SIMPLON
.co

PRÉDICTION DES CHARGES D'ASSURANCE

Fatima, Umberto, Flora

Objectif

- Analyse de données pour un assureur
- Objectif : produire un modèle de régression linéaire
- Prédiction des charges d'assurance en fonction de critères démographiques et médicaux
  - Démographiques → âge, nombre d'enfants, genre, région
  - Médicaux → BMI (Indice de Masse Corporelle), fumeur

Set de données

- 7 variables
- Cible : charges
- 3 variables numériques :
  - age
  - bmi
  - children
- 3 variables catégorielles :
  - sex
  - region
  - smoker

- Pas de valeurs manquantes
- 1 doublon supprimé

Ajout de 2 catégories

age_category

Tranches d'âges par décennie :

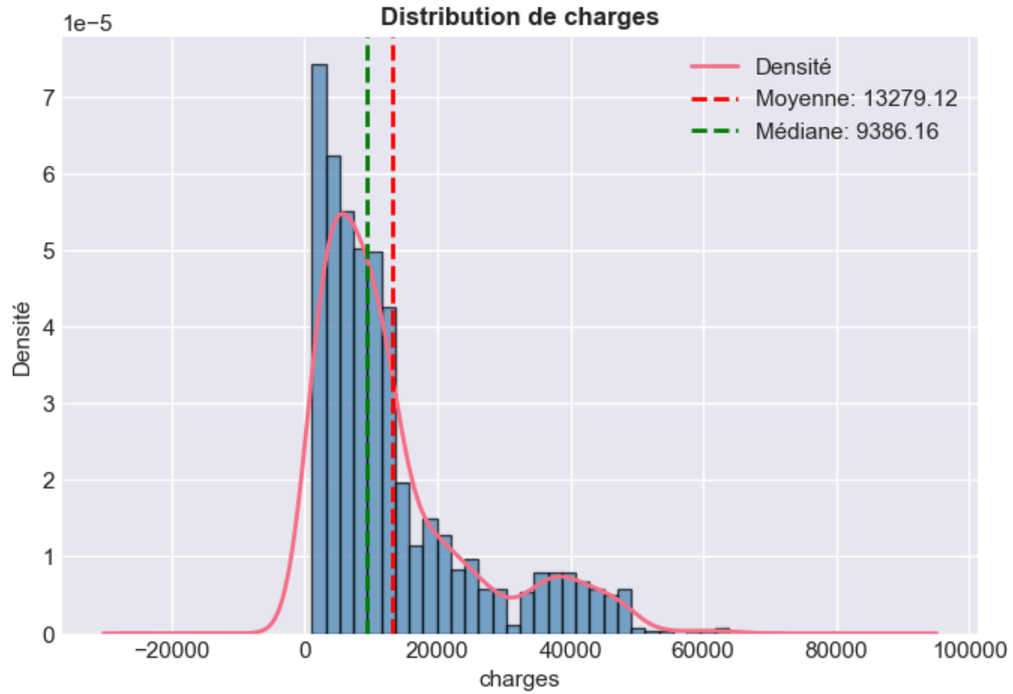
- 18s (18 et 19 ans)
- 20s
- 30s
- 40s
- 50s
- 60s

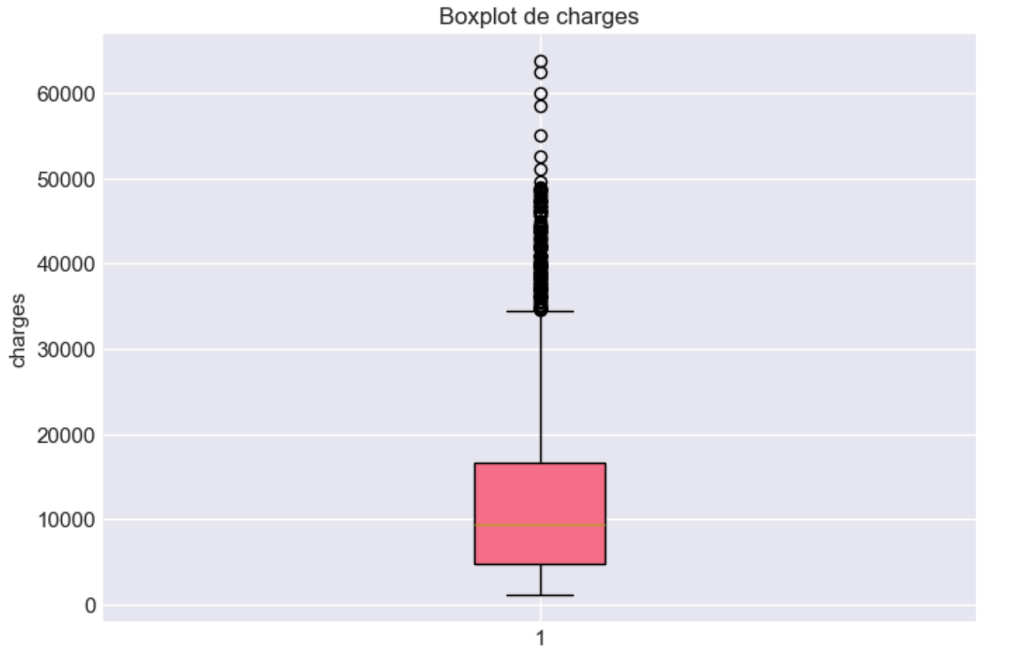
bmi_category

Catégories de l'OMS :

