Assur'aimant

Sommaire

Introduction

Base de données

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

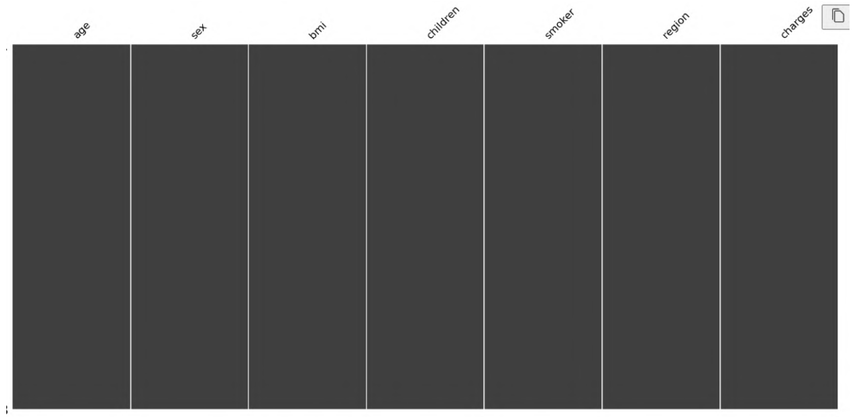
1338 rows × 7 columns

Personnes:1337

Moyenne d'age : 39

BMI : 24-25

Nettoyage de données



Vérifier si il y'a des doublons et les supprimer

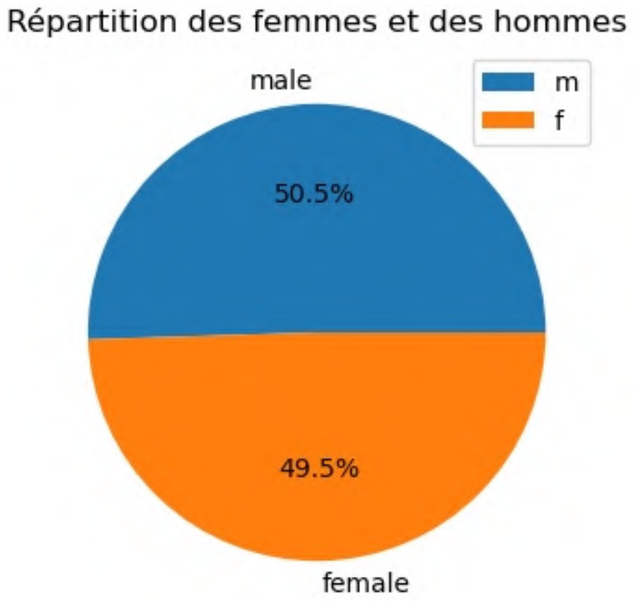
```
data.loc[data.duplicated(keep=False),:]
```

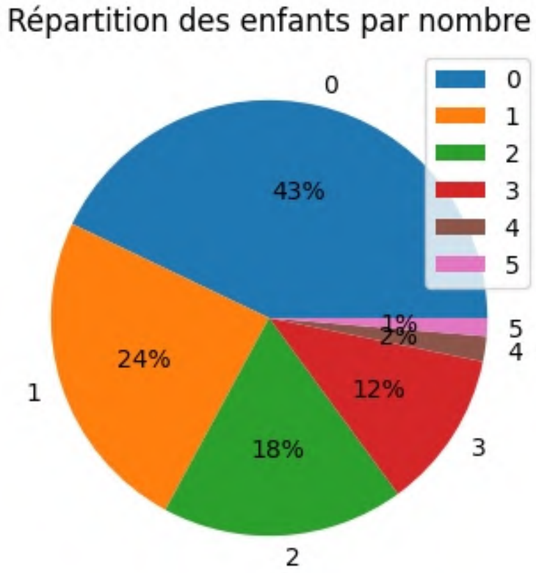
[11] ✓ 0.8s

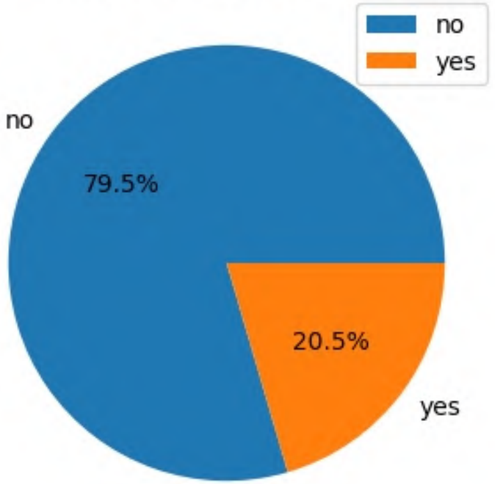
	age	sex	bmi	children	smoker	region	charges
195	19	male	30.59	0	no	northwest	1639.5631
581	19	male	30.59	0	no	northwest	1639.5631

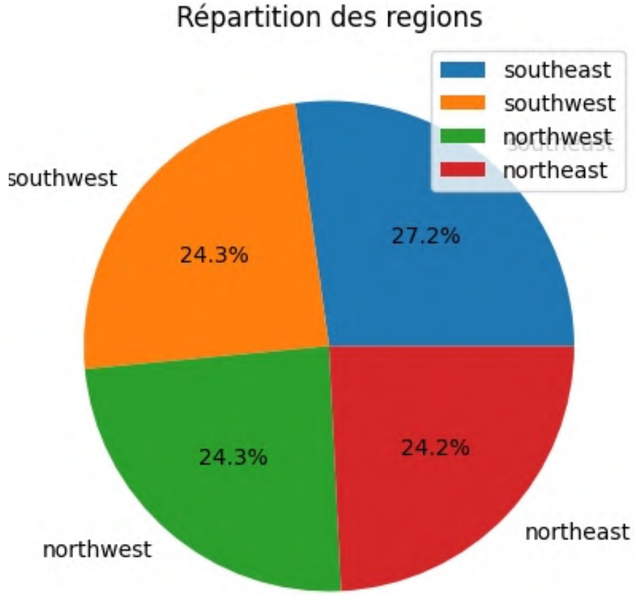
- Vérification de valeurs nulles et les doublons :
  - On remarque qu'il ne y'a pas de valeurs nulles
  - On voit qu'il y'a un seul doublon

