Université de Grenoble Alpes

Projet SAS

Prédiction des coûts d'assurance médicale

Auteurs :

N'DOYE El Hadrami
RAMDÉ Ismaïl

Master 1 Statistique et Science des Données
19 mars 2021

Table des matières

Table des figures

Liste des tableaux

1 Contexte et Objectifs

Dans le cadre du module de formation logiciel spécialisé (SAS), nous nous intéressons à l'étude d'un ensemble de données personnelles sur les coûts médicaux dans 4 régions des USA. En effet, de cet ensemble de données découle une problématique qui est de prévoir avec le plus de précision possible les coûts d'assurance. Notre objectif général sera de prédire les coûts d'assurance à l'aide de la régression linéaire multiple. Nos objectifs spécifiques sont les suivants :

- Comprendre le mieux possible nos données
- Prédire le coût d'assurance d'une personne
- Améliorer le score

Dans les lignes qui suivent nous commencerons par visualiser nos données, ensuite faire un pré-traitement en passant par des statistiques de bases et en fin nous utiliserons un modèle de régression linéaire pour prédire les coûts d'assurance.

2 Dictionnaire de données

Notre jeu de donnée est issu de Kaggle https ://www.kaggle.com/mirichoi0218/insurance et téléchargé le 11/01/2021. Il est composé de 07 variables définis de la manière suivante :

age :

- Définition : âge du bénéficiaire principal
- Granularité : âge
- Type : integer

sex :

- Définition : sexe de l'entrepreneur en assurance
- Granularité : femme, homme
- Type : string

bmi :

- Définition : indice de masse corporelle, permettant de comprendre le corps, poids relativement élevé ou faible du bénéficiaire
- Type : float
- Unité : Kg/m^2

children :

- Définition : nombre d'enfants couverts par l'assurance maladie / nombre de personnes à charge
- Type : integer

smoker :

- Définition : fumeur
- Granularité : oui, non
- Type : booléan

region :

- Définition : la zone résidentielle du bénéficiaire aux États-Unis
- Granularité : nord-est, sud-est, sud-ouest, nord-ouest
- Type : string

charges (target variable) :

- Définition : Frais médicaux individuels facturés par l'assurance maladie
- Type : float
- Unité : dollar

3 Statistiques descriptives

3.1 Lecture et préparation du jeux de données

Nom de la table	MABIBLIO.DATA	Observations	1338
Type de membre	DATA	Variables	7
Moteur	V9	Index	0
Créée	11/03/2021 18:10:50	Longueur d'observation	56
Dernière modification	11/03/2021 18:10:50	Observations supprimées	0
Protection		Compressée	NON
Type de table		Triée	NON
Libellé			
Représentation des données	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Codage	utf-8 Unicode (UTF-8)		

Liste alphabétique des variables et des attributs					
#	Variable	Type	Long.	Format	Informat
1	age	Num.	8	BEST12.	BEST32.
3	bmi	Num.	8	BEST12.	BEST32.
7	charges	Num.	8	BEST12.	BEST32.
4	children	Num.	8	BEST12.	BEST32.
6	region	Texte	9	$9.	$9.
2	sex	Texte	6	$6.	$6.
5	smoker	Texte	3	$3.	$3.

TABLE 1 – Informations sur le jeu de données

On compte 1338 observations sur notre jeu de données dont chacune des observations est composée de 07 variables. Il y a 04 variables quantitatives et 03 variables qualitatives comme on peut le voir sur la figure ci-dessus.

3.2 Macros-programmes

A fin de nous faciliter la tâche quant à l'utilisation de certaines procédures que nous réutilisons plusieurs fois ou pouvant l'être par un autre utilisateur, nous avons implémenté des macros-programmes. Ces macros sont :

- **univariate :** pour la statistique univariée d'une variable ou plusieurs variables (dans notre cas la target variable qui est "charges").
- **correlation :** pour le tableau de corrélation des variables de choix.

- **regressionL :** pour la régression linéaire sur la variable à expliquer et les variables explicatives.

3.3 La target variable : charges

Dans notre étude la variable d'intérêt est "charges". C'est elle que nous prédirons en fonctions des autres variables dites explicatives.

