PROJET D'ANALYSE

COÛTS MÉDICAUX

Comment votre profil affecte vos frais médicaux?

SOMMAIRE

majdi_hannachi@outlook.fr

CONTEXTE

Pour qu'une société d'assurance gagne de l'argent, elle doit en collecter davantage chaque année, des primes qu'elle dépense sur les soins médicaux à ses bénéficiaires. En conséquence, les assureurs investissent beaucoup de temps et d'argent pour développer des modèles qui **prédisent** avec précision les **frais médicaux**.

Les frais médicaux sont difficiles à estimer car les conditions les plus coûteuses sont rares et apparemment aléatoires. Néanmoins, certaines conditions prévalent davantage pour certains segments de la population. Par exemple, le cancer du poumon est plus probable chez les fumeurs que chez les non-fumeurs et les maladies cardiaques chez les obèses.

Dans le cadre d'un projet de fin de semestre, en matière d'analyse, on va mener une analyse sur ensemble de données sur les patients inscrits sous une firme d'assurance.

Le but de cette analyse est d'utiliser les données des patients pour estimer la moyenne des dépenses de soins médicaux pour ces segments de population. Ces estimations pourraient être utilisées pour créer des tables actuarielles fixant le prix des primes annuelles à la hausse ou à la baisse en fonction des coûts de traitement attendus.

Institut supérieur des études technologiques de Radès | majdi_hannachi@outlook.fr

SOURCE DES DONNÉES

Pour cette analyse, nous utiliserons un jeu de données simulé contenant les frais médicaux des patients aux États-Unis. Ces données ont été créées en utilisant les statistiques démographiques du **Bureau de recensement américain** et reflètent donc approximativement les conditions du monde réel.

Les données sont présentés en fichier csv 'insurance.csv' de taille 53.3 Ko téléchargé du site **Github,** et publié sur la plateforme du DataScience **Kaggle,** dont on a **1338 lignes et 7 colonnes** indiquant les caractéristiques du patient ainsi que le total des **frais médicaux imputés** au régime pour l'année civile.

kaggle

Les variables sont comme suit:

- **age**: Il s'agit d'un nombre entier indiquant l'âge du premier bénéficiaire (à l'exclusion des personnes de plus de 64 ans, car elles sont généralement couvertes par le gouvernement).
- **sex**: il s'agit du sexe du titulaire de la police, homme ou femme.
- **bmi**: il s'agit de l'indice de masse corporelle (iMC), qui donne une idée de la densité ou de l'insuffisance pondérale d'une personne par rapport à sa taille. L'IMC est égal au poids (en kilogrammes) divisé par la taille (en mètres) au carré. Un iMC idéal se situe entre 18,5 et 24,9.
- **children**: Il s'agit d'un nombre entier indiquant le nombre d'enfants / de personnes à charge couverts par le régime d'assurance.
- **smoker**: oui ou non selon que l'assuré fume régulièrement du tabac.
- **region**: il s'agit du lieu de résidence du bénéficiaire aux États-Unis, divisé en quatre régions géographiques: nord-est, sud-est, sud-ouest ou nord-ouest.
- **charges**: Frais médicaux individuels facturés par l'assurance maladie