Analyse des données



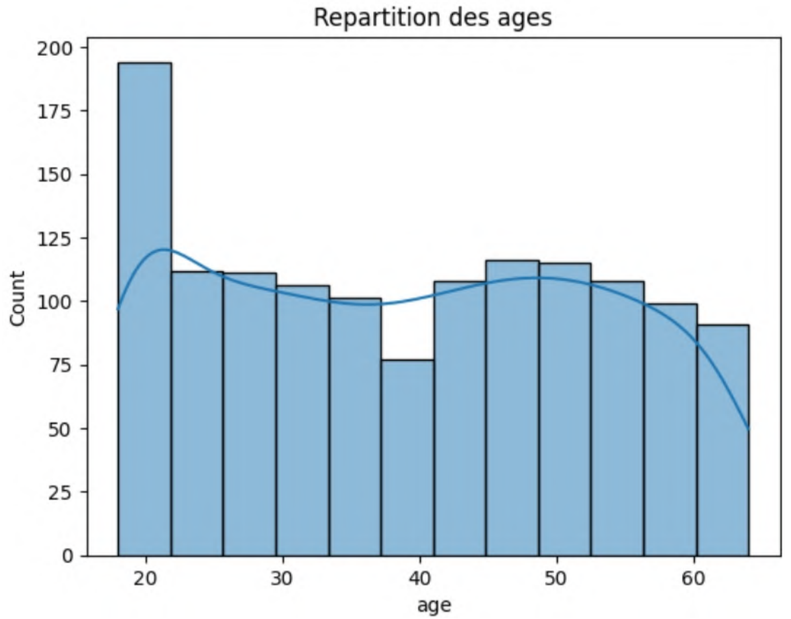




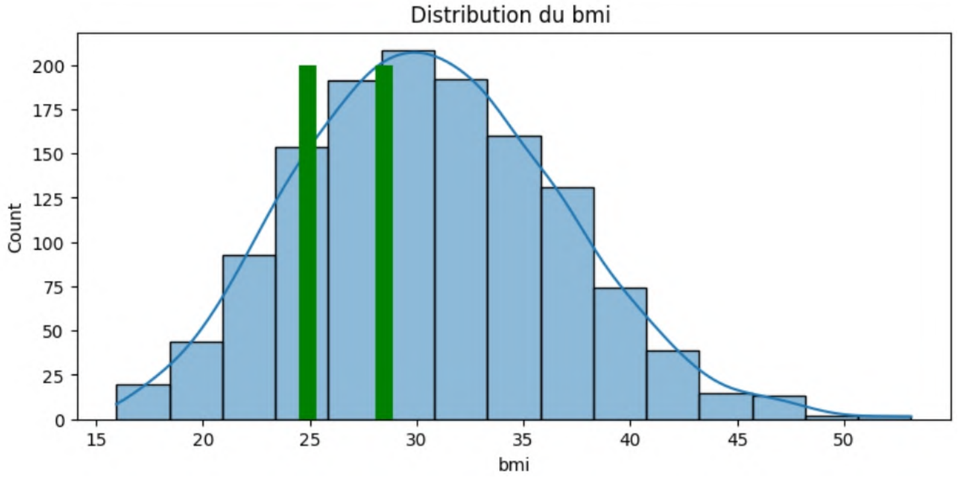


- On remarque que la distribution des personnes selon leurs sexe et selon leurs région est équitable.
- On remarque sur la répartitions des fumeurs, que la majorité des personnes sont non fumeurs.
- On remarque sur la répartitions des enfants, que presque la moitié des personnes sont sans enfants et le reste est reparti entre 1/2/3/4/5 enfants.

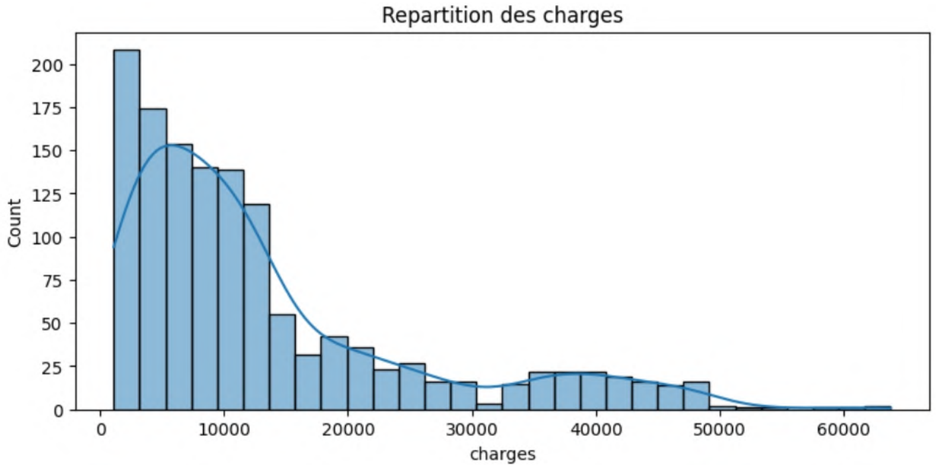
Analyse des données



Ce graphique représente la répartitions des ages



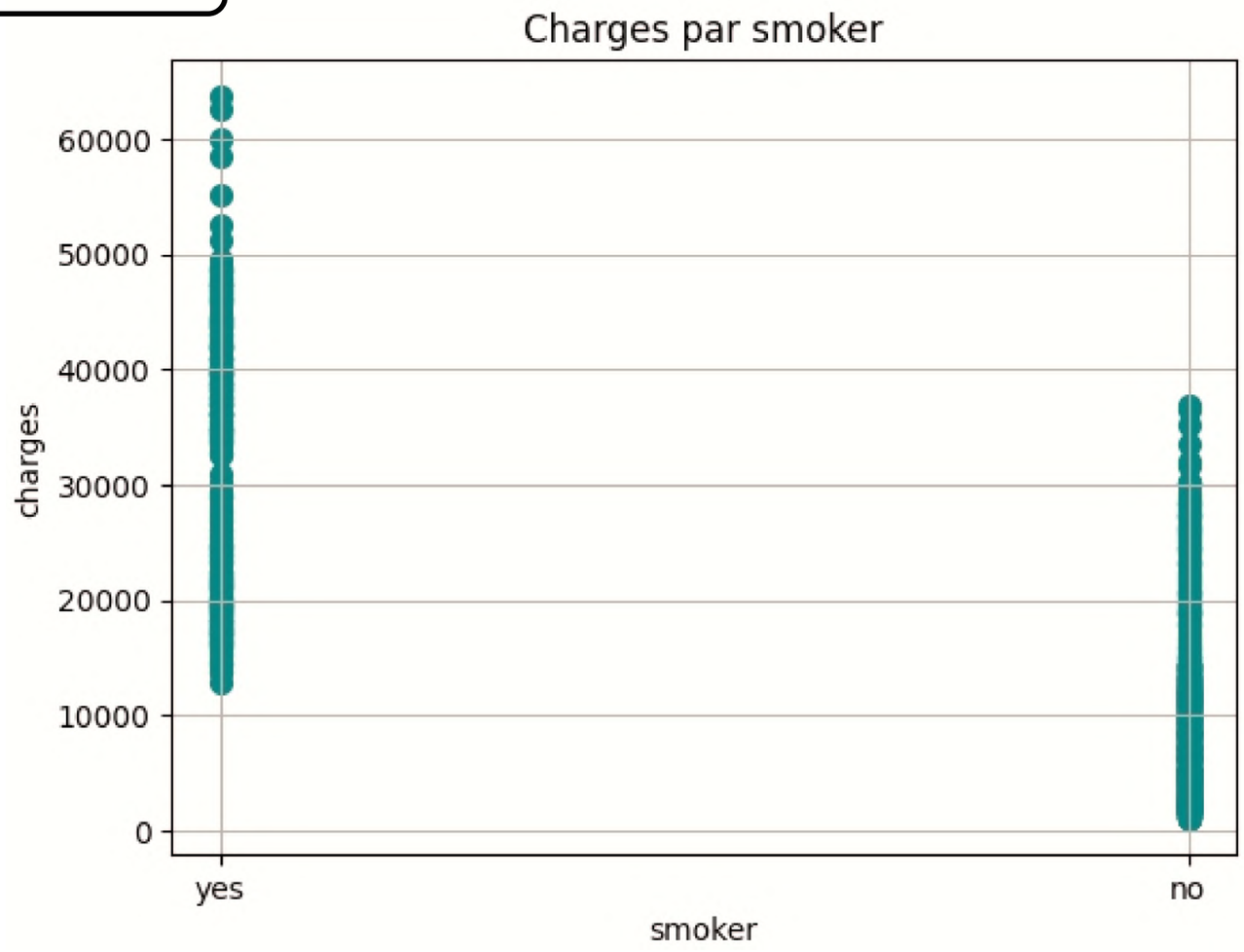
Ce graphique représente la distributions du bmi