Procedure Résumé

La procédure UNIVARIATE
Variable : charges

Moments			
N	1338	Somme des poids	1338
Moyenne	13270.4223	Somme des observations	17755825
Ecart-type	12110.0112	Variance	146652372
Skewness	1.51587966	Kurtosis	1.60629865
Somme des carrés non corrigée	4.31702E11	Somme des carrés corrigée	1.96074E11
Coeff Variation	91.2556586	Std Error Mean	331.067454

Mesures statistiques de base			
Location		Variabilité	
Moyenne	13270.42	Ecart-type	12110
Médiane	9382.03	Variance	146652372
Mode	1639.56	Intervalle	62649
		Ecart interquartile	11919

TABLE 2 – Statistiques de base de la variable charges

3.4 Histogramme et nuage de points

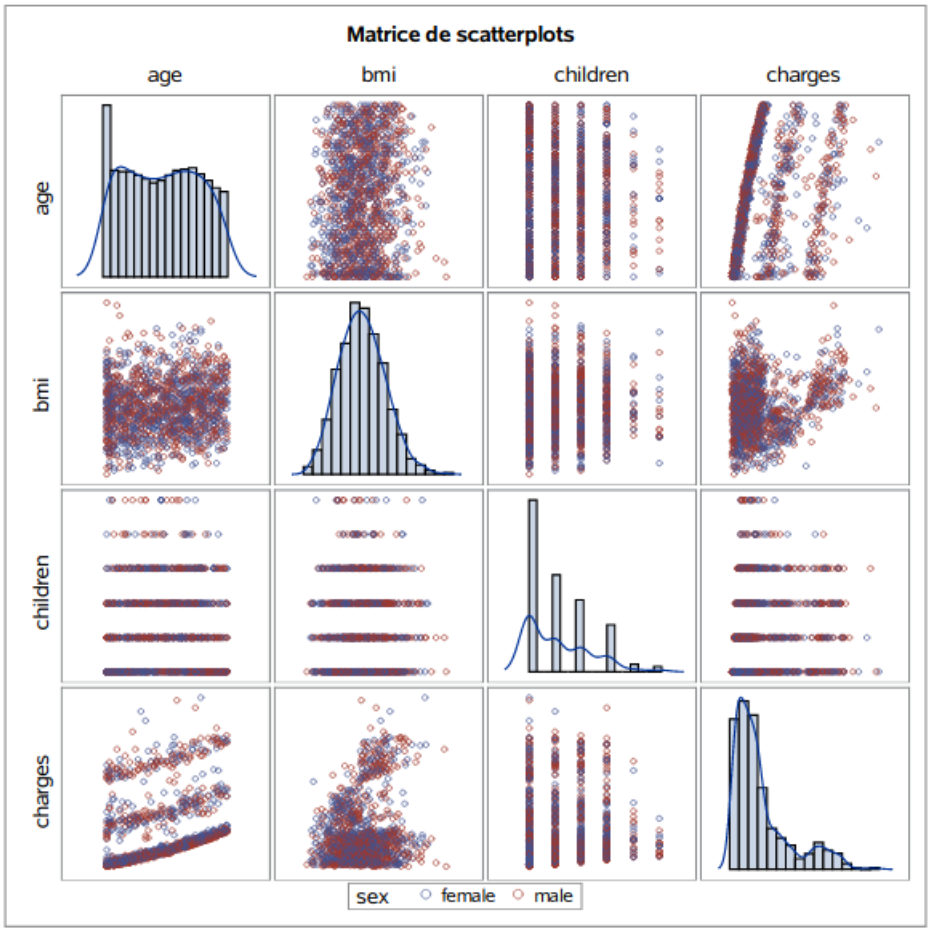

FIGURE 1 – Histogramme et nuage de points

On remarque dans un premier temps que l'indice de masse corporelle (Body Mass Index/BMI) a tendance à suivre une loi normale à travers son nuage de points et son histogramme. Dans un second temps, la forme des différents nuages de points nous montre que les variables les plus corrélées à notre variable d'intérêt(charges) sont "bmi" et "age".

