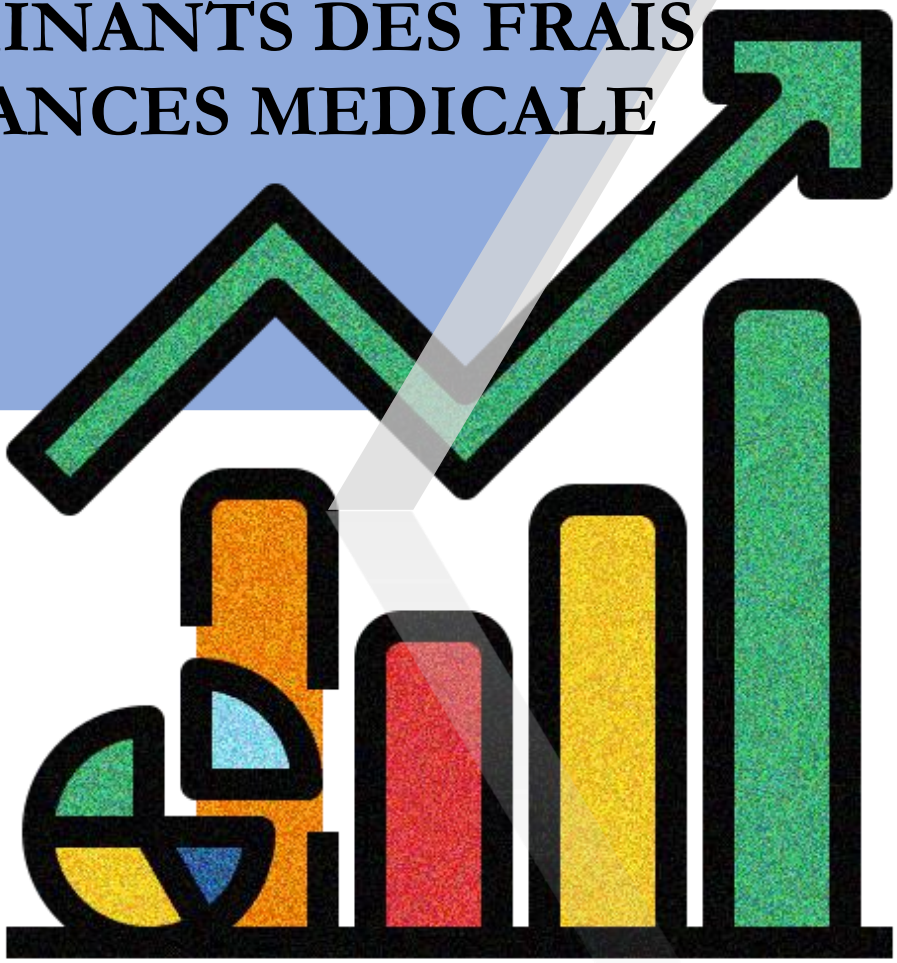
INSTITUT UNIVERSITAIRE
SAINT JEAN

Thème : LES DETERMINANTS DES FRAIS D'ASSURANCES MEDICALE



Projet de Statistique

Enseignant : Pr. NGUEFACK-TSAGUE Georges

Réalisé par : OWONA FOUDA Jean Edouard

Etudiant en Master 1 Option : « **Data science** »

Année Académique : 2022 - 2023

Les déterminants des frais d'assurance médicale

Table des matières

Liste des tableaux

Introduction

Le dictionnaire Larousse définit une assurance comme étant : « un contrat par lequel l'assureur s'engage à indemniser l'assuré, moyennant une prime ou une cotisation, de certains risques ou sinistres éventuels. Par extension, l'assurance est le service économique qui regroupe les activités de conception, de production et commercialisation de ce type de service. Ce secteur économique représente donc une mine de données qui sont générées par les souscripteurs en général et ceux du domaine de la santé en particulier. La préoccupation du futur Data Scientist est la mise en place d'une démarche de valorisation et d'exploitation à partir des données générées par des processus informatiques, humains ou hybrides. Ainsi il peut s'agir pour ce dernier de déterminer les facteurs qui influencent le coût total des frais d'assurance médicale pour un souscripteur. Telle est la tâche à laquelle nous nous attèlerons dans ce travail.

Ce travail est structuré en quatre (04) parties. La première est une brève description des données exploitées. La deuxième est relative à la définition de l'objectif général et des objectifs spécifiques. Le troisième et sans doute le plus important est la démarche proposée pour répondre à la question suivante : **quels sont les déterminants des frais d'assurance pour un souscripteur** ? En dernier lieu, nous discuterons les résultats obtenus. Le présent travail a été réalisé à l'aide du langage de programmation Python dans l'environnement Jupyter.

Les déterminants des frais d'assurance médicale

I. Description des données

Les données utilisées dans le cadre de ce travail proviennent du site www.kaggle.com , Le jeu de données « **Insurance Cost Prediction** » comporte 1338 individus et 7 variables. Ces dernièrues Les variables sont principalement de deux types à savoir : quantitatives (l'âge, l'indice de masse corporelle, le nombre d'enfants, le montant total des frais d'assurances) et qualitatives (le sexe, la région de résidence et la consommation de tabac).

Comme le montre le tableau ci-après, aucune des variables ne comporte de valeurs manquantes.

Tableau 1 : Descriptif des variables

Variable	Nature	Nombre de valeurs manquantes
Age	Quantitative	0
IMC		
Nombre d'enfants		0
Frais d'assurances		0
Sexe	Qualitative	0
Fumeur		0
Région		0

Source : nos traitements statistiques

II. Objectifs

II. A. Objectif général

On s'intéresse aux variables qui susceptibles d'influencer le coût global des frais d'assurance pour un souscripteur. L'objectif est donc de ressortir ces dernières

II. B. Objectifs spécifiques

L'atteinte de cet objectif, nécessite de réaliser les activités ci-après :

a. Effectuer des analyses descriptives univariées ;
b. Effectuer des analyses descriptives croisées à deux variables ;

c. Déterminer et tester l'existence d'éventuelles corrélation/liaison de toutes les autres variables avec la variable.