Ce graphique représente la distributions des charges

Analyse des données

- On remarque qu'entre les fumeurs et les non-fumeurs il y a une grosse différence de répartition

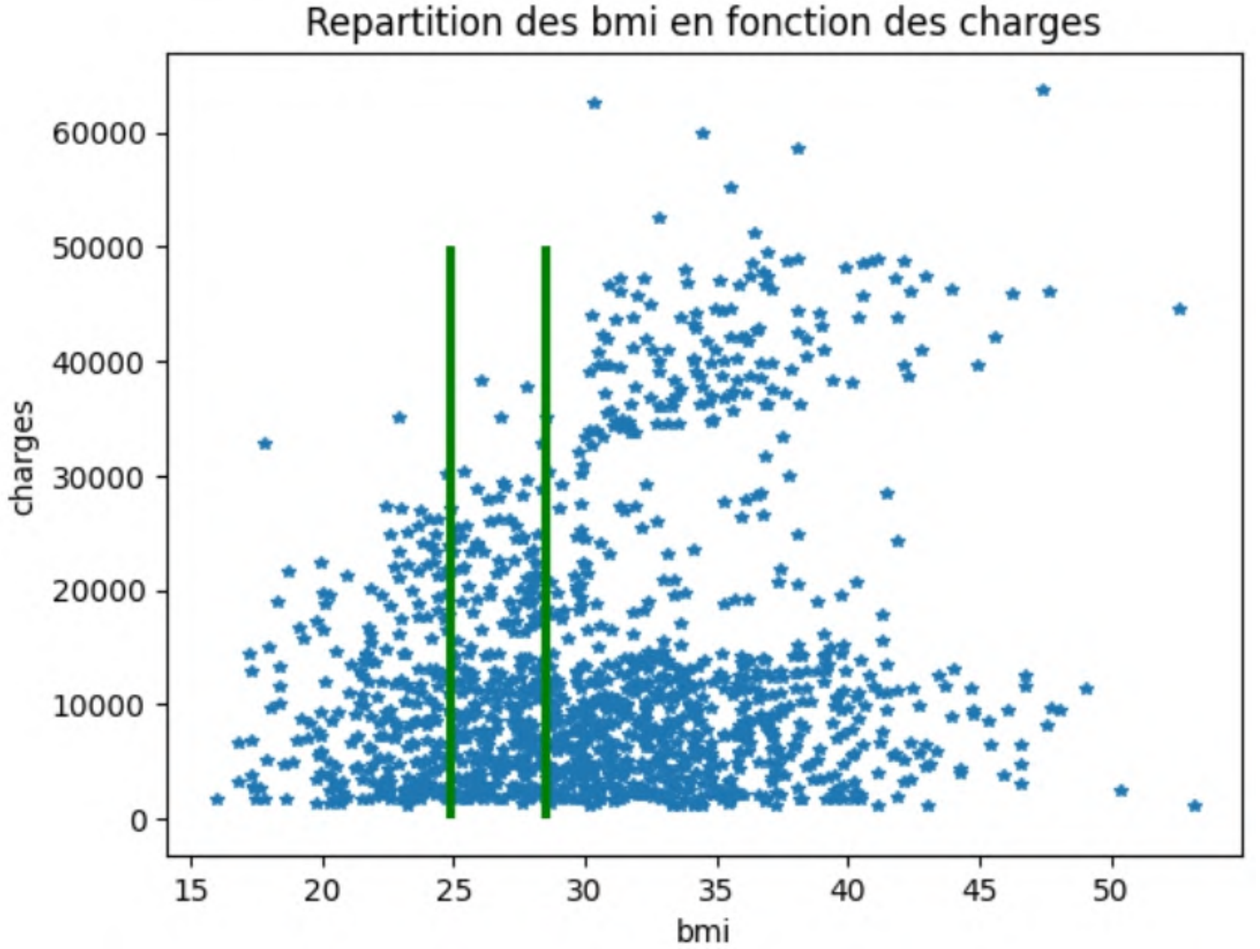


Ce graphique représente les charges en fonction de la caractéristique fumeur

Analyse des données

- On remarque dans ce graphique que le bmi des personnes en bonne santé est compris entre 25~28, or, on voit dans notre répartition qu'il y'a une quantité importante de personnes qui sont en dehors de cet intervalle.