- Underweight < 18.5
- Healthy < 25
- Overweight < 30
- Obesity I < 35
- Obesity II < 40
- Obesity III > 40

	age	sex	bmi	children	smoker	region	charges	age_category	bmi_category
0	19	female	28.0	0	yes	southwest	16885.0	18s	Overweight
1	18	male	34.0	1	no	southeast	1726.0	18s	Obesity I
2	28	male	33.0	3	no	southeast	4449.0	20s	Obesity I
3	33	male	23.0	0	no	northwest	21984.0	30s	Healthy
4	32	male	29.0	0	no	northwest	3867.0	30s	Overweight





Asymétrie et valeurs aberrantes des charges

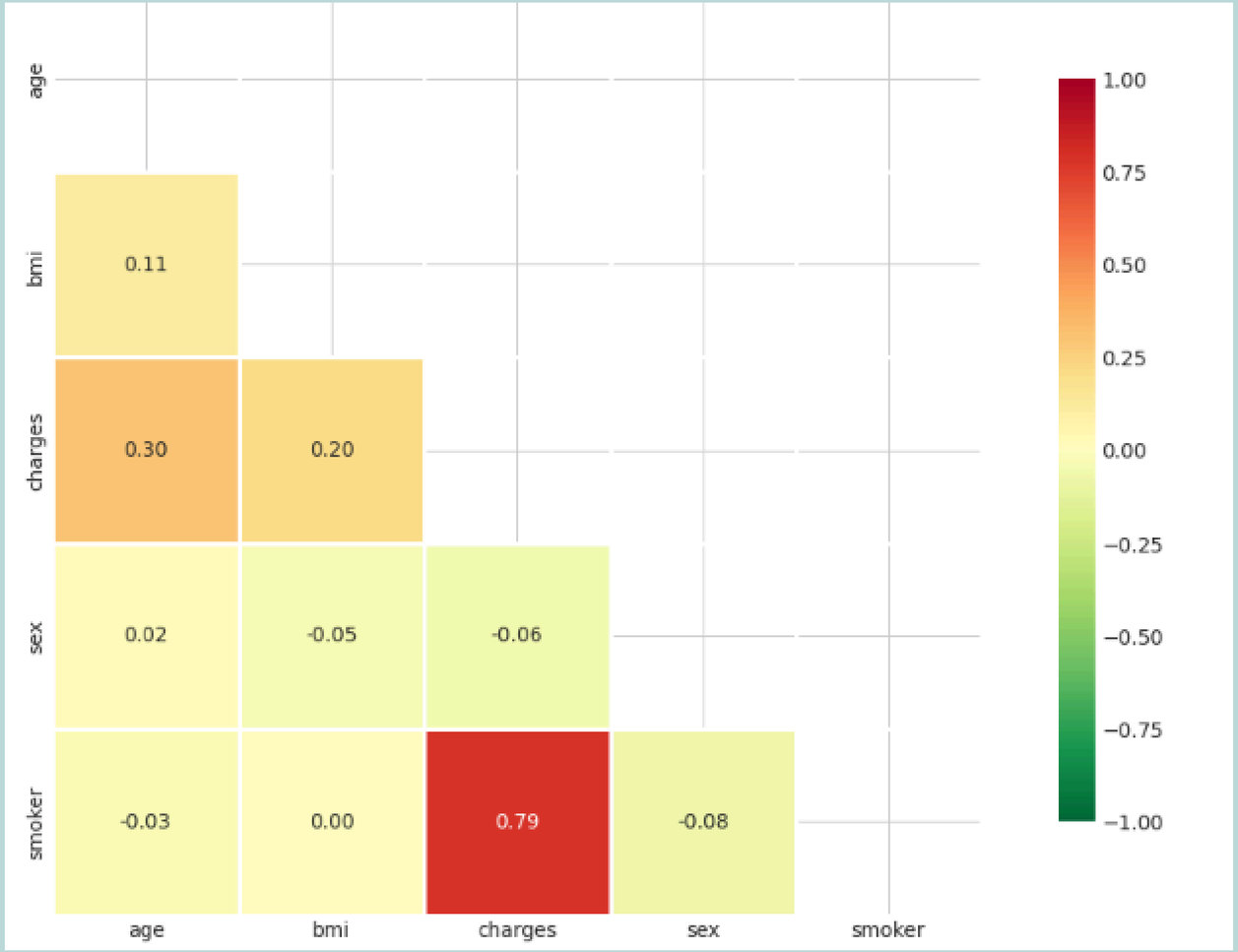
Skewness : 1.52
Données condensées avec quelques valeurs fortement élevées