III. Résultats

III. A. Partite descriptive

Dans cette partie nous explorons chaque variable individuellement afin avoir une idée de sa distribution ou de la répartition de ses modalités

a. Variables quantitatives

Le tableau ci-après illustre les statistiques descriptives des variables quantitatives :

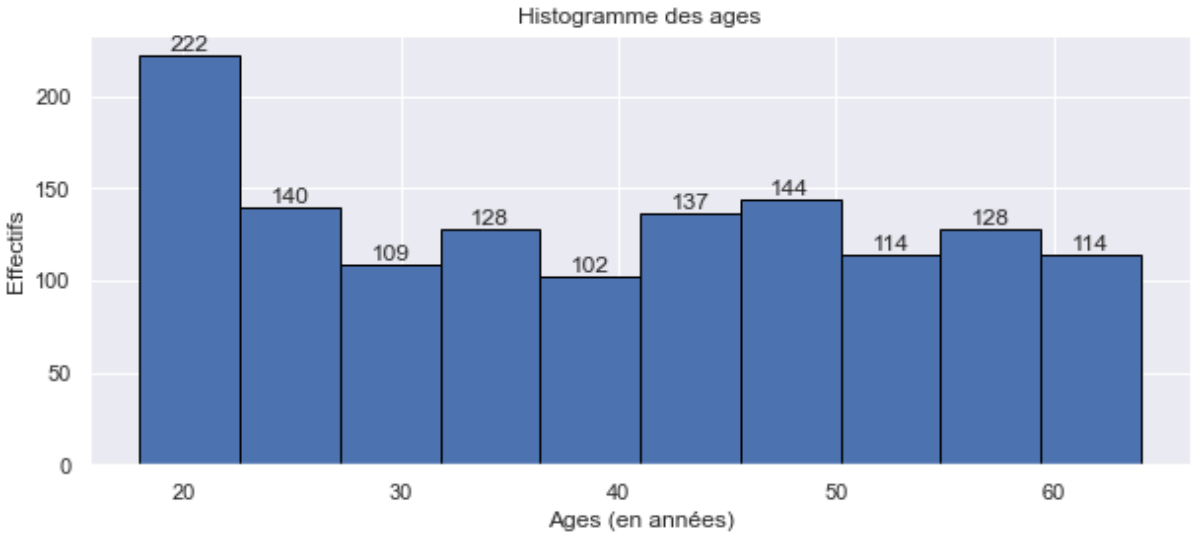
Tableau 2 : Récapitulatif des indicateurs des variables quantitatives

Indicateurs	Age (en années)	IMC (en Kg/m²)	Nombre d'enfants	Frais d'assurance (en $)
N			1338	
Moyenne ± ET	39,21 ± 14,05	30,66 ± 6,10	1,09 ± 1,21	13270,42 ± 12110,01
Min	18,00	15,96	0,00	1121,87
Max	27,00	26,30	0,00	4740,29
Q1	39,00	30,40	1,00	9382,03
Q2	51,00	34,69	2,00	16639,91
Max	64,00	53,13	5,00	63770,43

Source : nos traitements statistiques

Globalement, le nombre d'enfants des répondants et leurs frais d'assurance étaient les plus dispersés. Les figures ci-après données plus de détails sur la répartition de ces variables.

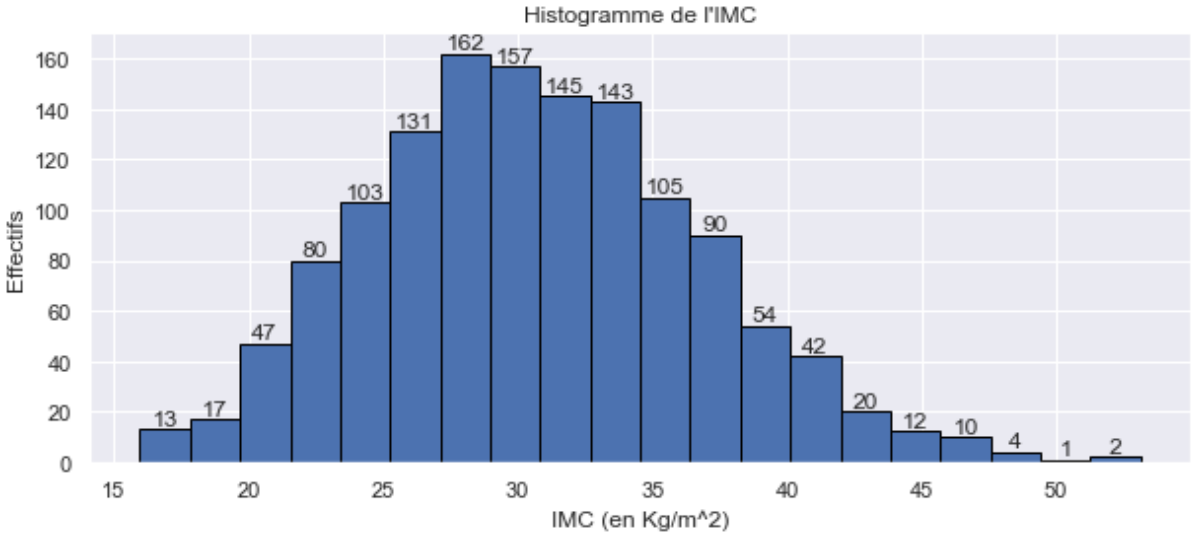
☞ Répartition des individus selon l'âge



Source : nos traitements statistiques

Les individus étaient âgés d'au plus 64 ans, ces derniers avaient en moyenne moins de 40 ans et 50% d'entre eux avaient moins de 51 ans. D'après le graphique ci-dessous, la majorité des individus sont dans la vingtaine.

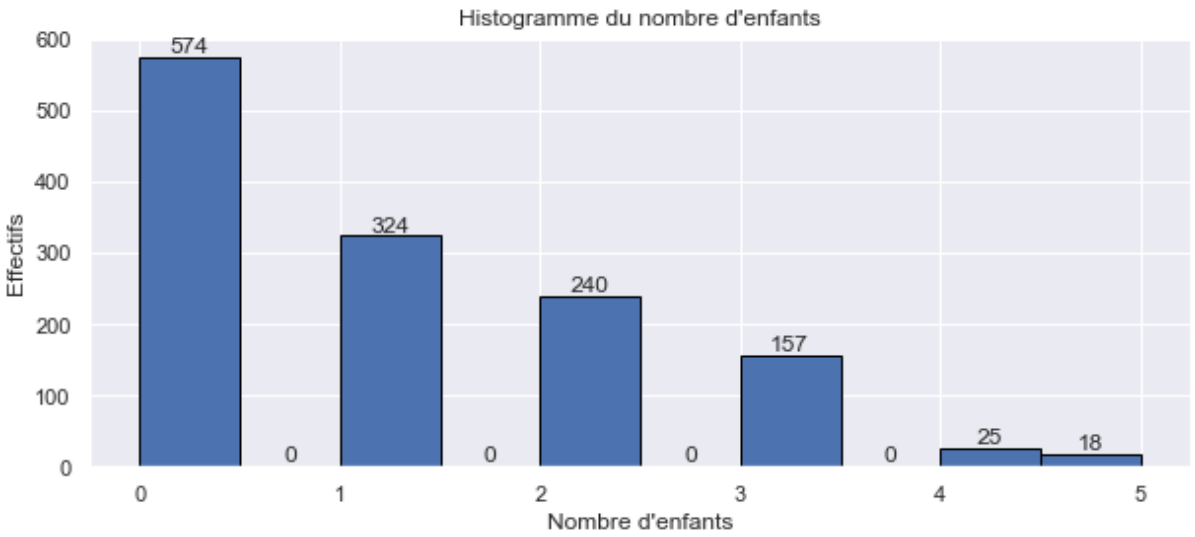
☞ Répartition des individus selon l'IMC



Source : nos traitements statistiques

S'agissant de l'IMC (Indice de Masse Corporelle), la moitié des individus avaient in indice supérieur à 30 kg/m² et donc étaient obèses (selon l'OMS (Organisation Mondiale de la Santé)). Ce graphique suggère une distribution qui semble normale.