3.5 Boites à moustaches

3.5.1 Variable charges en fonction du sexe

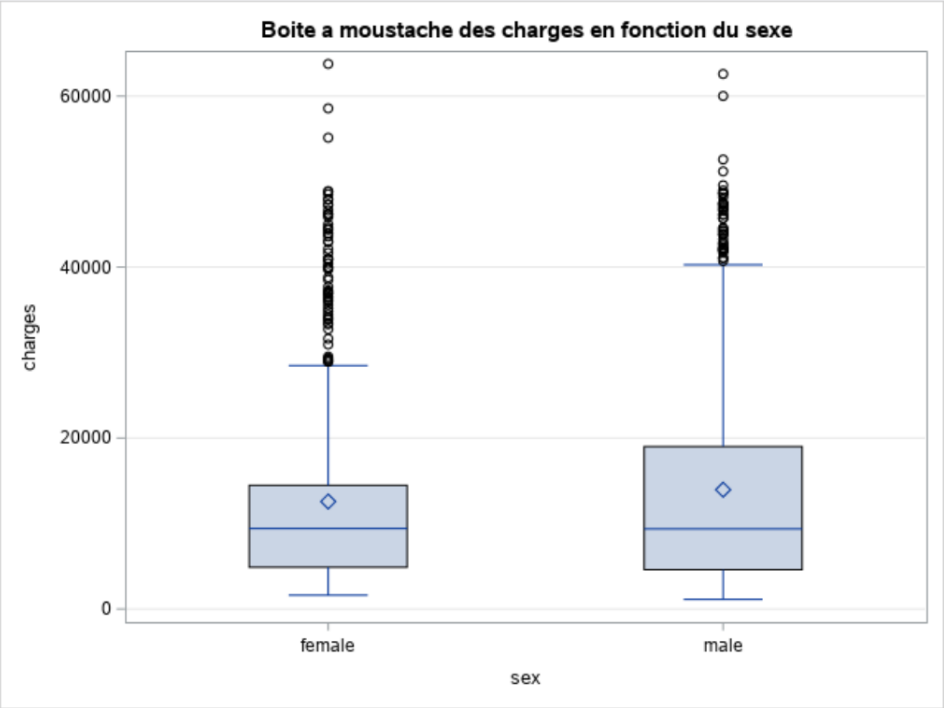

FIGURE 2 – Boite à moustache de la charges en fonction du sexe

On remarque de façon générale que les hommes ont des charges beaucoup plus élevées par rapport aux femmes (charge moyenne des hommes supérieur à celle des femmes). En effet plus de 85% des femmes ont des charges en-dessous de 20000, presque 15% entre 20000 et 30000 et des outliers qui vont de 30000 à plus de 60000. Quant aux hommes, 75% sont sous une charge de 20000, 25% entre 20000 et 40000 et des outliers situés entre 40000 et plus de 60000.