Outliers : 139 (≈ 10 %)
Nombreuses valeurs extrêmes

Données déséquilibrées → modèle de régression linéaire moins performant

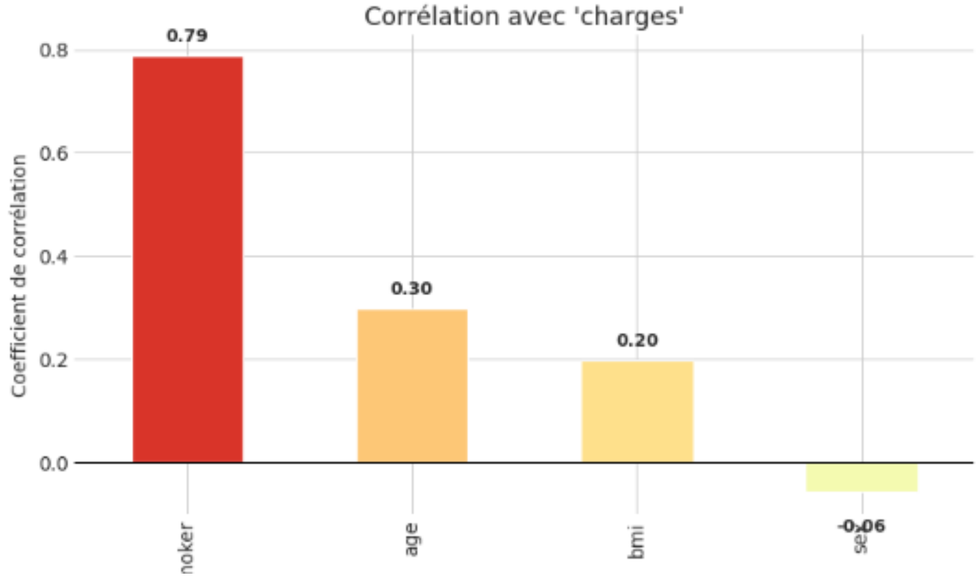
- Conservation des outliers pour ne pas biaiser les données
- Observation à prendre en compte dans les analyses

Matrice de corrélations

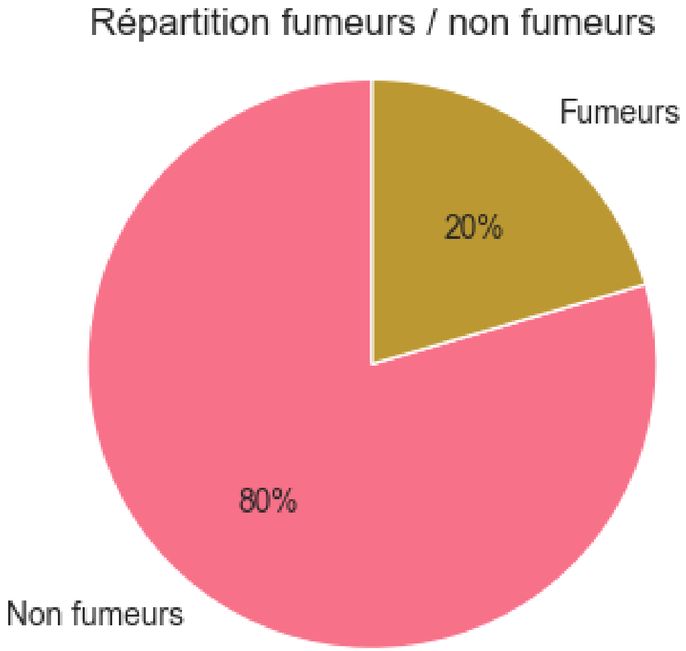


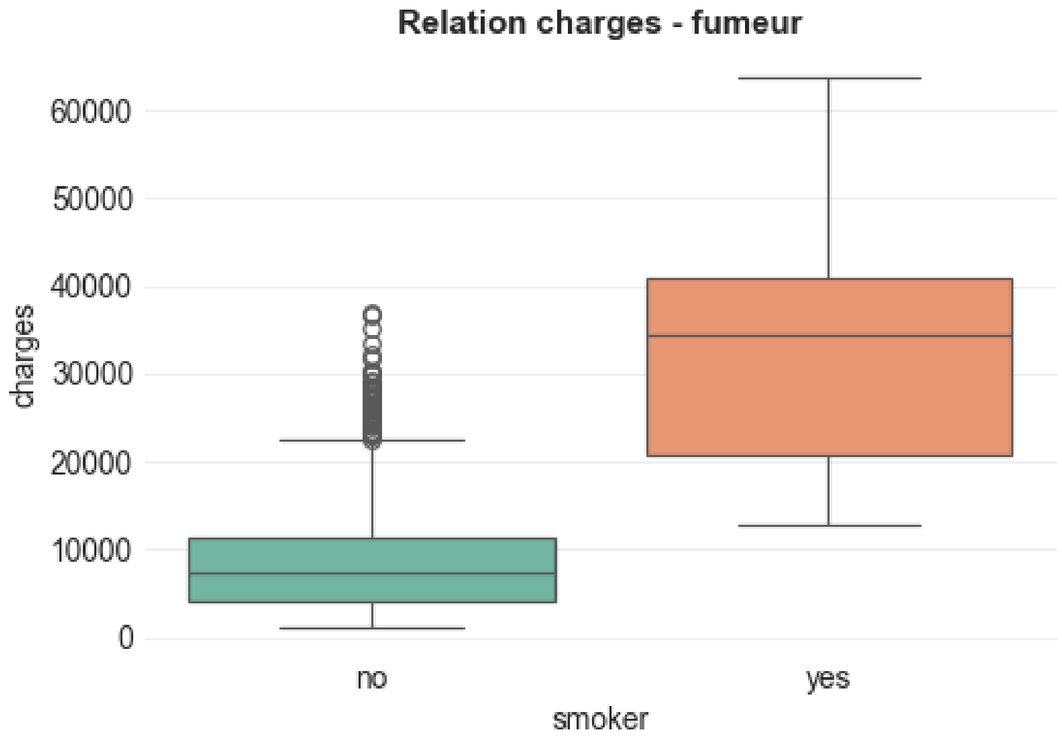
- Fumer est fortement associé à des charges plus élevées.
- L'âge et le bmi ont un effet modéré

Les autres variables sont indépendantes.