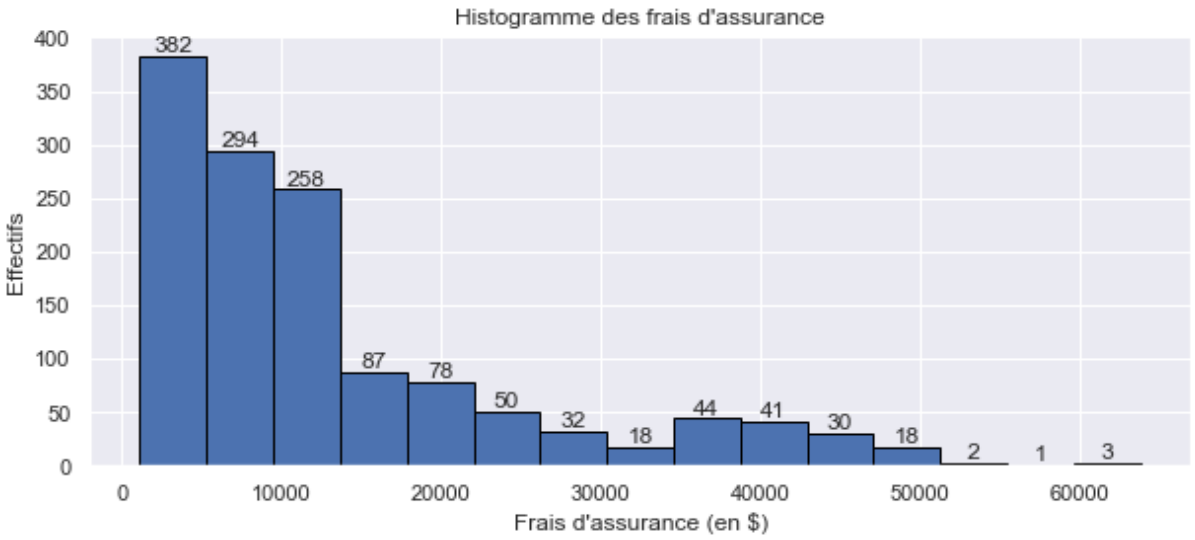
☞ Répartition des individus selon le nombre d'enfants



Source : nos traitements statistiques

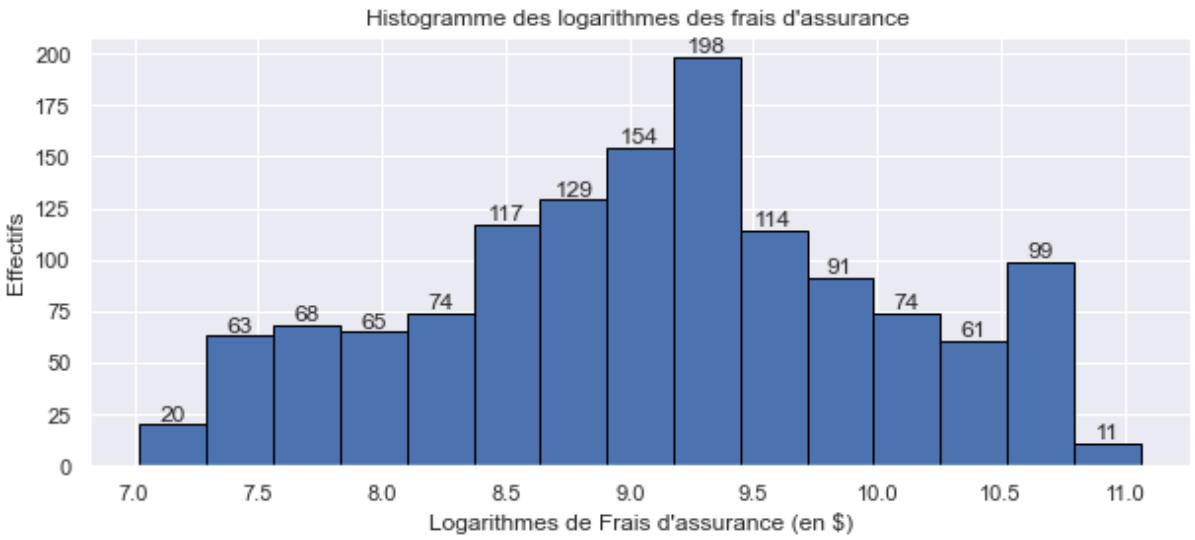
La plupart des individus n'avaient aucun enfant en charge.

☞ Répartition des individus selon le montant des frais d'assurances



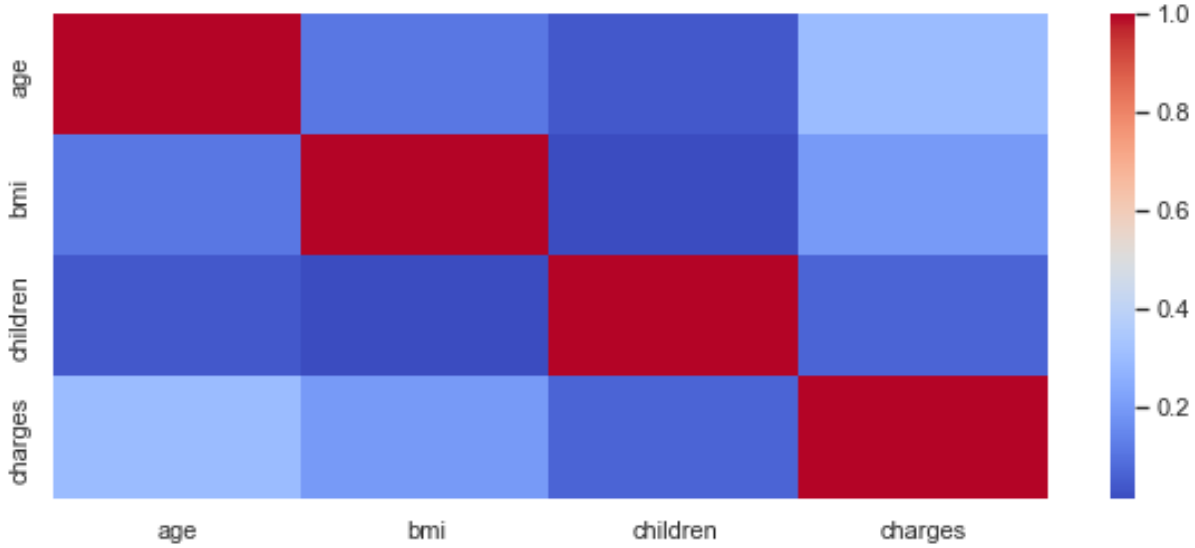
Source : nos traitements statistiques

Les individus avaient principalement des frais d'assurances inférieurs ou égaux à 15000 $. La distribution observée est fortement asymétrique et aplatie à gauche, il serait intéressant de calculer les logarithmes des frais d'assurance afin d'obtenir une distribution plus normalisée et réduire les outliers. Le graphique ci-après présenté la distribution obtenue. Cette variable représente la cible ou la variable dépendante.