3.5.2 Variable charges en fonction du statut fumeur

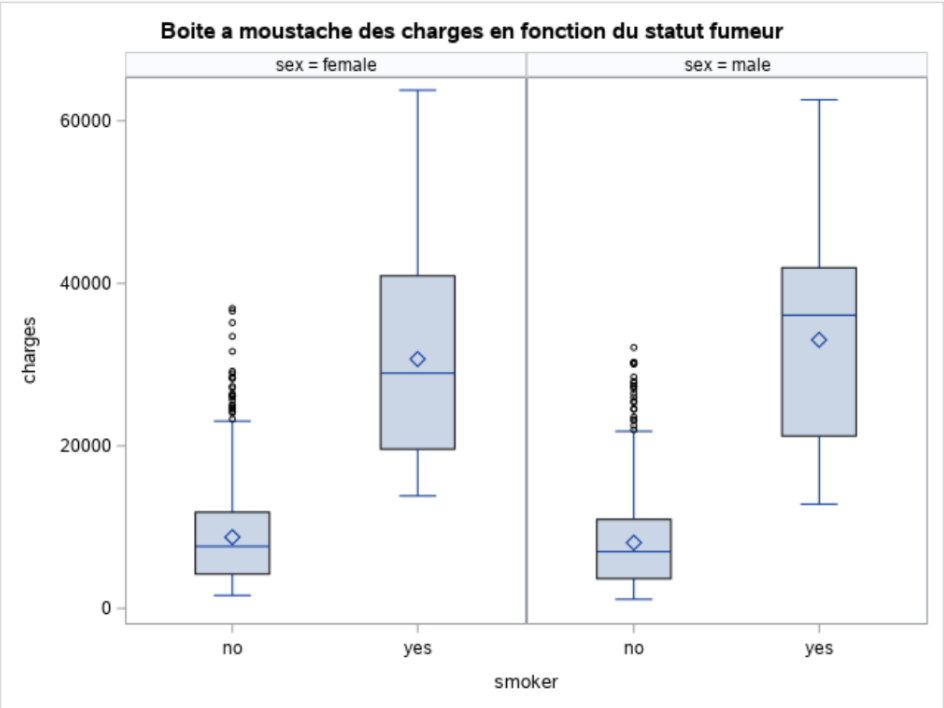

FIGURE 3 – Boite à moustache de la charges en fonction du statut fumeur

On remarque que quelque soit le sexe, les charges sont beaucoup plus élevées chez les fumeurs par rapport aux non fumeurs. On voit aussi des outliers chez les non fumeurs pour les deux sexes.

3.6 Graphiques à bulles

3.6.1 Variable charges en fonction des régions

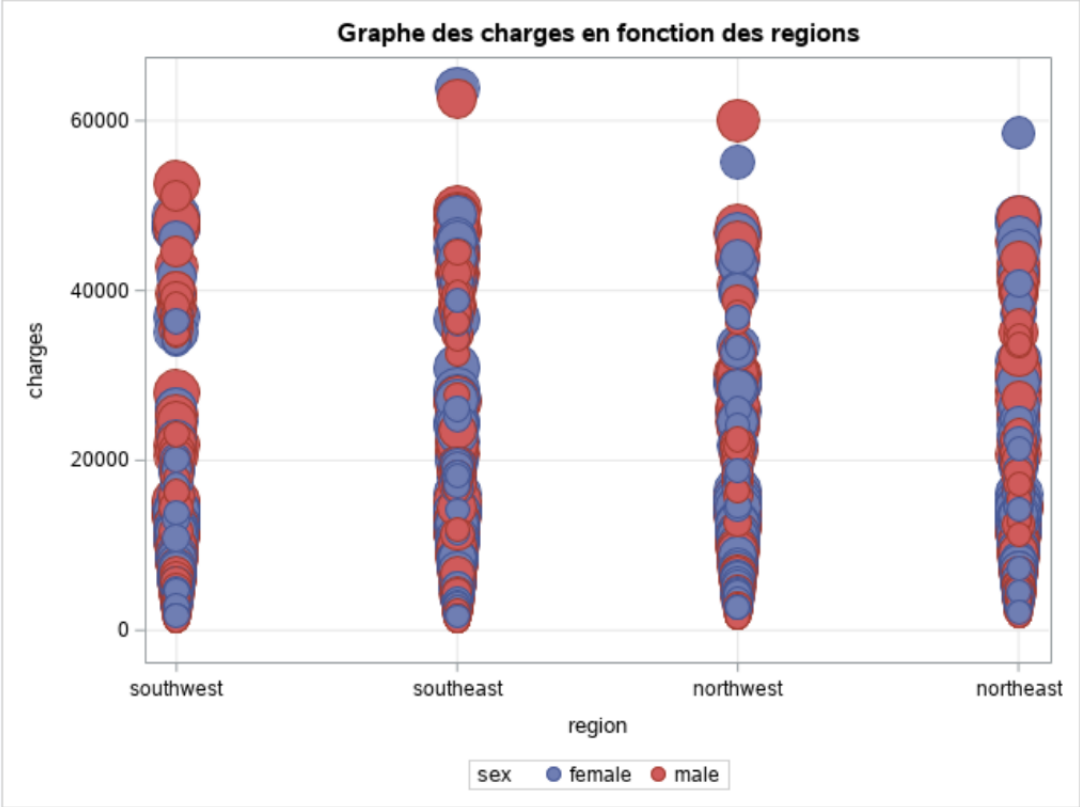

FIGURE 4 – Graphique à bulles de la charges en fonction des régions

Dans ce graphique on constate que les charges des différentes régions sont pratiquement au même niveau avec quelque légères différences. On note également la presence de valeurs aberrantes pour les régions southeast, northwest et northeast.