Relation charges - fumeur

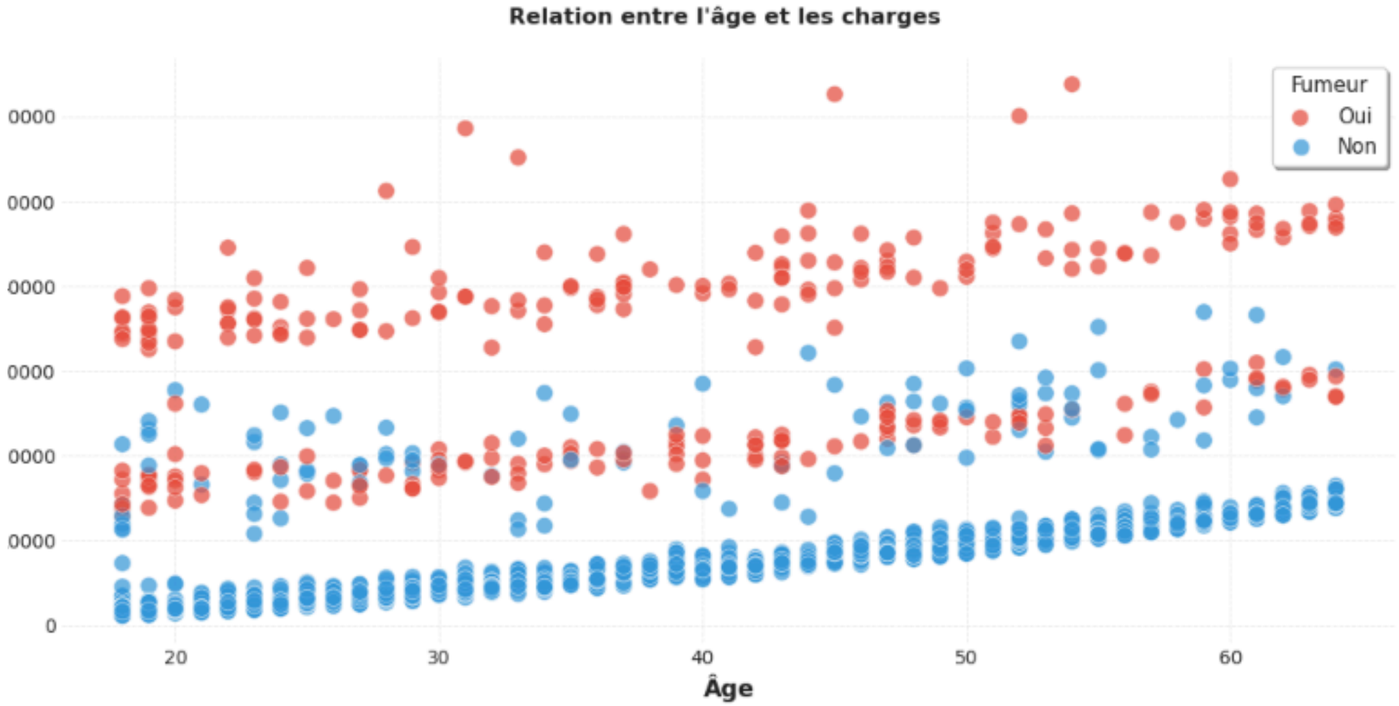


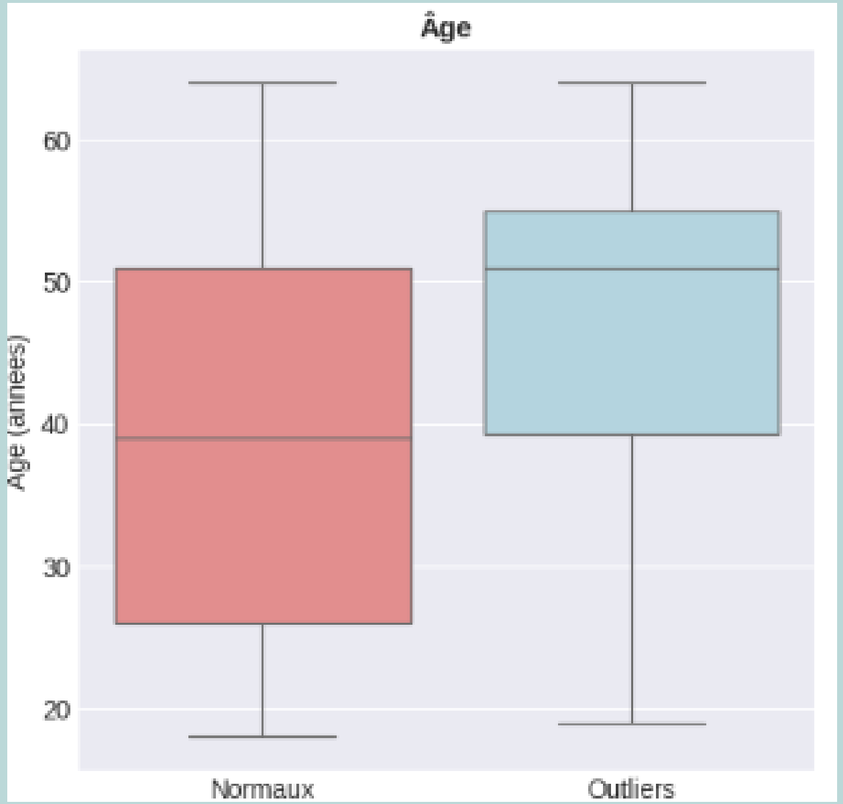


- Moyenne charges non fumeurs : 8 440 (médiane : 7346)
- Moyenne charges fumeurs : 32 050
- Les fumeurs paient en moyenne 3,8 fois plus

