre['profession'] = df\_propre['profession'].replace({1.0: 'agriculteur', 2.0: 'artisan', 3.0: 'cadre', 4.0 : 'prof\_intermédiaire', 5.0: 'employé', 6.0: 'ouvrier', 7.0: 'sans emploi', 8.0: 'autre'}) |
| **Analyse univariée** |
| # Créer un boxplot avec seaborn  sns.boxplot(data=df\_propre, x='age', color='skyblue')  # Ajouter la moyenne en tant que ligne rouge  mean\_value = np.mean(df\_propre['age'])  plt.axvline(mean\_value, color='red', linestyle='dashed', linewidth=2, label='Moyenne')  # Afficher les valeurs caractéristiques  plt.text(mean\_value, 0.5, f'Moyenne: {mean\_value:.2f}', color='red', verticalalignment='baseline', horizontalalignment='right')  # Ajouter la médiane  median\_value = np.median(df\_propre['age'])  plt.axvline(median\_value, color='green', linestyle='dashed', linewidth=2, label='Médiane')  # Afficher les valeurs caractéristiques  plt.text(median\_value, 0.5, f'Médiane: {median\_value:.2f}', color='green', verticalalignment='baseline', horizontalalignment='left')  # # Ajouter d'autres caractéristiques si nécessaire (médiane, etc.)  # min\_value = np.min(df\_propre['age'])  # plt.axvline(min\_value, color='blue', linestyle='dashed', linewidth=2, label='Minimum')  # Personnaliser le graphique  plt.title('BOXPLOT DE LA VARIABLE AGE')  plt.legend()  # Afficher le graphique  plt.show()# 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()  # Isolement des variables qualitatives  quali=[] # Initialisation de quali en tant que liste vide  for col in df\_propre.columns:  if df\_propre[col].dtype == 'category':  quali.append(col)    print(quali) |
| **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()  # Effectif des variables  colQuali = ['sexe', 'profession', 'recommander', 'service']  for var in colQuali:  print(f"L'effectif par {var} est:\n", round(df\_propre[var].value\_counts(), 2),"\n") |
| **Statistique inférentielle** |
| ## 1- \*La proportion de sujets recommandant le service dans lequel ils sont passés est-il le même quel que soit le genre ?\*  # Test d'Independance pour les variables catégorielles  # Effectuer le test du Chi-2  chi2\_stat, p\_value, dof, expected = chi2\_contingency(TabEff1)  # Imprimer les résultats  print(f"Statistique de test du Chi-2 : {chi2\_stat}")  print(f"P-value : {p\_value}")  print(f"Degrés de liberté : {dof}")  print("Tableau des effectifs attendus :")  print(expected)  # Interpréter les résultats  alpha = 0.05  if p\_value < alpha:  print("On rejette l'hypothèse nulle. Les variables ne sont pas indépendantes.")  else:  print("On ne rejette pas l'hypothèse nulle. Les variables sont indépendantes.") """    ### 2- \*La moyenne du score de relation est-il significativement différent chez les hommes et chez les femmes ?\*  data = df\_propre[['age', 'score.relation']].dropna()  # Effectuer le test de corrélation de Spearman  correlation\_coefficient, p\_value = spearmanr(data['age'], data['score.relation'])  # Afficher les résultats  print(f"Coefficient de corrélation de Spearman : {correlation\_coefficient}")  print(f"P-valeur : {p\_value}")  # Interprétation du résultat  alpha = 0.05  if p\_value < alpha:  print("La corrélation est statistiquement significative.")  else:  print("La corrélation n'est pas statistiquement significative.")  ### 4) Peut-on affirmer que la qualité de l’information reçue est la même quelle que soit la profession?  # Effectuer le test de Wilcoxon sur les rangs  statistique, p\_valeur = kruskal(\*[df\_propre['ScoreInfo'][df\_propre['profession'] == prof] for prof in df\_propre['profession'].unique()])  # Afficher les résultats  print("Statistique de Kruskal :", statistique)  print("P-valeur :", p\_valeur)  # Vérifier si le résultat est statistiquement significatif  alpha = 0.05  if p\_valeur < alpha:  print("Rejeter l'hypothèse nulle : Il y a une différence significative entre les groupes.")  else:  print("Ne pas rejeter l'hypothèse nulle :\nIl n'y a pas de différence significative entre les groupes.")  ### 5) La qualité des relations avec le personnel soignant est-il fonction du service ayant accueilli le patient?  # Test de Kruskal-Wallis  result\_kruskal = kruskal(\*[df\_propre['score.relation'][df\_propre['service'] == prof] for prof in df\_propre['service'].unique()])  # Interprétation  alpha = 0.05  if result\_kruskal.pvalue < alpha:  print(f"la P-value est : {round(result\_kruskal.pvalue, 2)}, alors on rejette H0,\nil y a une différence significative entre au moins deux groupes.")  else:  print(f"la P-value est : {round(result\_kruskal.pvalue, 2)}, alors on ne rejette pas H0,\nil n'y a pas de différence significative entre les groupes.") |
|  |

# Bibliographie et Webographie

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

ChatGPT

DataScientest. *DataScientest.* 10 Octobre 2023. https://datascientest.com/data-quality-erreurs (accès le Janvier 17, 2023).