Source : nos traitements statistiques

☞ Matrice de corrélation entre les variables quantitatives



Source : nos traitements statistiques

D'après la matrice ci-dessus, il n'existe pas de corrélation forte entre les différentes variables.

b. Variables qualitatives

Le tableau ci-après illustre les principaux indicateurs des variables quantitatives :

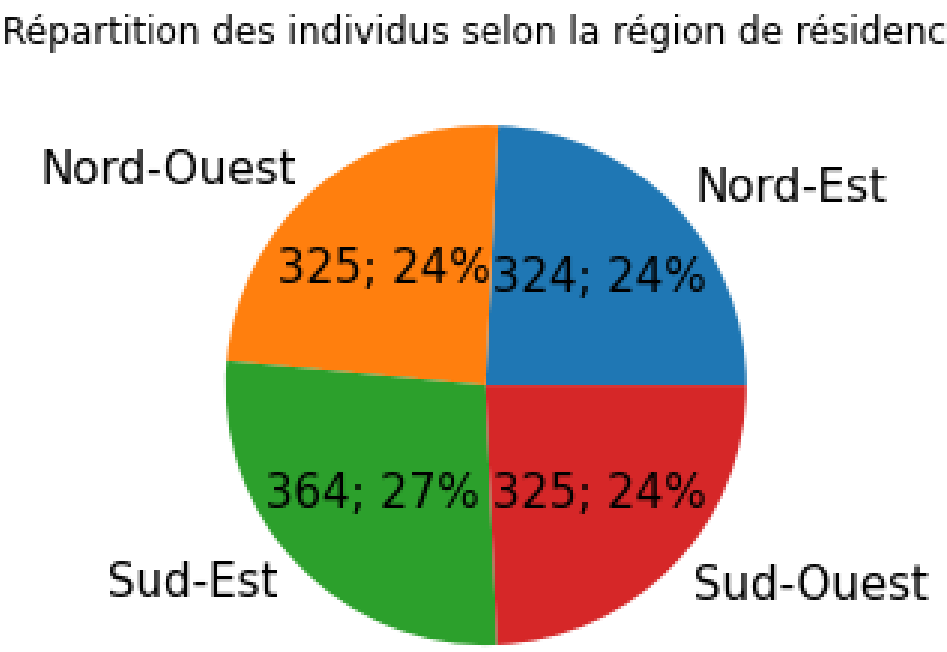
Tableau 3 : Les principaux indicateurs des variables qualitatives

Indicateurs	Sexe	Fumeur	Région
N		1338	
Modalités	2	2	4
Mode	Masculin	No	Sud-Est
Effectif du mode	676	1064	364

Source : nos traitements statistiques

Selon le sexe, la répartition était quasi égale entre les deux sexes. Et la sex-ratio valait 1,02

Les consommateurs de tabac représentaient 20% de l'ensemble des individus.



Source : nos traitements statistiques

À l'exception de la région « **Sud-Est** » qui regroupait le plus d'individus, la répartition dans les autres régions était égale.

III. B. Partie analytique

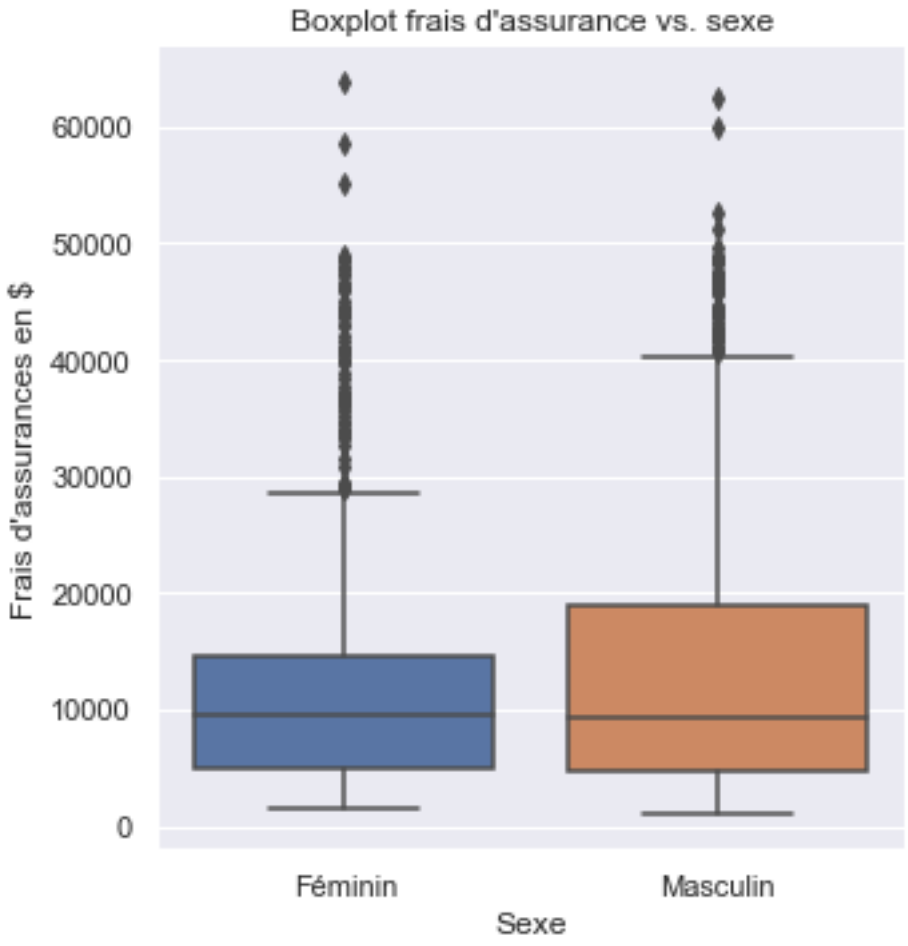
NB : Le seuil de significativité est fixé à 5%.

Cette partie regroupe les tests statistiques qui seront réalisés afin d'identifier les variables susceptibles d'influencer le coût global des frais d'assurance médicale pour un individu.