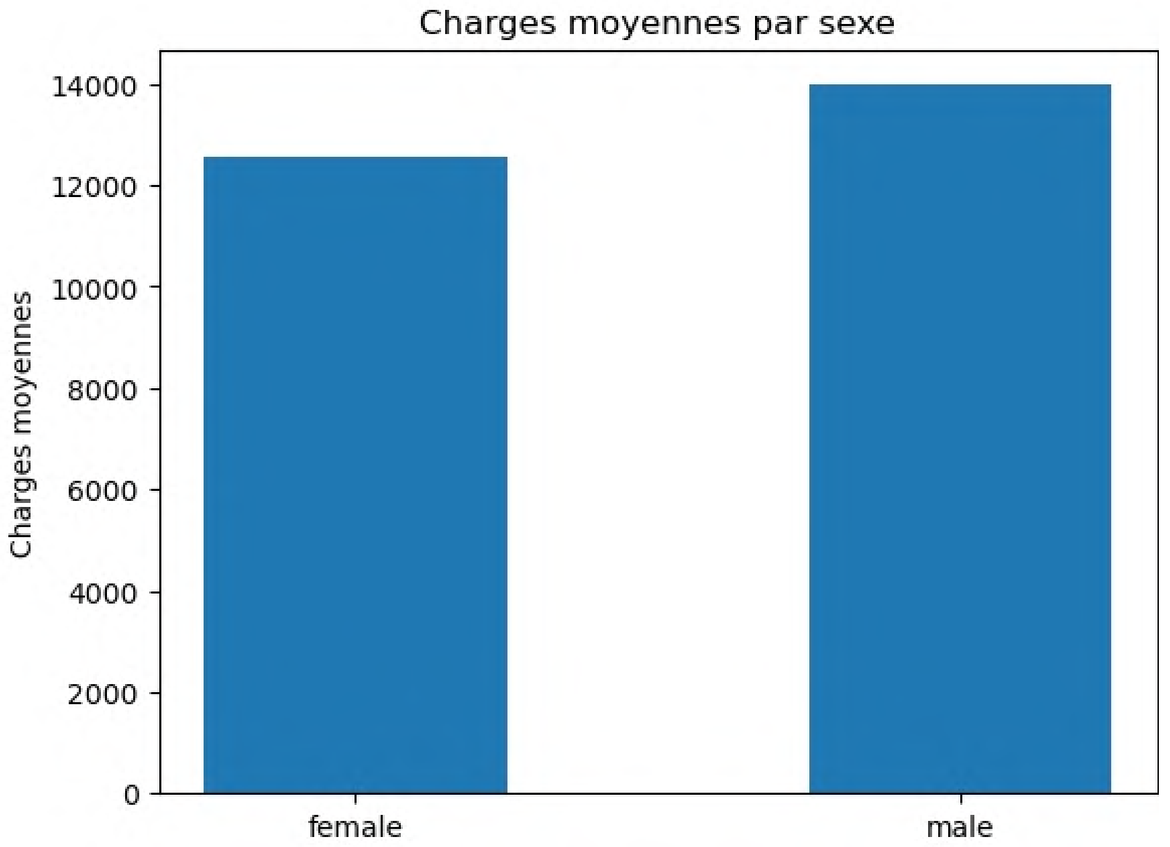
- D'une manière générale, les valeurs aberrantes voire extrêmes sont en dehors de notre intervalle "bonne santé".



Ce graphique représente les charges en fonction du bmi

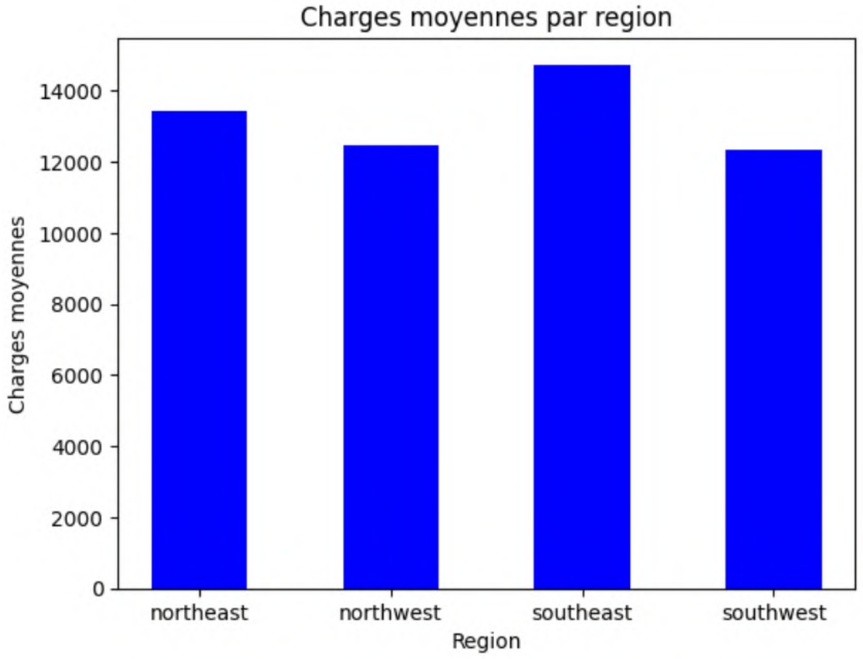
Analyse des données

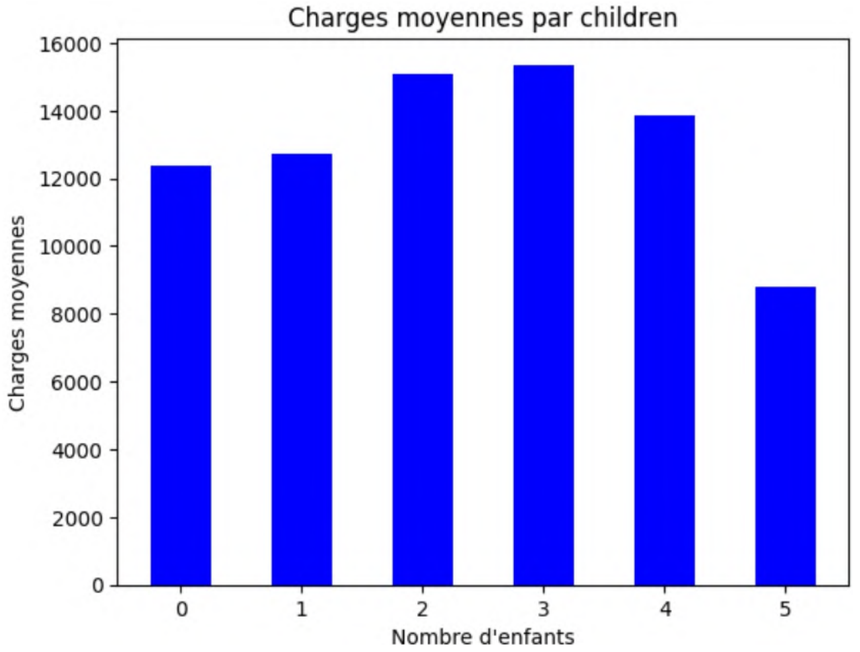
- On remarque que sur les charges moyennes les hommes payent plus que les femmes.