Fumeur : variable ayant le plus d'impact sur les charges

Relation charges - fumeur - âge





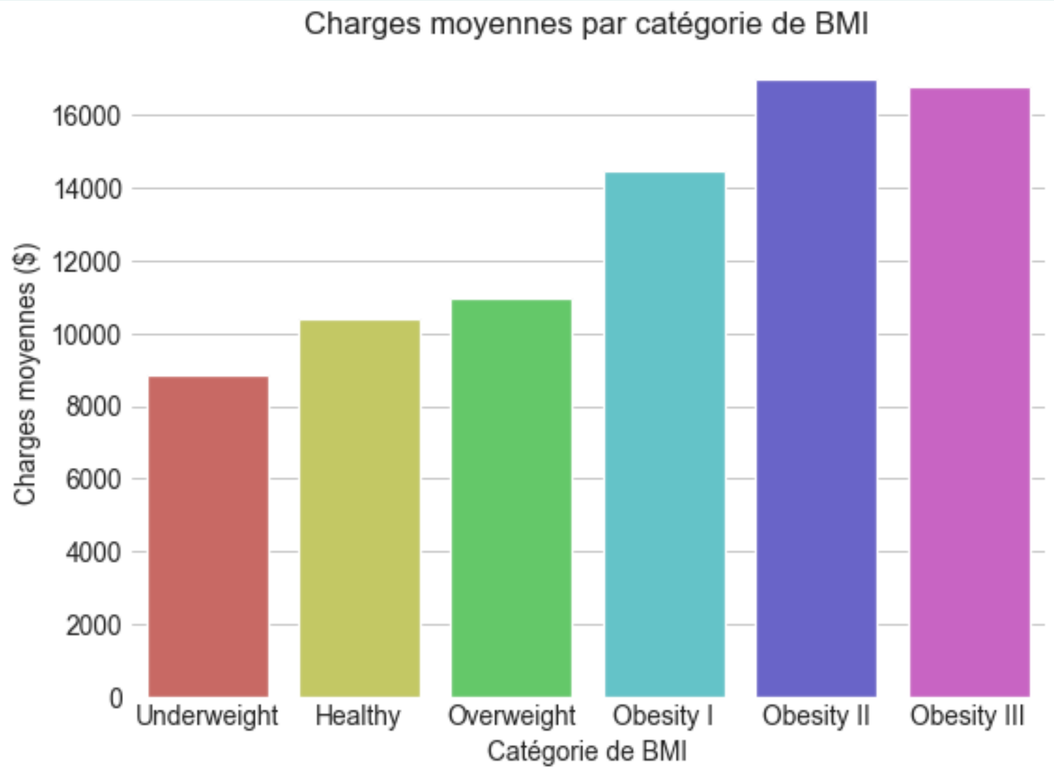
Corrélation charges - âge :

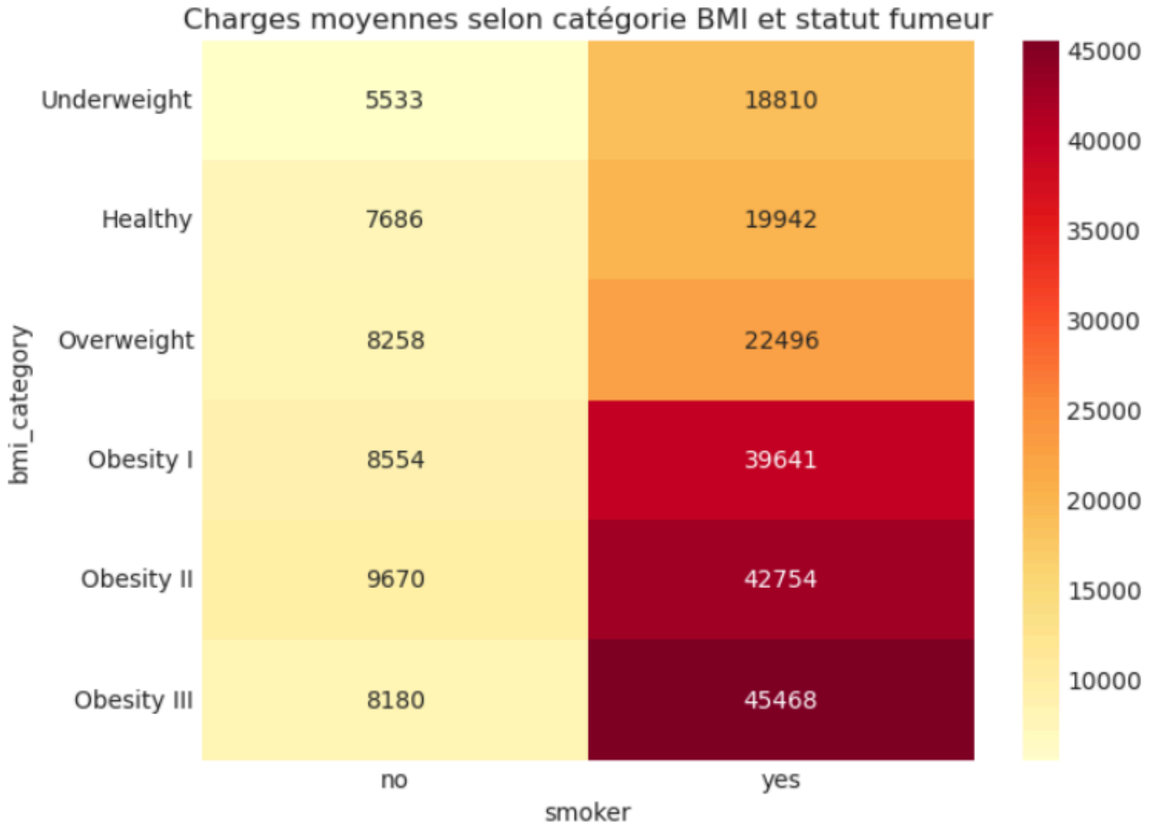
- chez les fumeurs : 0.37
- chez les non fumeurs : 0.63