- ☞ Tests de normalité de la variable IMC

Le graphique de l'IMC à la page 4 suggérait une distribution normale, afin de vérifier cette hypothèse, nous avons réalisé un test de normalité et la **p-value (<0.01)**, de ce dernier ne permet pas de conclure que cette distribution est normale.

- ☞ Les frais d'assurance selon le sexe



Source : nos traitements statistiques

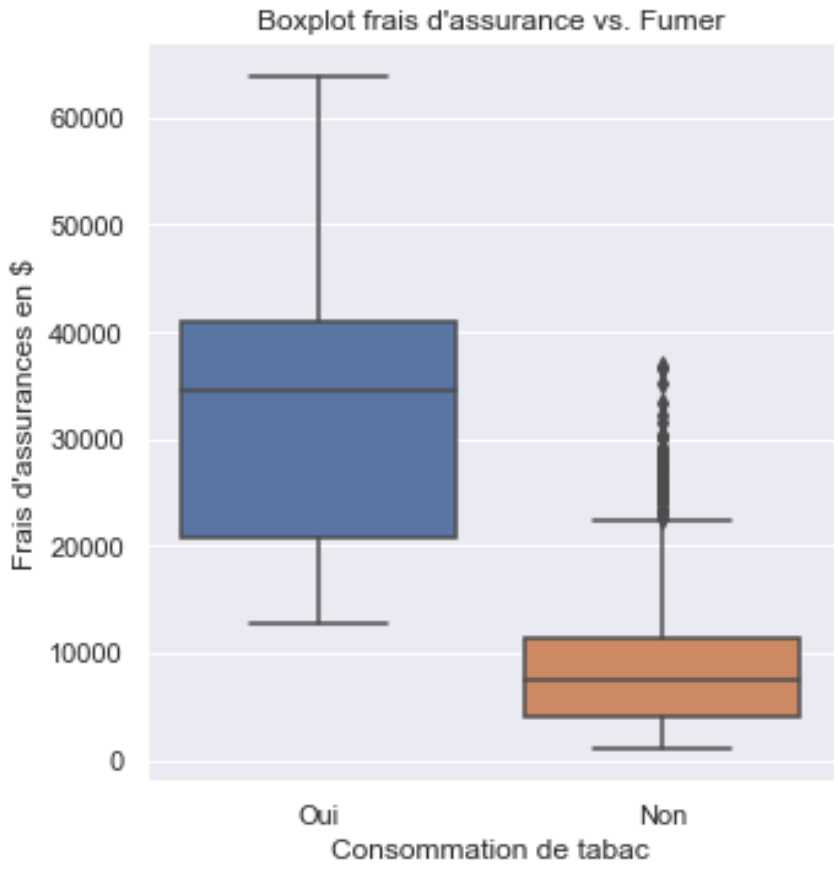
Les médianes des frais d'assurances sont quasi égales chez les hommes et chez les femmes, ces frais sont approximativement dans la même fourchette pour les deux sexes. Cependant ils ont tendance à être plus élevés chez les hommes. Le tableau ci-dessous résume les résultats du test de Student afin de comparer les moyennes des frais d'assurances entre les deux sexes.

Tableau 4 : Comparaison des moyennes des frais selon le sexe

T	DLL	IC (95%)	P-value	Conclusion
2.10	1313,36	[91.86, 2682.49]	***0,034***	Tout porte à croire que le sexe a une influence significative sur le coût des frais d'assurance

Source : nos traitements statistiques

☞ Les frais d'assurance selon la consommation de tabac



Source : nos traitements statistiques

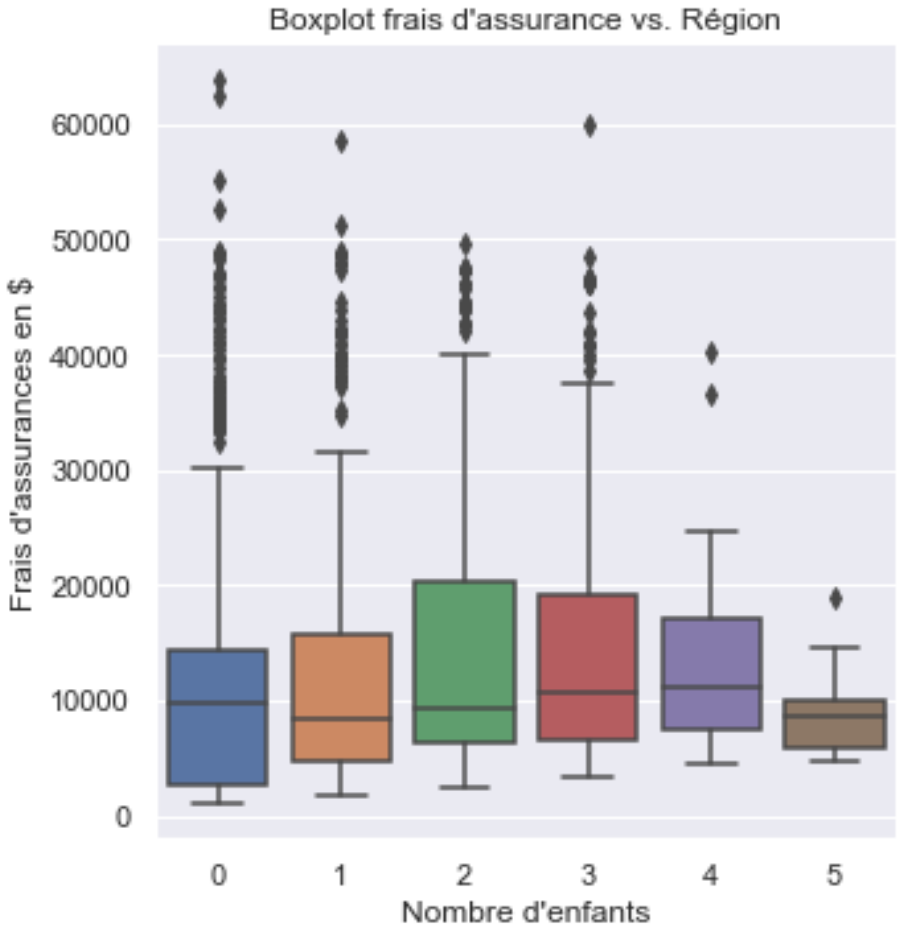
Les frais d'assurances chez les fumeurs sont très dispersés comparés à ceux des non-fumeurs chez qui la distribution est plus ou moins équilibrée. Il faut aussi noter la moitié des fumeurs dépensent plus du maximum de ce que dépensent les non-fumeurs. Il serait naturel de soupçonner une relation entre ces deux variables. Le tableau ci-dessous résume les résultats du test de Student afin de d'éclaircir ce soupçon.