Ce graphique représente les charges moyennes par sexe.

Analyse des données



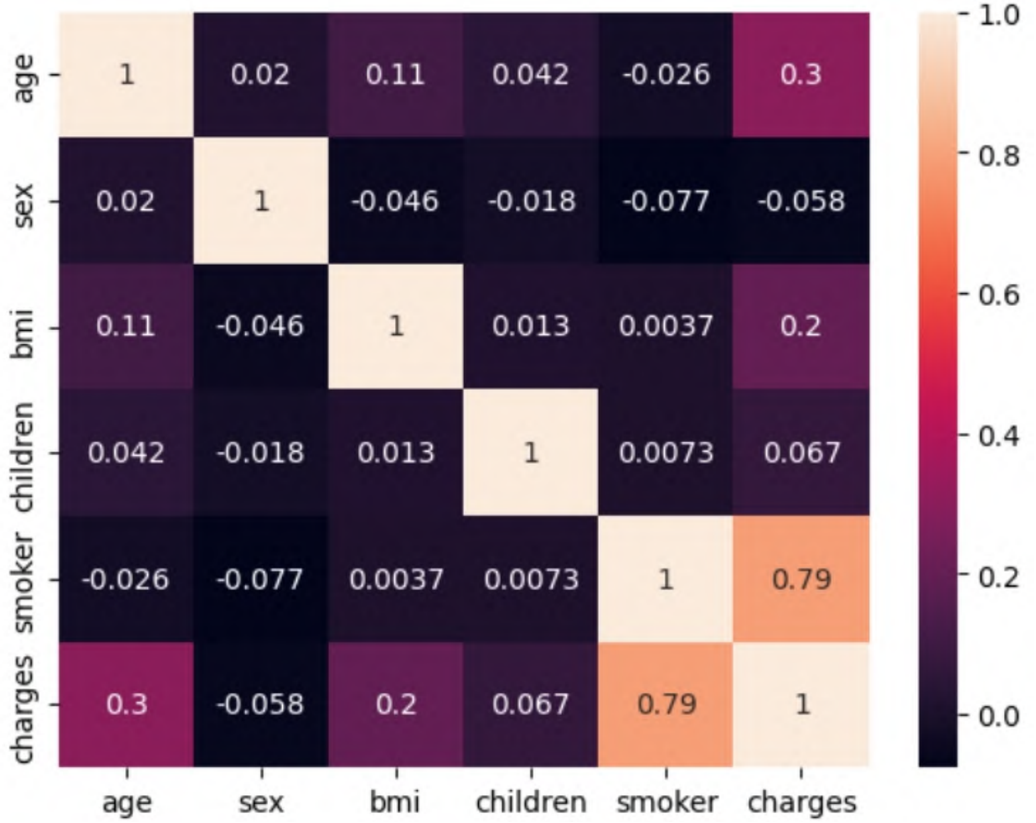


Ces 2 graphiques représentent les charges moyennes en fonction de nombres d'enfants ,
et en fonction de la répartions des gens sur les régions

Correlation entre les caracteristiques et la cible

Coefficient de corrélation de Pearson

```
charges     1.000000
smoker      0.787234
age         0.298308
bmi         0.198401
children    0.067389
sex        -0.058044
dtype: float64
```



La **corrélation** mesure à quel point deux variables sont liées. Si deux variables sont fortement corrélées, une augmentation d'une variable entraînera généralement une augmentation ou une diminution de l'autre.

Analyse des données

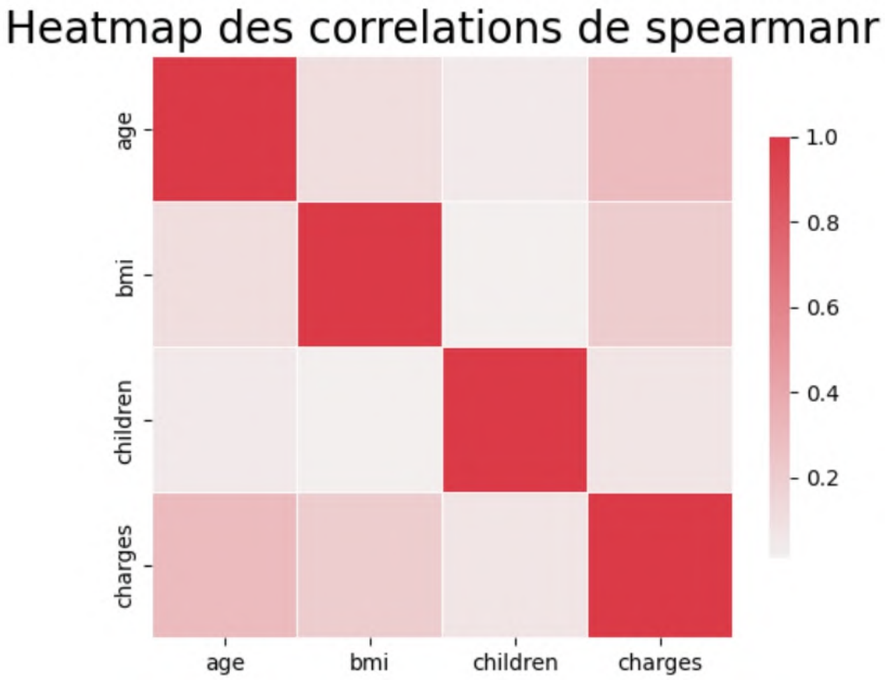
Test de corrélation de Spearmanr

Outil permettant de mesurer la **relation** entre deux variables

- hypothèse nulle:
  les caractéristiques sont indépendantes et n'ont aucune influence sur la cible.
- hypothèse alternative:
  les caractéristiques ont une influences sur la cible.