3.6.2 Variable charges en fonction du statut fumeur

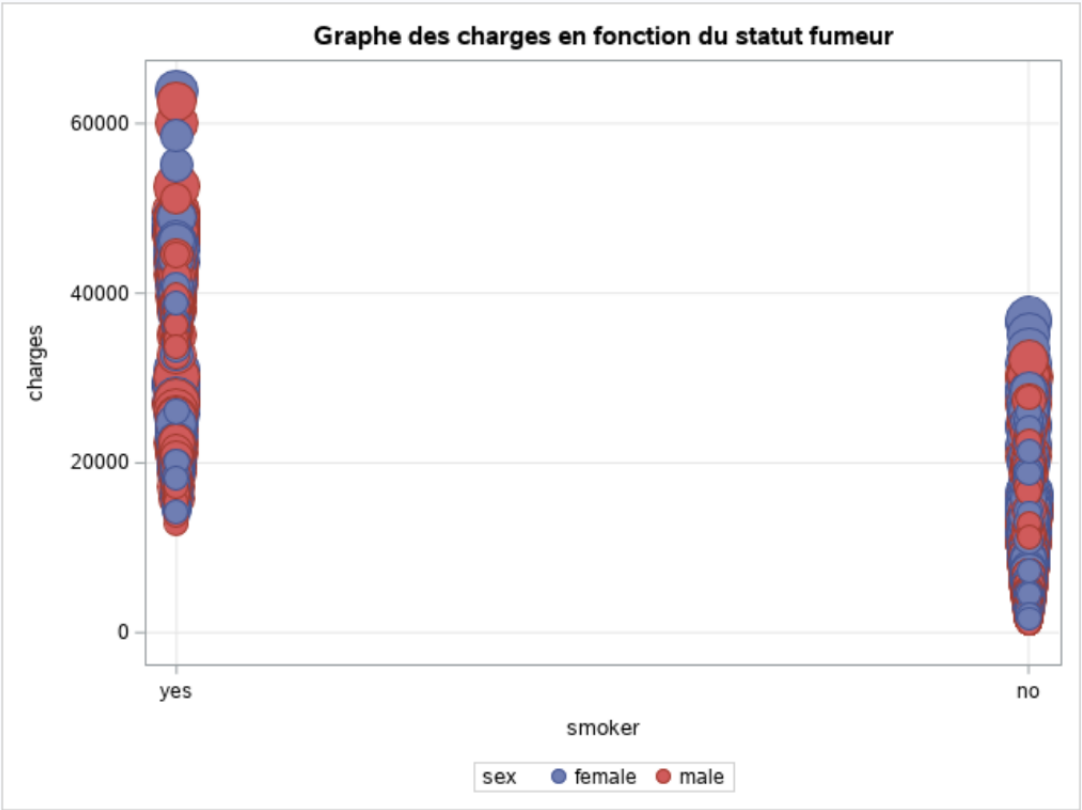

FIGURE 5 – Graphique à bulles de la charges en fonction du statut fumeur

Les charges des fumeurs sont très importants par rapport celles des non fumeurs. En effet chez les fumeurs les charges vont de 10000 à plus de 60000 et chez les non fumeurs elles vont de 0 à 40000. Il y a donc une différence significative.