Tableau 5 : Comparaison des moyennes des frais selon la consommation de tabac

T	DLL	IC (95%)	P-value	Conclusion
32,75	311,85	[22197.21, 25034.71]	***<0,01***	Tout porte à croire que consommer du tabac à une influence significative sur le coût des frais d'assurance

Source : nos traitements statistiques

- ☞ Les frais d'assurance selon le nombre d'enfants



Source : nos traitements statistiques

La dépense médiane individus ayant quatre (4) enfants est supérieure à celles des autres. Il faut aussi noter que les individus n'ayant aucun enfant ont une fourchette de dépense plus large que ceux ayant cinq (5) enfants, ceci est sans doute du au fait que ces deux groupes sont (respectivement) les plus (les moins) prépondérants. L'ANOVA réalisée entre ces deux variables permet de conclure que le nombre d'enfants a une influence significative sur le coût des frais d'assurance (**p-value = 0,005**).

- ☞ Nuage de points des frais d'assurance en fonction de l'âge

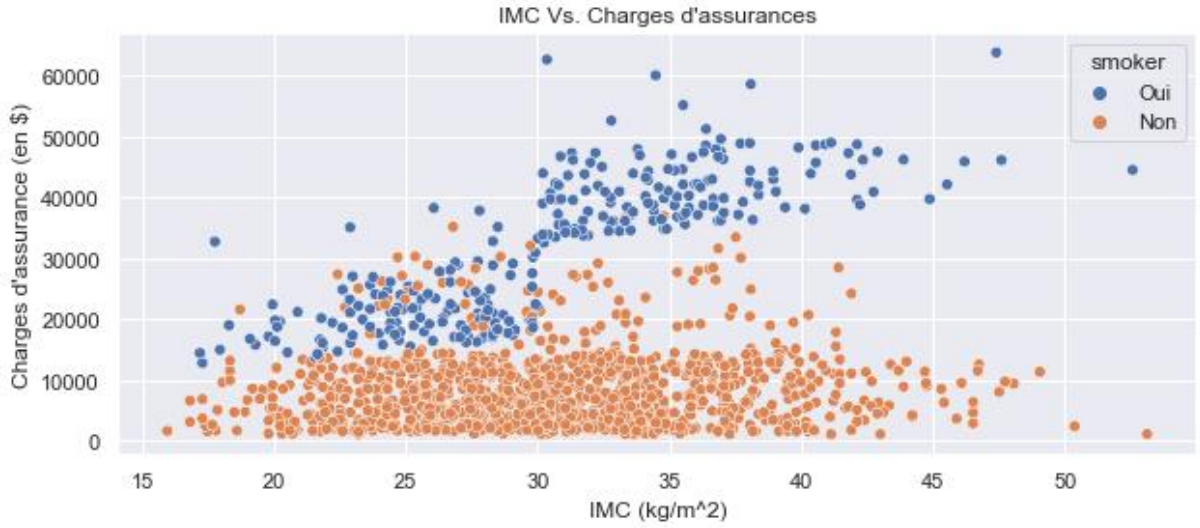
Source : nos traitements statistiques

Le graphique ci-dessus peut être subdivisé en trois (03) région du bas vers le haut. D'une part, la majorité des non-fumeurs se situent dans la première région et on des dépenses de moins de15 000 $, d'autre part les fumeurs sont majoritairement situés dans la troisième région et les plus dépensiers vont jusqu'à plus de 60 000 $. Globalement on peut observer une relation linéaire positive entre ces deux variables. Le tableau ci-dessous illustre les résultats du test du coefficients de corrélation.

N	r	IC (95%)	P-value	Conclusion
1338	0,30	[0.25, 0.35]	***<0,01***	La relation linéaire entre l'âge et les charges d'assurance est significative

Source : nos traitements statistiques

☞ Nuage de points des frais d'assurance en fonction de l'IMC



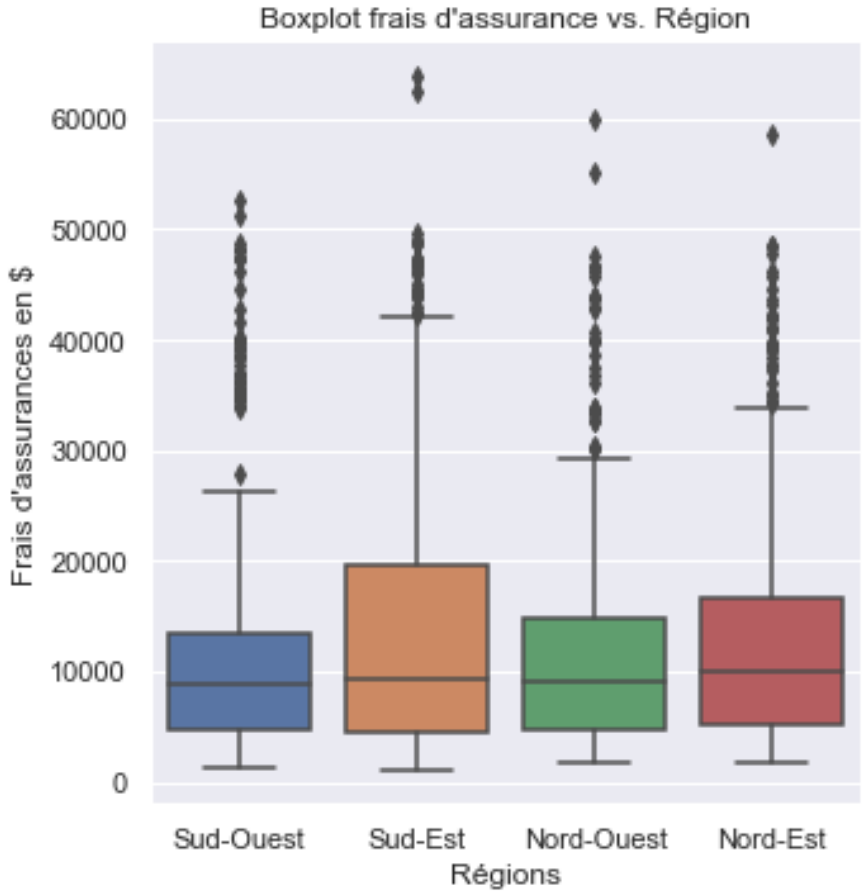
Source : nos traitements statistiques

Le nuage de points ci-dessus suggère une relation linéaire entre les deux variables. Globalement une faible relation linéaire s'observe. Le tableau ci-dessous illustre les résultats du test du coefficients de corrélation.

N	r	IC (95%)	P-value	Conclusion
1338	0,20	[0.15, 0.25]	***<0,01***	La relation linéaire entre l'IMC et les charges d'assurance est significative

Source : nos traitements statistiques

☞ La région selon le coût des frais d'assurance



Source : nos traitements statistiques

Les charges d'assurances minimales et médianes sont quasi similaires dans les différentes régions, cependant la fourchette des dépenses dans la région Sud-Est est plus large, c'est d'ailleurs la région la plus peuplée de l'échantillon. Les résultats de l'ANOVA réalisée entre ces deux variables permettent de conclure (avec un faible risque de se tromper) à l'influence significative de la région sur le coût total des frais d'assurance (**p-value = 0,03**).