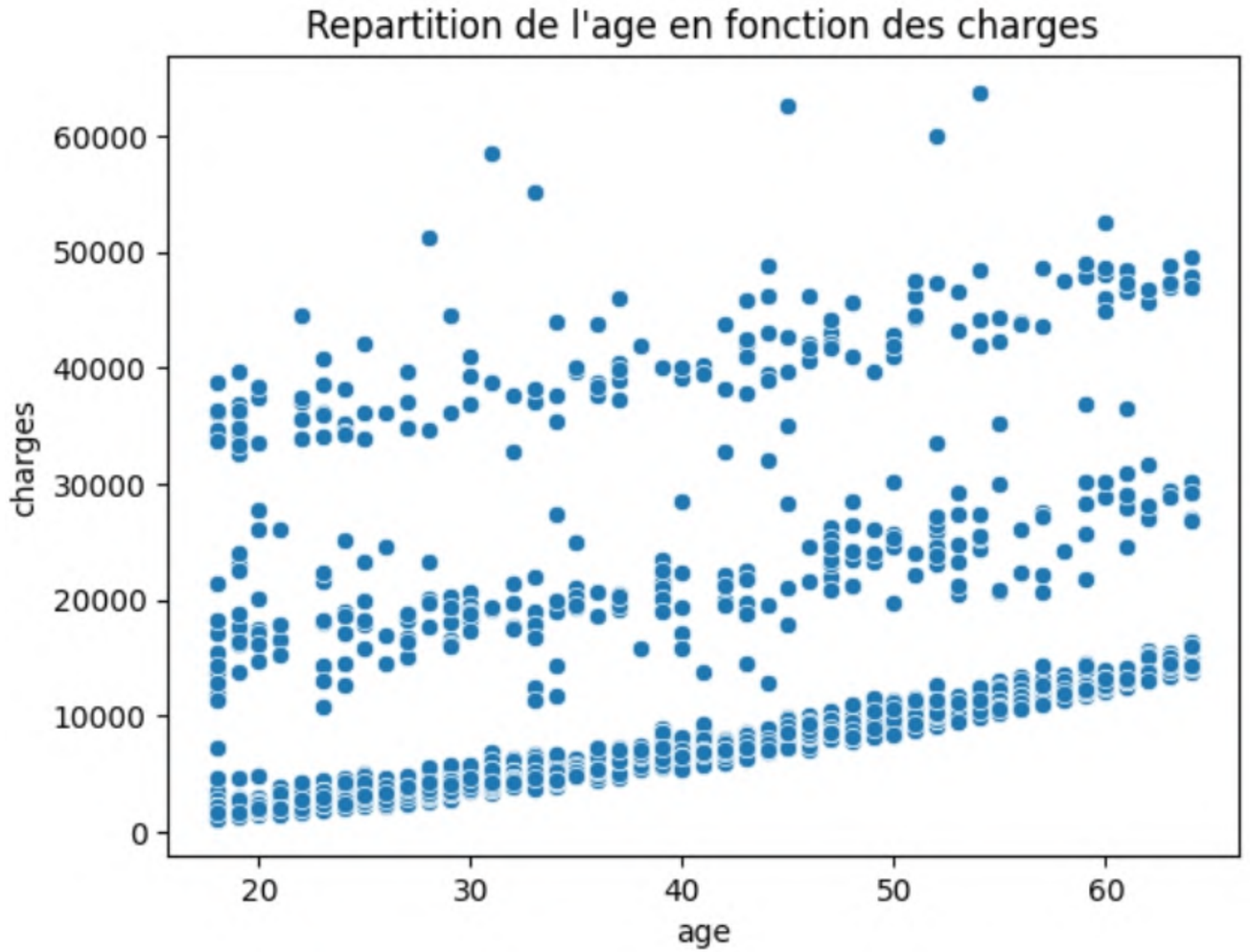
Seul de signification pour rejeter l'hypothèse nulle p-value = 0.05

Dans ce cas, **on rejette l'hypothèse nulle**
les caractéristiques sont toutes liés à la cible.



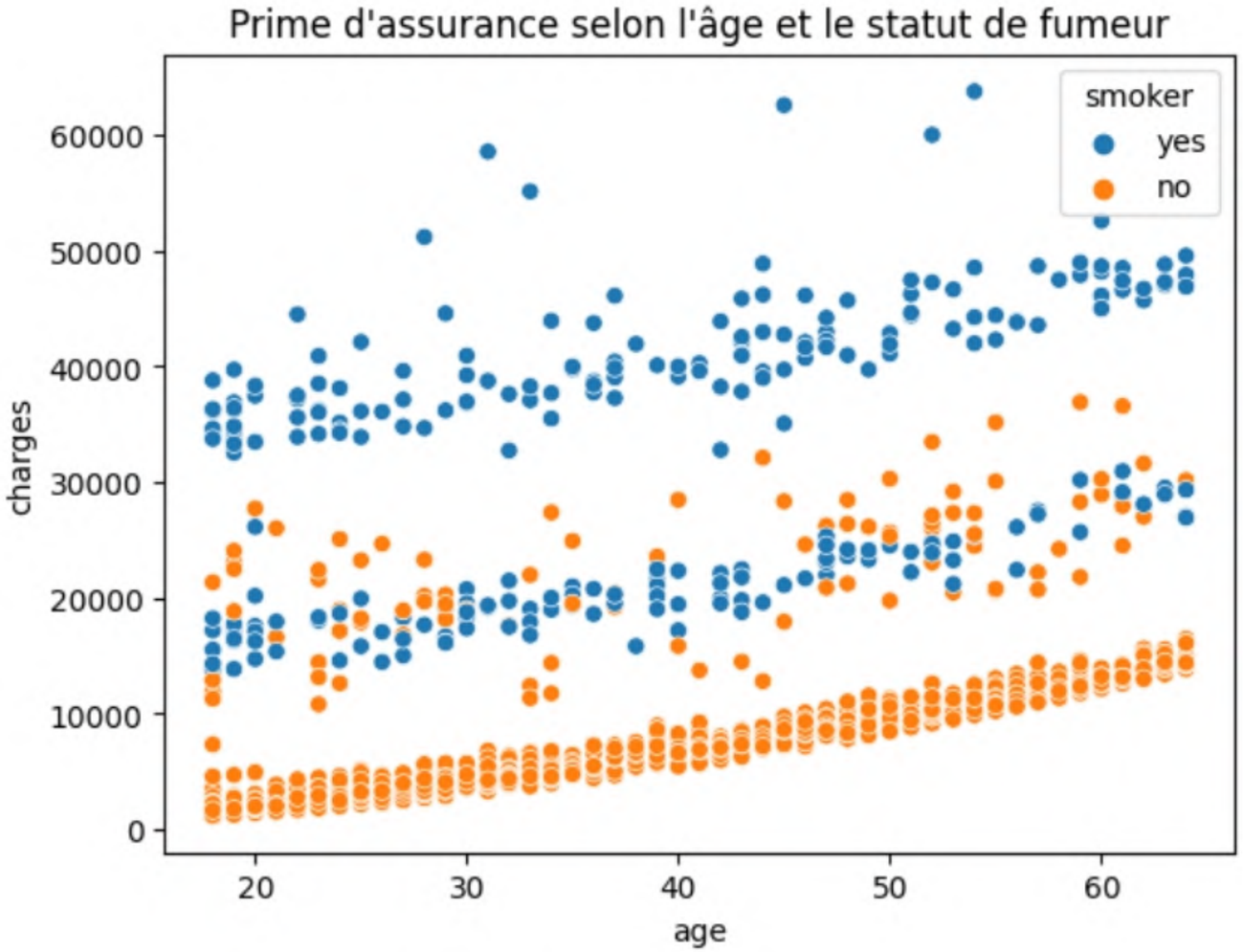

```
Correlation for age: 0.5335232787189862, p-value: 3.1877556503224748e-99
Correlation for bmi: 0.11958495819244366, p-value: 1.1637179203181515e-05
Correlation for children: 0.13220013322835855, p-value: 1.2303764274728685e-06
```