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](#_Toc156779497)

[Sigles 2](#_Toc156779498)

[Liste des illustrations 3](#_Toc156779499)

[Listes des figures 3](#_Toc156779500)

[Liste des tableaux 3](#_Toc156779501)

[Sommaire 4](#_Toc156779502)

[Résumé 6](#_Toc156779503)

[Abstract 7](#_Toc156779504)

[Introduction Générale 9](#_Toc156779505)

[PARTIE I :](file:///C:\Users\THINKPAD\OneDrive\Bureau\Mini-Projets-INSSEDS\MiniProjet3\MiniProjet3_SavanéMory_2023.docx#_Toc156779506) [PRETRAITEMENT ET STATISTIQUE DESCRIPTIVE 11](file:///C:\Users\THINKPAD\OneDrive\Bureau\Mini-Projets-INSSEDS\MiniProjet3\MiniProjet3_SavanéMory_2023.docx#_Toc156779507)

[I. PRETRAITEMENT DU JEU DE DONNEES 12](#_Toc156779508)

[Introduction 12](#_Toc156779509)

[1. Présentation des données 12](#_Toc156779510)

[2. Traitement des valeurs manquantes et doublons 14](#_Toc156779511)

[3. Traitement des valeurs aberrantes 16](#_Toc156779512)

[Conclusion 18](#_Toc156779513)

[II. STATISTIQUE DESCRIPTIVE UNIVARIÉE 19](#_Toc156779514)

[Introduction 19](#_Toc156779515)

[1. Analyse univariée des variables quantitatives 19](#_Toc156779516)

[2. Analyse univariée des variables qualitatives 21](#_Toc156779517)

[Conclusion 22](#_Toc156779518)

[PARTIE II :](file:///C:\Users\THINKPAD\OneDrive\Bureau\Mini-Projets-INSSEDS\MiniProjet3\MiniProjet3_SavanéMory_2023.docx#_Toc156779519) [STATISTIQUE INFERENTIELLE 23](file:///C:\Users\THINKPAD\OneDrive\Bureau\Mini-Projets-INSSEDS\MiniProjet3\MiniProjet3_SavanéMory_2023.docx#_Toc156779520)

[I- TEST DE COMPARAISON DE DEUX POPULATIONS INDEPENDANTES 24](#_Toc156779521)

[Introduction 24](#_Toc156779522)

[1- La proportion de sujets recommandant le service dans lequel ils sont passés est-il le même quel que soit le genre ? 24](#_Toc156779523)

[2- La moyenne du score relation est-il significativement différent chez les hommes et chez les femmes ? 26](#_Toc156779524)

[Conclusion 29](#_Toc156779525)

[II- TEST DE COMPARAISON DE PLUSIEURS POPULATIONS INDEPENDANTES 30](#_Toc156779526)

[Introduction 30](#_Toc156779527)

[1- Peut-on affirmer que la qualité de l’information reçue est la même quelle que soit la profession ? 30](#_Toc156779528)

[2- La qualité des relations avec le personnel soignant est-il fonction du service ayant accueilli le patient ? 35](#_Toc156779529)

[Conclusion 38](#_Toc156779530)

[III- TEST DE LIAISON OU D’INDEPENDANCE :](#_Toc156779531)[Le score de relation est-il significativement corrélé à l’âge ? 39](#_Toc156779532)

[Introduction 39](#_Toc156779533)

[Conclusion générale 42](#_Toc156779534)

[Annexe 44](#_Toc156779535)

[Bibliographie et Webographie 50](#_Toc156779536)

1. Se dit d'un trouble, d’une maladie provoqués par un acte médical ou par les médicaments, même en l'absence d'erreur du médecin. (larousse.fr) [↑](#footnote-ref-1)
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6211818/#cit0001 [↑](#footnote-ref-2)
3. https://www.kaggle.com/datasets/samawel97/satisfaction-hospital/code [↑](#footnote-ref-3)
4. 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-4)