IV. Discussion

Au regard des résultats obtenus précédemment, le constat suivant peut être signalé, toutes les variables de notre base de données peuvent être utilisées afin d'expliquer significativement le coût global des charges d'assurance médicale pour un individu.

- ☞ La variable « **Sexe** » impacte significativement sur le coût global des frais d'assurances au seuil de 5%. En effet les individus de sexe masculin ont tendance à dépenser plus d'argent pour les charges, comparées à ceux de sexe féminin. Ce résultat reflète à juste titre la réalité vécue dans la majorité des sociétés dans lesquelles les hommes sont appelées à contribuer de manière plus importante aux besoins de la famille. Une variable comme le statut matrimonial des enquêtés aurait aidé à corroborer cette remarque.
- ☞ S'agissant des variables « **Consommation du tabac** » et « **Indice de masse corporelle** », elles ont tendance à accroitre les dépenses d'assurances effectuées et ceci significativement au seuil de 1% chacune. En d'autres termes, un individu qui déclare consomment du tabac ou un autre qui est obèse ont généralement tendance à plus dépenser. Ces résultats peuvent être justifiés par une majoration appliquée à la prime d'assurance pour les personnes exposées aux nombreux problèmes de santé encourus lorsqu'on dans l'une ou l'autre de ces situations ou les deux (cancer, maladies cardio-vasculaire, diabète, maladies pulmonaires entre autres).
- ☞ Le « **nombre d'enfants** », influence significativement lui aussi la montant des frais d'assurances, bien qu'il soit probable que ce résultat soit plutôt influencé ici par la répartition du nombre d'enfants, dans laquelle les individus sans enfants sont prépondérants.
- ☞ Enfin, la « **région** » influence significativement sur le montant des charges d'assurance. Nous avons trouvé qu'elle a tendance à prendre des fourchettes de valeurs plus grandes dans le Sud-Est. Ce résultat semble indiquer un niveau de vie moyen dans cette région par rapport aux autres, ce qui justifie l'installation majoritaire des personnes dans cette région.

Conclusion et recommandations

Ce travail qui sanctionne la maitrise des connaissances et compétences acquises dans le cadre du premier cours de statistique qui meuble notre formation de Data Scientist, avait pour objectif principal la détermination des variables susceptibles d'influencer le coût global des frais d'assurances médicale pour un souscripteur. Spécifiquement, il s'agissait : d'effectuer des analyses descriptives univariées, d'effectuer des analyses descriptives croisées à deux variables, enfin déterminer et tester l'existence d'éventuelles corrélation/liaison de toutes les autres variables avec la variable. Nous avons mobilisé les techniques d'analyses statistiques descriptives et analytiques apprises en cours pour atteindre ces objectifs. Il s'est avéré au terme de la partie analytique que toutes les variables dites indépendantes avaient une influence sur les charges d'assurance médicale. La principale recommandation serait de poursuivre ce travail en faisant intervenir un modèle statistique (la régression linéaire multiple en l'occurrence) qui aiderait à prédire les charges d'assurances sur la base de certaines valeurs des prédicteurs passés sous forme de paramètres à ce dernier. Ce travail n'a pas été sans bénéfice car il a permis de mettre en pratique les connaissances théoriques acquises dans le cadre du premier cours de statistique.

Références

[1]. Andy. (2019, 04 19). *Logartithmic Transformation in Linear Regression Models: Why & When*. Récupéré sur https://dev.to/: https://dev.to/rokaandy/logarithmic-transformation-in-linear-regression-models-why-when-3a7c

[2]. *Assurances et tabac.* (2015, 06 2). Récupéré sur tabacstop.be: tabacstop.be/nouvelles/assurances-et-tabac

[3]. *Comment communiquer les résultats d'analyses statistiques?* (s.d.). Récupéré sur https://cescup.ulb.be: https://cescup.ulb.be/comment-communiquer-les-resultats-danalyses-statistiques/

[4]. EduTechWikiFr. (2023, 01 23). *Principes de base d'analyse statistique.* Récupéré sur https://edutechwiki.unige.ch/fr: https://edutechwiki.unige.ch/fr/Principes_de_base_d%27analyse_statistique#:~:text=Le%20but%20de%20l'analyse,points)%20ou%20par%20plusieurs%20variables.

[5]. KUMAR, S. (2020). *Linear Regression Tutorial.* Récupéré sur https://www.kaggle.com/: https://www.kaggle.com/code/sudhirnl7/linear-regression-tutorial/notebook

[6]. LAROUSSE. (2022, 01 23). *https://www.larousse.fr/dictionnaires/francais/assurance/5915#586.* Récupéré sur larousse.fr: https://www.larousse.fr/dictionnaires/francais/assurance/5915

[7]. Musacchio, F. (2021, 09 02). *Chapter 13: Statistical Analysis with Pingouin.* Récupéré sur https://www.fabriziomusacchio.com/: https://www.fabriziomusacchio.com/teaching/python_course/13_pingouin

[8]. Santé, O. M. (2020, 08 20). *Obésité et surpoids.* Récupéré sur https://www.who.int/fr: https://www.who.int/fr/news-room/fact-sheets/detail/obesity-and-overweight#:~:text=kg%2Fm2.-,Adultes,%C3%A9gal%20ou%20sup%C3%A9rieur%20%C3%A0%2030.

[9]. Swamynathan, M. (2017). *Mastering Machine Learning with Python in Six Steps : A Practical Implementation Guide to Predictive Data Analytics Using Python.* Bangalore, Karnataka, India: APRESS.

[10]. *Test d'indépendance avec Python.* (s.d.). Récupéré sur https://asardell.github.io/: https://asardell.github.io/statistique-python/

[11]. UsherbrookeCA. (s.d.). *Corrélation.* Récupéré sur https://spss.espaceweb.usherbrooke.ca/: https://spss.espaceweb.usherbrooke.ca/correlation/