Analyse des données



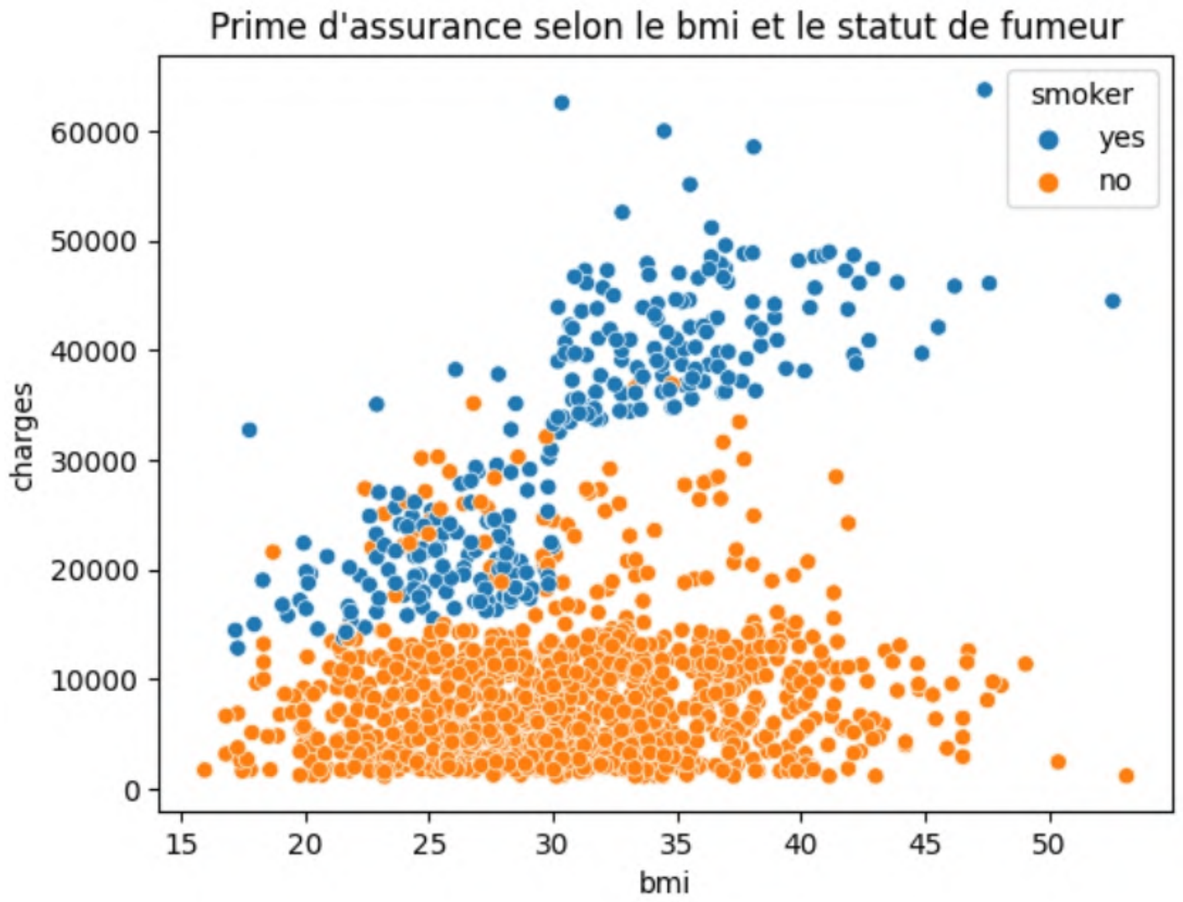
Cette figure représente les charges en fonction d'age

Analyse des données



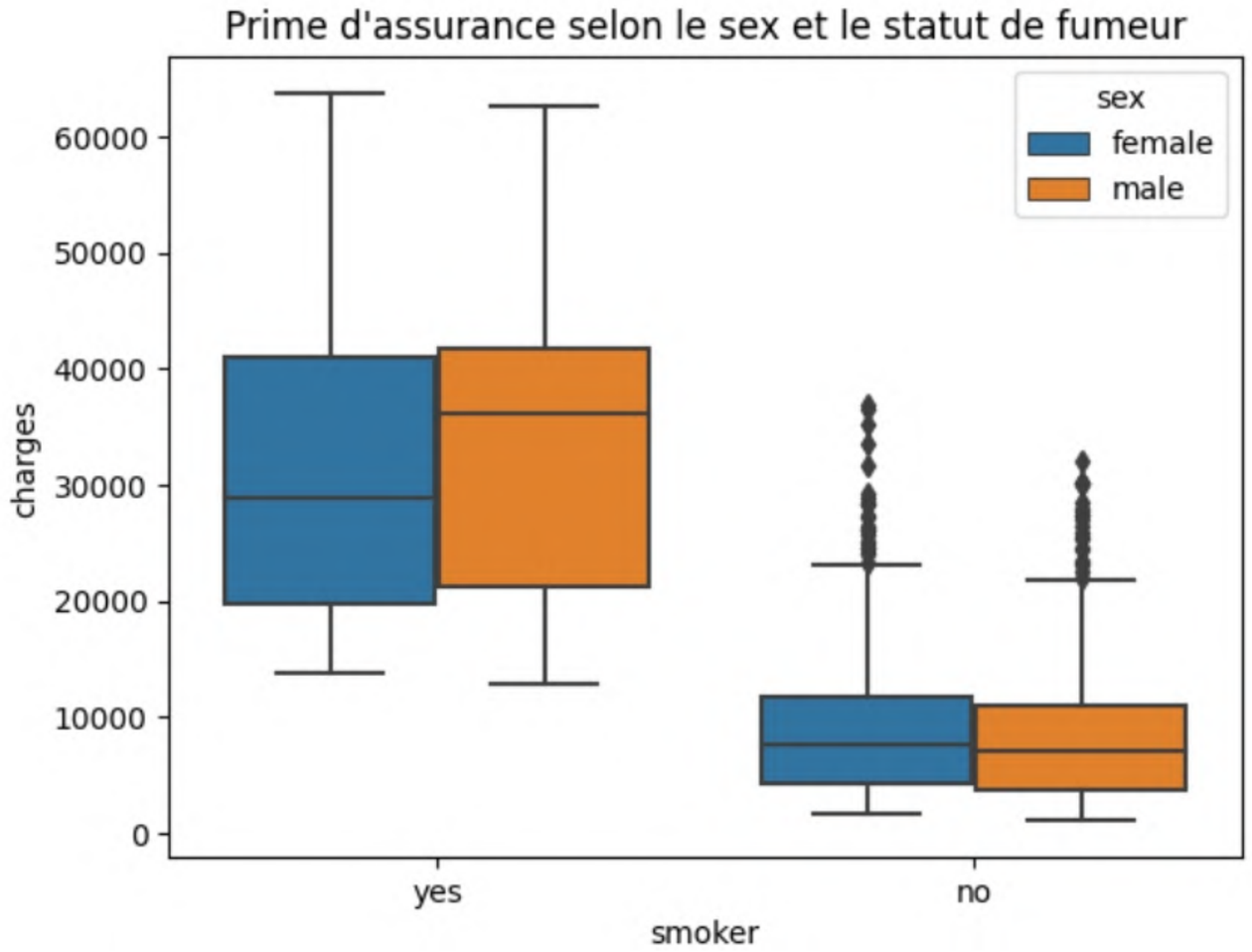
Cette figure représente les charges en fonction d'age et le statut fumeur ou non fumeur