L'âge explique ~40 % de la variation des charges

Relation charges - BMI

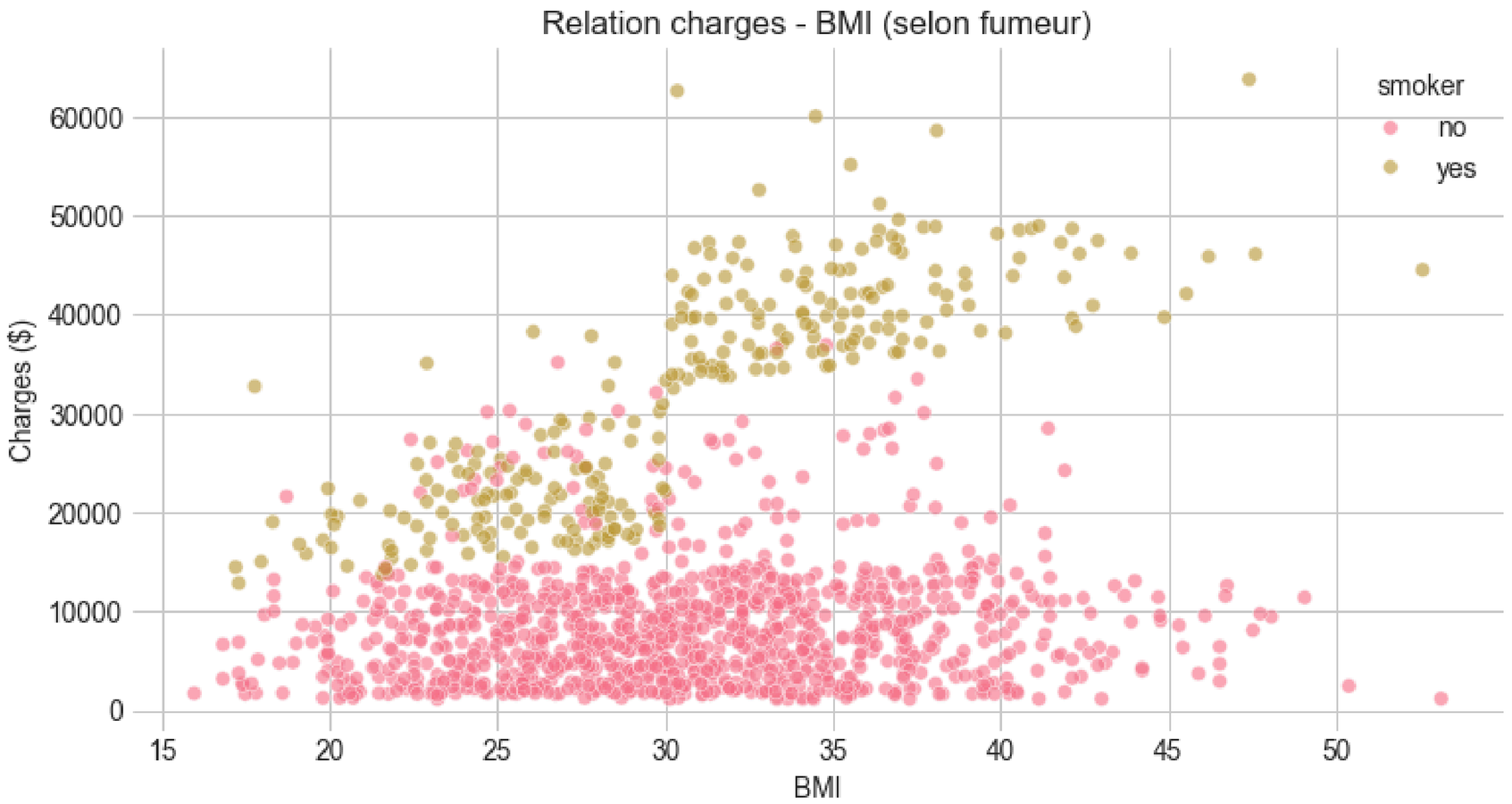
- Corrélation globale modérée mais non négligeable (0.30)