Appendices

Les codes pythons

```
# importation des packages et modules nécessaires
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import pingouin as pg
import numpy as np
import scipy
import math

# lecture des données
data = pd.read_csv("insurance.csv", index_col="index")

# apperçu rapide des données
data.head()

# affichage du nombre total d'observations et de variables
n,p = data.shape
print(f"Le jeu de données comporte : {n} individus et {p} variables")

# affichage du nombre de valeurs manquantes par variable
data.isna().sum()

# affichages des types de chaque variable
data.dtypes

# description générales des variables quantitatives
data.describe()

# description générale des variables qualitatives
data.describe(include='object')

# renommage des modalités des variables sex, region et smoker
data["sex"] = data["sex"].replace(["male","female"],["Masculin","Féminin"])
data["smoker"] = data["smoker"].replace(["yes","no"],["Oui","Non"])
data["region"]                                                    =
data["region"].replace(["southwest","southeast","northwest","northeast"],["
Sud-Ouest","Sud-Est","Nord-Ouest","Nord-Est"])

# définition de la fonction qui génère les labels des graphiques
def label_function(val):
    return f'{val/100*len(data):.0f}; {val:.0f}%'

# construction du diagramme pour la région
data.groupby('region').size().plot(kind="pie", autopct=label_function,
                                   textprops={'fontsize':15})
plt.title("Répartition des individus selon la région de résidence")
plt.ylabel("")

# définition de la fonction qui permet d'ajouter des effecifs sur les
# barres de l'histogramme
def hist_annotate(freq, bins, patches):
    # coordonnées x pour les labels
    bins_centers = np.diff(bins)*0.5+bins[:-1]
    i=0
    # on crée un itérable composé de freq, bins_centers et patches pour
```

```
    # le parcours
    for fr, x, patch in zip(freq, bins_centers, patches):
        height = int(freq[i])
        plt.annotate("{}".format(height),
                     xy = (x, height), # coin supérieur gauche de la barre
                     xytext = (0,0.2),
                     textcoords = "offset points",
                     ha = "center", va = "bottom"
                     )

        i = i+1


# histogrammes pour les variables quantitatives

# cas de l'âge
freq, bins, patches = plt.hist(data["age"], alpha=1, edgecolor='black')
hist_annotate(freq = freq, bins = bins, patches = patches)
plt.xlabel("Ages (en années)")
plt.ylabel("Effectifs")
plt.title("Histogramme des ages")
plt.show()


# cas de l'IMC
freq,    bins,    patches    =    plt.hist(data["bmi"],    bins=20,alpha=1,
edgecolor="black")
hist_annotate(freq = freq, bins = bins, patches=patches)
plt.xlabel("IMC (en Kg/m^2)")
plt.ylabel("Effectifs")
plt.title("Histogramme de l'IMC")
plt.show()

# cas du nombre d'enfants
freq, bins, patches = plt.hist(data["children"], edgecolor="black")
hist_annotate(freq, bins, patches)
plt.xlabel("Nombre d'enfants")
plt.ylabel("Effectifs")
plt.title("Histogramme du nombre d'enfants")
plt.show()


# cas des charges
freq, bins, patches = plt.hist(data["charges"], bins=15, edgecolor="black")
hist_annotate(freq, bins, patches)
plt.xlabel("Frais d'assurance (en $)")
plt.ylabel("Effectifs")
plt.title("Histogramme des frais d'assurance")
plt.show()

# Création de la colonne des logarithmes des charges
data["logCharges"] = np.log(data["charges"])
# Appercu des 5 premières lignes des logCharges
data["logCharges"].head()

# histogramme des logCharges
freq,    bins,    patches    =    plt.hist(data["logCharges"],    bins=15,
edgecolor="black")
hist_annotate(freq, bins, patches)
plt.xlabel("Logarithmes de Frais d'assurance (en $)")
plt.ylabel("Effectifs")
```

```
plt.title("Histogramme des logarithmes des frais d'assurance ")
plt.show()

# matrice des corrélations pour les variables quantitatives
sns.set(rc = {'figure.figsize':(10,4)})
data_corr = data.corr()
ax = sns.heatmap(data_corr, xticklabels=data_corr.columns,
                 yticklabels=data_corr.columns, cmap='coolwarm')

# diagrammes de boites à moustaches des charges selon le sexe
sns.catplot(x='sex', y = 'charges', data = data, kind = 'box')
plt.xlabel("Sexe")
plt.ylabel("Frais d'assurances en $")
plt.title("Boxplot frais d'assurance vs. sexe")

# diagrammes de boites à moustaches des charges selon smokder
sns.catplot(x='smoker', y = 'charges', data = data, kind = 'box')
plt.xlabel("Consommation de tabac")
plt.ylabel("Frais d'assurances en $")
plt.title("Boxplot frais d'assurance vs. Fumer")

# diagrammes de boites à moustaches des charges selon la région
sns.catplot(x='region', y = 'charges', data = data, kind = 'box')
plt.xlabel("Régions")
plt.ylabel("Frais d'assurances en $")
plt.title("Boxplot frais d'assurance vs. Région")


# diagrammes de boites à moustaches des charges selon le nombre d'enfants
sns.catplot(x='children', y = 'charges', data = data, kind = 'box')
plt.xlabel("Nombre d'enfants")
plt.ylabel("Frais d'assurances en $")
plt.title("Boxplot frais d'assurance vs. Région")


# test de normalité pour l'âge
pg.normality(data['age'])

# test de normalité pour l'IMC
pg.normality(data['bmi'])

# test de normalité pour le frais d'assurances
pg.normality(data['charges'])

# test de normalité pour le nombre d'enfants
pg.normality(data['children'])

# nuage de points entre les charges et l'âge
sns.scatterplot(x= 'age', y = 'charges', data = data, hue = 'smoker')
plt.xlabel("Age (en années)")
plt.ylabel("Charges d'assurance (en $)")
plt.title("Age Vs. Charges d'assurances")


# nuage de points entre les charges et l'IMC
sns.scatterplot(x= 'bmi', y = 'charges', data = data, hue = 'smoker')
plt.xlabel("IMC (kg/m^2)")
plt.ylabel("Charges d'assurance (en $)")
plt.title("IMC Vs. Charges d'assurances")
```

```
# test de student entre le sexe et les charges
pg.ttest(x = data[data['sex']=='male']['charges'],y = data[data['sex']=='female']['charges'])

# test de student entre le smoker et les charges
pg.ttest(x = data[data['smoker']=='yes']['charges'],y = data[data['smoker']=='no']['charges'])

# entre la région et les charges
scipy.stats.f_oneway(
    data[data['region']=='Sud-Ouest']['charges'],
    data[data['region']=='Sud-Est']['charges'],
    data[data['region']=='Nord-Ouest']['charges'],
    data[data['region']=='Nord-Est']['charges'])
# on peut affirmer avec un faible risque de se tromper qu'au moins deux moyennes
# diffèrent. Autrement dit il y'a une influence significative de la région sur les
# dépenses d'assurances


# entre le nombre d'enfants et les charges
scipy.stats.f_oneway(
    data[data['children']==0]['charges'],
    data[data['children']==1]['charges'],
    data[data['children']==2]['charges'],
    data[data['children']==3]['charges'],
    data[data['children']==4]['charges'],
    data[data['children']==5]['charges'])
# on peut affirmer avec un faible risque de se tromper qu'au moins deux moyennes diffèrent. Autrement dit
# il y'a une influence significative du nombre d'enfants sur les dépenses d'assurances
```