3.7 Graphique en barre (barplots) des fumeurs en fonction du sexe

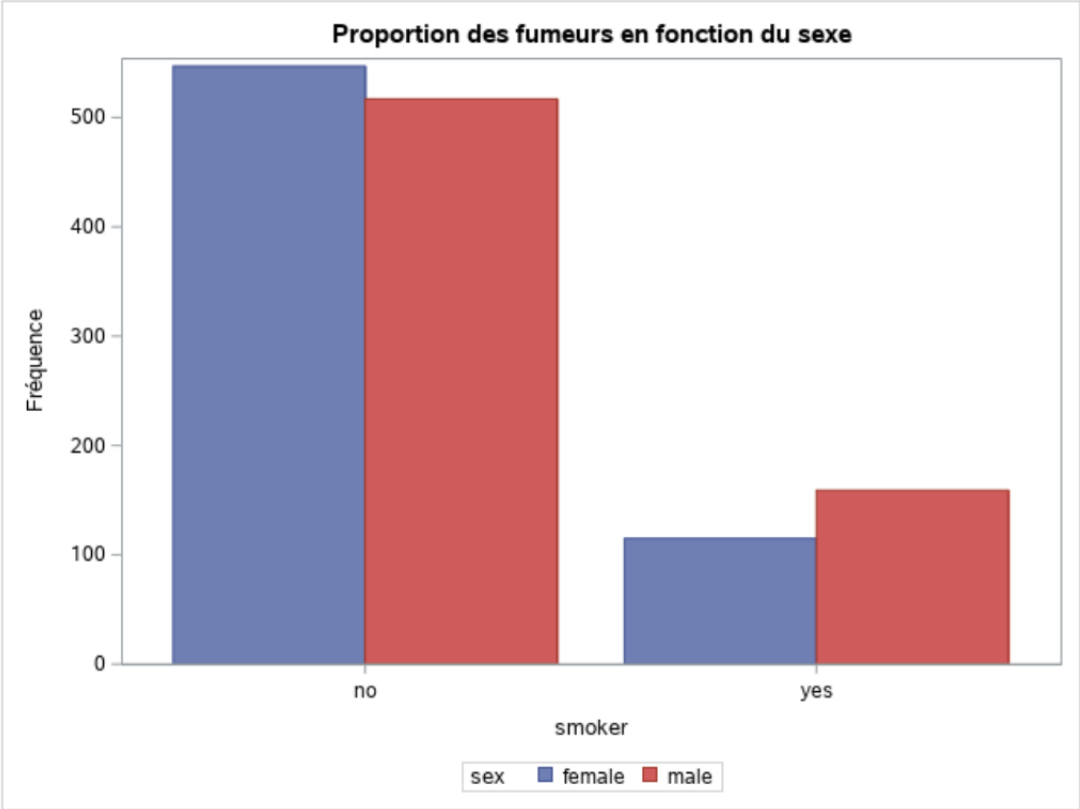

FIGURE 6 – Barplots des fumeurs en fonction du sexe

Ce graphe nous enseigne de façon globale qu'il y a moins de fumeurs par rapport au non fumeurs. Dans le groupe des fumeurs la proportion des hommes est plus élevée que celle des femmes. Et dans celui des non fumeurs c'est la proportion des femmes qui est plus élevée que celle des hommes.

4 Pré-traitement

4.1 Encodage des variables qualitatives

Pour faire un modèle de machine learning il est préférable de convertir toutes les variables qualitatives en quantitatives pour avoir une bonne prédiction. Notre jeu de données comporte trois variables qualitatives : *"sex"*, *"regions"*, *"smoker"*. À l'issu de la transformation (recodage), on obtient : *"sex_Recode"*,

"region_Recode", "smoker_Recode" comme on peut le voir dans le tableau ci-dessous.

Liste alphabétique des variables et des attributs					
#	Variable	Type	Long.	Format	Informat
1	age	Num.	8	BEST12.	BEST32.
3	bmi	Num.	8	BEST12.	BEST32.
7	charges	Num.	8	BEST12.	BEST32.
4	children	Num.	8	BEST12.	BEST32.
6	region	Texte	9	$9.	$9.
10	region_Recode	Num.	8	BEST12.	BEST32.
2	sex	Texte	6	$6.	$6.
8	sex_Recode	Num.	8	BEST12.	BEST32.
5	smoker	Texte	3	$3.	$3.
9	smoker_Recode	Num.	8	BEST12.	BEST32.

TABLE 3 – Liste alphabétique des variables et des attributs

4.2 Partition du jeu de données

Avant l'apprentissage automatique nous divisons notre jeu de données en deux parties :

- la première (train set) sera utilisée pour entraîner le modèle
- la seconde (validation set) pour tester et valider le modèle. Elle représente 30% du jeu de données total et servira pour faire des prédictions à partir du modèle choisi.

Cette manière de procéder nous permettra non seulement de vérifier la qualité de notre modèle sur le validation set mais aussi d'éviter un éventuel sur-apprentissage (Overfitting).

Resumé train set

Variable	N	Moyenne	Ec-type	Minimum	Maximum
age	401	39.1970075	14.4604146	18.0000000	64.0000000
bmi	401	30.7470948	6.3446196	17.1950000	50.3800000
children	401	1.1246883	1.2306153	0	5.0000000
charges	401	13041.57	12000.48	1131.51	63770.43
sex_Recode	401	0.5087282	0.5005483	0	1.0000000
smoker_Recode	401	0.1895262	0.3924156	0	1.0000000
region_Recode	401	1.4638404	1.1132449	0	3.0000000