- Corrélation importante à partir de la catégorie Obesity I

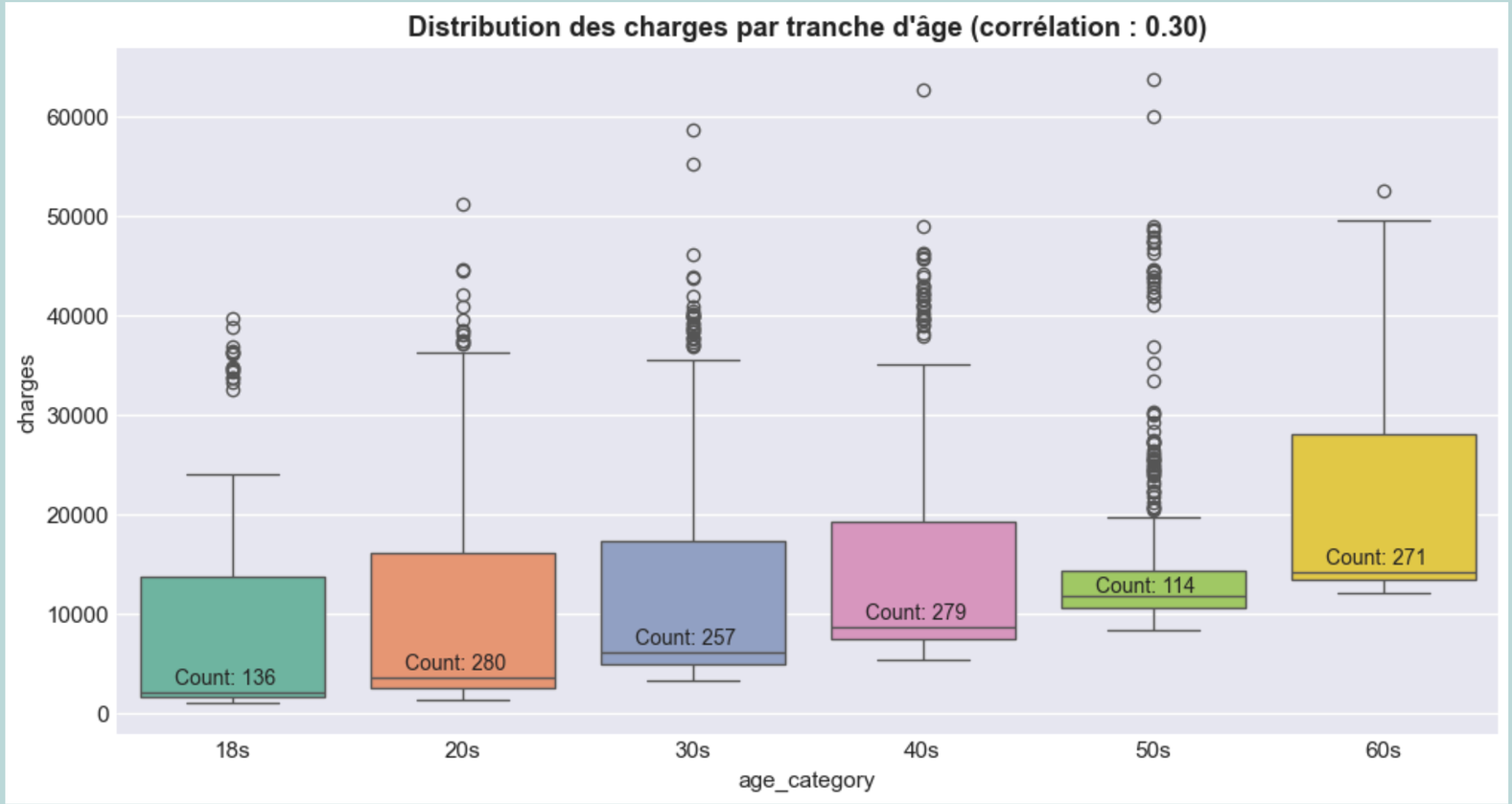
Relation charges - BMI



- Corrélation très forte entre les charges et le BMI pour les personnes fumeuses (0,80)
- Corrélation négligeable pour les personnes non fumeuses (0,08)

Le BMI impacte les charges, surtout chez les fumeurs

Relation charges - âge



- Corrélation linéaire positive
- 50s : peu de données

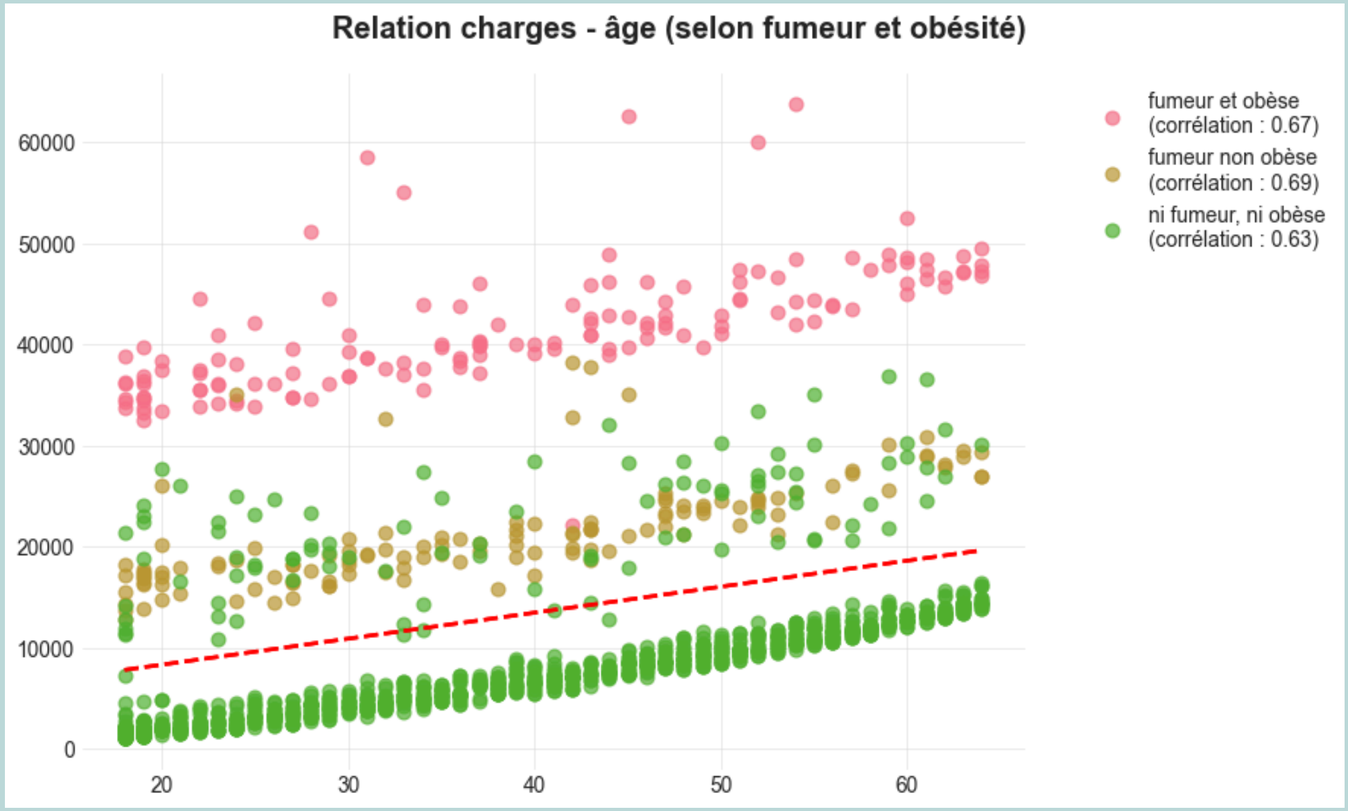
- Disparité dans les valeurs supérieures due à l'influence d'autres variables (fumeur + BMI)