Analyse des données



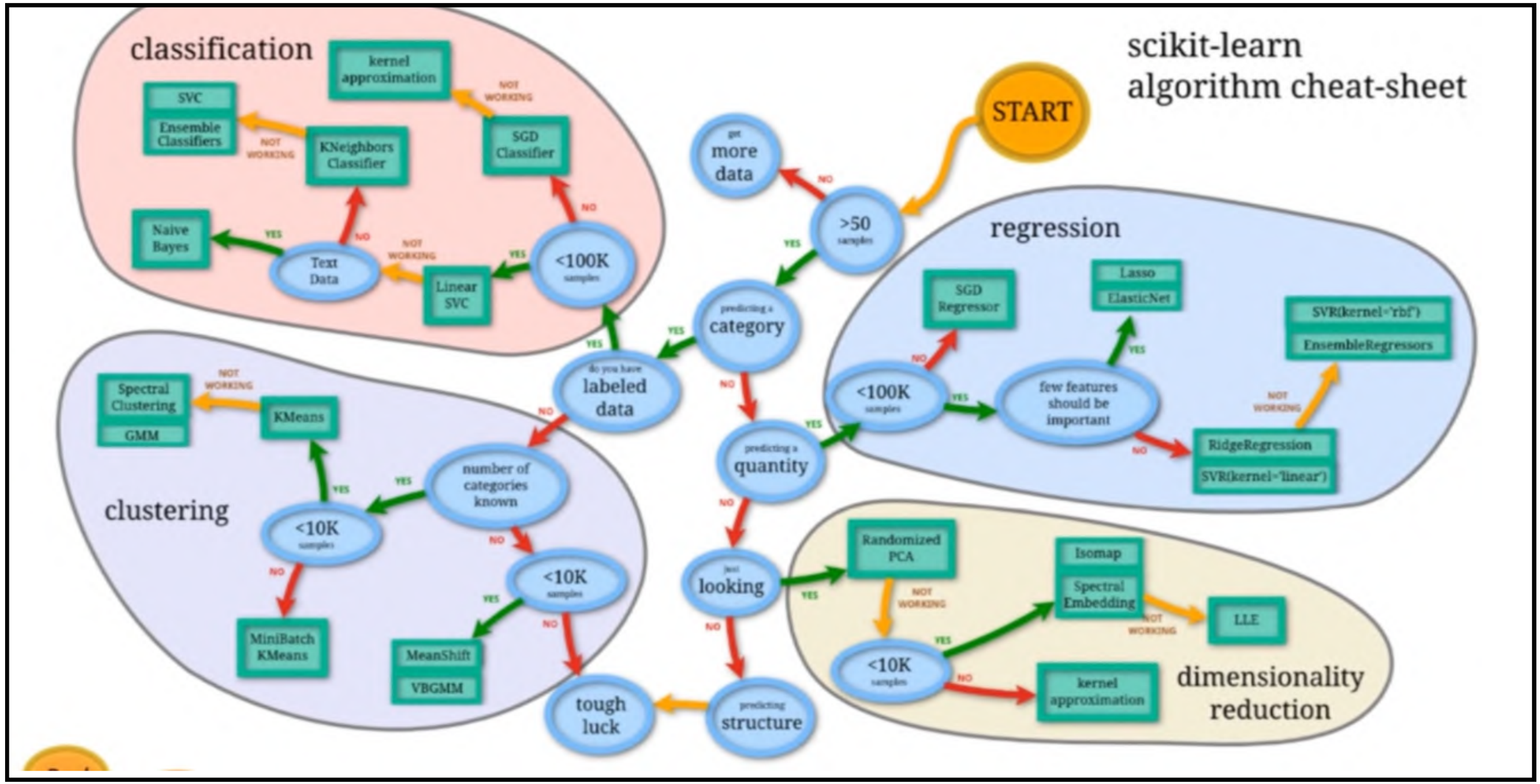
Cette figure représente les charges en fonction du bmi et le statut fumeur ou non fumeur

Analyse des données



Cette figure représente les charges en fonction de sexe et le statut fumeur ou non fumeur

Modélisation



Modélisation

Régression Linéaire

Lasso

Ridge

ElasticNet

Modèles utilisés

Régression Linéaire

Simple, relation linéaire entre les variables

Lasso

Régression linéaire, avec une régularisation des coeff des variables

Ridge

Régression linéaire, pénalité pour réduire le nombre de variables. Cela entraîne une perte d'informations, et réduit la précision du modèle, mais peut aussi le rendre plus stable et moins sujet aux fluctuations.

ElasticNet

Combinaison de RL et de Ridge

Scores de Modélisation

- Lasso : 83.17%
- Ridge : 83.17%
- EasticNet : 83.28%
- Régression Linéaire : 85.54 %

Conclusion

On conclut qu'avec le travail accompli, on a pu mieux connaître la clientèle tout en étudiant et en s'intéressant de près à leurs données démographique. Ce qui nous a amené à prédire la prime d'assurance qu'une personne peut payer d'une manière plus précise, on vous laisse découvrir cela sur notre Streamlit.

Piste d'amélioration

01

Plus d'informations

- Santé
- Mode de vie
- Base de données plus grande

02

Plus de temps

- Étudier plus de modeles

03

Plus de compréhension

- Faire des sous-groupes pour le bmi
- Comprendre les valeurs atypiques influentes

Merci