Resumé validation set

Variable	N	Moyenne	Ec-type	Minimum	Maximum
age	401	39.1970075	14.4604146	18.0000000	64.0000000
bmi	401	30.7470948	6.3446196	17.1950000	50.3800000
children	401	1.1246883	1.2306153	0	5.0000000
charges	401	13041.57	12000.48	1131.51	63770.43
sex_Recode	401	0.5087282	0.5005483	0	1.0000000
smoker_Recode	401	0.1895262	0.3924156	0	1.0000000
region_Recode	401	1.4638404	1.1132449	0	3.0000000

TABLE 4 – Train-test/validation-set

5 Analyse statistique et modélisation

5.1 Corrélations

Coefficients de corrélation de Pearson, N = 1338 Proba > |r| sous H0: Rho=0							
	age	bmi	children	charges	sex_Recode	smoker_Recode	region_Recode
age	1.00000	0.10927 <.0001	0.04247 0.1205	0.29901 <.0001	-0.02086 0.4459	-0.02502 0.3605	-0.00213 0.9380
bmi	0.10927 <.0001	1.00000	0.01276 0.6410	0.19834 <.0001	0.04637 0.0900	0.00375 0.8910	-0.15757 <.0001
children	0.04247 0.1205	0.01276 0.6410	1.00000	0.06800 0.0129	0.01716 0.5305	0.00767 0.7792	-0.01657 0.5448
charges	0.29901 <.0001	0.19834 <.0001	0.06800 0.0129	1.00000	0.05729 0.0361	0.78725 <.0001	0.00621 0.8205
sex_Recode	-0.02086 0.4459	0.04637 0.0900	0.01716 0.5305	0.05729 0.0361	1.00000	0.07618 0.0053	-0.00459 0.8668
smoker_Recode	-0.02502 0.3605	0.00375 0.8910	0.00767 0.7792	0.78725 <.0001	0.07618 0.0053	1.00000	0.00218 0.9365
region_Recode	-0.00213 0.9380	-0.15757 <.0001	-0.01657 0.5448	0.00621 0.8205	-0.00459 0.8668	0.00218 0.9365	1.00000

TABLE 5 – Tableau de corrélation entre les différentes variables

Sur cette table on voit les différentes corrélations possibles entre les variables. Toutes les variables explicatives sont corrélées positivement avec la variable à expliquer (charges). La variable qui a la plus forte corrélation positive avec la variable charge est "smoker" (r= 0.7825).

5.2 Régression linéaire multiple (train set) avec toutes les variables explicatives

Après avoir fait le pré-traitement nous allons choisir un modèle de régression linéaire multiple entre la variable a expliquer (charges) avec le reste des variables dites explicatives.
La formule théorique du modèle de régression linéaire multiple s'écrit de la manière suivante :

$$Y_{charges} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \cdots + \beta_p X_n + \epsilon$$

Avec :
$Y_{charges}$: Représente la variable a expliquer
β_0 : Représente l'intercepte ou constante de régression

β_i : Représente l'effet du i-ème variable explicative sur la variable $Y_{charges}$
X_i : Représente le i-ème variable explicative
ϵ : Représente l'erreur résiduelle.

Root MSE	6065.81954	R carré	0.7528
Moyenne dépendante	13368	R car. ajust.	0.7512
Coeff Var	45.37444		

Paramètres estimés					
Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	-13145	1219.01659	-10.78	<.0001
age	1	255.23813	14.38934	17.74	<.0001
bmi	1	337.72508	33.69372	10.02	<.0001
children	1	461.41480	166.12222	2.78	0.0056
smoker_Recode	1	23838	486.52317	49.00	<.0001
sex_Recode	1	-65.69378	398.15304	-0.16	0.8690
region_Recode	1	440.94185	182.08884	2.42	0.0156

TABLE 6 – Résultat du modèle de régression linéaire

Le résultat du modèle de régression linéaire avec toutes les variables donne un R^2 de 0.7528 soit 75.28% ce qui montre que le modèle est assez bon.
La statistique du test de la variable "sex_ Recode" donne une p-valeur supérieur à 5% , donc cette variable n'a pas d'effet significative sur la variable a expliquer (charges). Par la suite nous effectuerons une sélection de variable pour conserver les variables qui ont plus d'effets (significatives).