Relation charges - âge selon fumeur et obésité

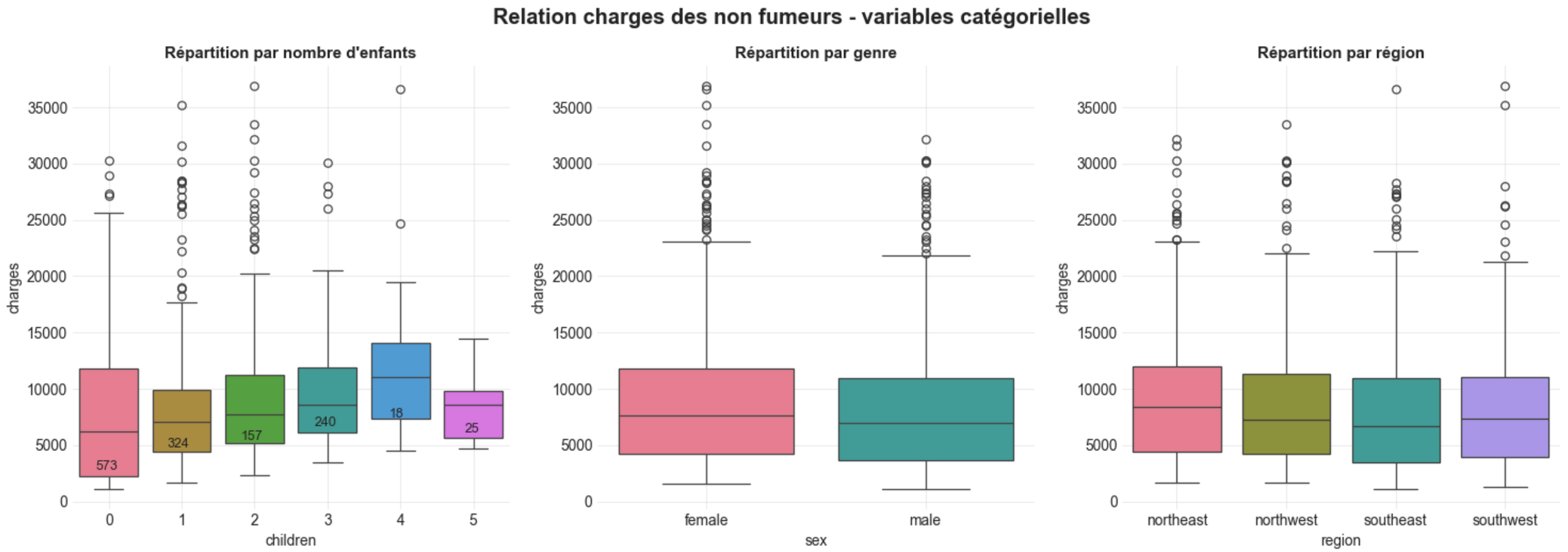
- 3 groupes de charges selon fumeur et obésité (BMI >= 30)
- Pour chaque groupe : même motif + augmentation de la corrélation âge-charges

↓

Forte corrélation linéaire entre l'âge et les charges après application des paramètres fumeur/obèse



Variables à faible impact



- Nombre d'enfants : léger impact sur les charges des non fumeurs
- Augmentation linéaire (sauf 5)
- 4-5 enfants : peu de données

- Genre : léger impact sur les charges des non fumeurs
- Charges plus élevées chez les femmes
- Idem pour les fumeurs

- Régions : pas de lien avec les charges ni avec les autres variables

Plan de modélisation

Feature engineering	Interaction smoker × BMI Transformation logarithmique des charges	→ Capture l'effet multiplicateur identifié → Réduit la forte asymétrie
Pré-traitement	Encodage One-Hot Standardisation	→ Variable région transformée en numérique → Variables numériques (age, bmi, children) normalisées via StandardScalers
Split stratifié	Respect de la proportion de fumeurs (20 %) dans train et test sets	→ Évite le biais dû au déséquilibre

Merci pour votre attention !

Des questions ?