5.3 Sélection de variables (train set)

Une sélection de variable avec la méthode AIC donne le résultat suivant :

Synthèse des sélections Stepwise				
Etape	Effet saisi	Effet supprimé	Nombre d'effets dans	AIC
0	Intercept		1	18566.9229
1	smoker_Recode		2	17644.6371
2	age		3	17365.9484
3	bmi		4	17276.6759
4	children		5	17271.2065
5	region_Recode		6	17267.3375*
* Valeur optimale du critère				

TABLE 7 – Sélection de variables

La sélection de variables par la méthode AIC conserve les variables explicatives qui ont plus d'effets sur la variable a expliquer (charges) comme on peut le voir dans le tableau ci-dessus. Les variables qui expliquent au mieux la variable charges sont : **smoker_Recode, age, bmi, children et region_Recode**.

5.4 Régression linéaire multiple avec les variables sélectionnées

Root MSE	6062.64971	R carré	0.7528
Moyenne dépendante	13368	R car. ajust.	0.7515
Coeff Var	45.35073		

Paramètres estimés					
Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	-13163	1213.55831	-10.85	<.0001
age	1	255.22112	14.38145	17.75	<.0001
bmi	1	337.34118	33.59572	10.04	<.0001
children	1	461.17290	166.02894	2.78	0.0056
smoker_Recode	1	23833	485.41701	49.10	<.0001
region_Recode	1	439.91953	181.88829	2.42	0.0158

TABLE 8 – Régression linéaire multiple avec les variables sélectionnées

Cette régression linéaire multiple donne un R^2 de 0.7528 semblable à celui obtenu avec toutes les variables. On peut conclure que le modèle obtenu après sélection de variables est meilleur même si le R^2 reste tel. Ce modèle minimise l'erreur quadratique moyenne (MSE) et réduit le nombre de variables. Il faut aussi noter que cela permettrai d'optimiser la mémoire ainsi que le temps d'exécution.

5.5 Modèle de régression prédictif (validation set)

Après avoir modélisé notre problème on peut maintenant prédire les nouvelles valeurs de charges sur lesquelles les individus des différentes régions sont confrontés. C'est seulement ici qu'intervient la deuxième partie du jeu de données : le validation set.
Le tableau ci dessous donne la prédiction des charges des 20 premiers individus. Cette prédiction est efficace à 75%.

Obs.	age	bmi	smoker	region	p_charges
1	46	33.44	no	southeast	11089.33
2	37	27.74	no	northwest	7979.33
3	52	30.78	no	northeast	11819.04
4	30	35.3	yes	southwest	30769.71
5	60	36.005	no	northeast	15031.89
6	30	32.4	no	southwest	6591.19
7	18	34.1	no	southeast	3483.16
8	34	31.92	yes	northeast	31264.65
9	59	27.72	no	southeast	13709.27
10	63	23.085	no	northeast	11759.78
11	23	17.385	no	northwest	54.29
12	31	36.3	yes	southwest	32369.55
13	19	28.6	no	southwest	4581.25
14	19	20.425	no	northwest	-547.35
15	62	32.965	no	northwest	16137.37
16	31	36.63	no	southeast	8691.96
17	18	38.665	no	northeast	5940.97
18	60	24.53	no	southeast	11431.01
19	36	35.2	yes	southeast	32815.40
20	36	34.43	yes	southeast	32061.12

TABLE 9 – Prédictions sur le validation set

6 Conclusion

Au terme de notre étude qui portait sur la prédiction des charges d'assurance en fonction de diverses variables notamment l'âge, le sexe, l'indice de masse corporelle, le nombre d'enfants, le statut fumeur et la région, a été faite sur la base d'un modèle de régression linéaire multiple. Bien avant cela, nous avons pratiqué des statistiques descriptives pour comprendre et aborder au mieux notre étude. Nous avons aussi utilisé le tableau de corrélation entre les variables et une méthode de sélection de variables (AIC) pour retenir les variables qui expliquaient le plus notre "target" variable (charges). Le modèle final ainsi obtenu est : **charges = age + bmi + smoker + region**.
À l'issu de tout cela, nous avons obtenu un score R^2 globalement bon qui est de 0.7528. Toutefois, notre étude n'a porté que sur deux méthodes à savoir la régression linéaire multiple classique et la sélection de variables. Ce travail que nous avons réalisé pourrait donc être complété et poursuivi avec d'autres méthodes notamment la validation croisée.