ENVIRONNEMENT DE TRAVAIL

1- Equipe et équipements

Majdi Hannachi
Etudiant en Master Business Intelligence

Lenovo

Ram 4 Go
Disque dur 1 To
Processeur 6éme generation i5

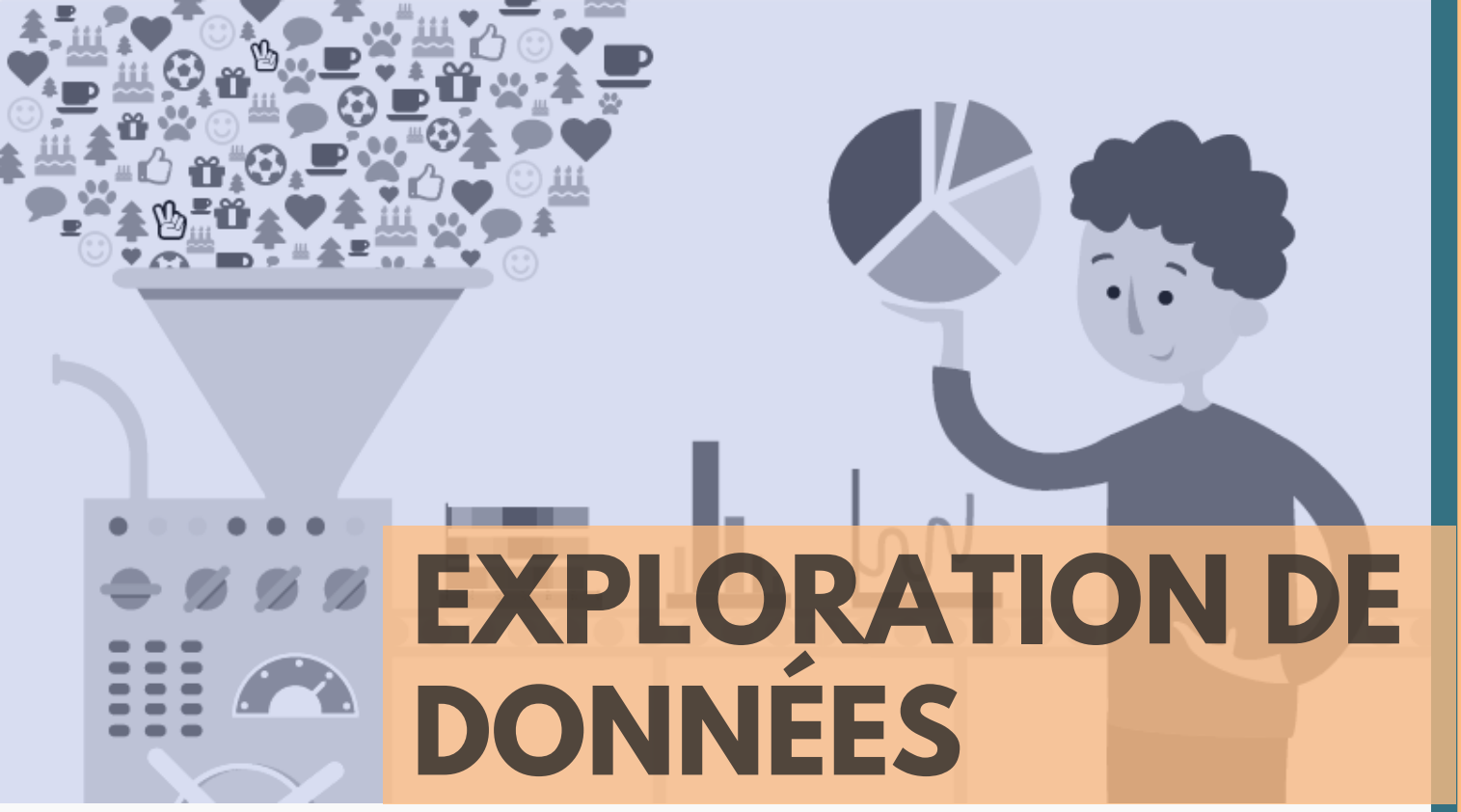
2- Logiciel:

R est un langage de programmation dont le but est de pouvoir traiter et organiser **des jeux de données** afin de pouvoir y appliquer des tests **statistiques** plus ou moins complexes et se représenter ces données graphiquement à l'aide d'une grande variétés de graphiques disponibles.

RStudio est un environnement de développement gratuit, libre et multiplateforme pour R, un langage de programmation utilisé pour le traitement de données et l'analyse statistique.

RStudio®

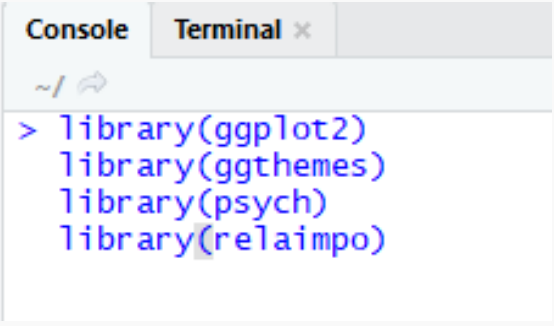
Institut supérieur des études technologiques de Radès | majdi_hannachi@outlook.fr



EXPLORATION DE DONNÉES

1- CHARGEMENT DES LIBRAIRIES

- **ggplot2** est une librairie R de visualisation de données.
- **ggthemes** fournit des thèmes et des échelles 'ggplot2' qui reproduisent l'aspect des parcelles.



- **psych** est une boîte à outils polyvalente pour la personnalité, la théorie psychométrique et la psychologie expérimentale. Les fonctions sont principalement destinées à l'analyse multivariée et à la construction d'échelles à l'aide d'une analyse factorielle, d'une analyse en composantes principales, d'une analyse par grappes et d'une analyse de fiabilité, bien que d'autres fournissent des statistiques descriptives de base.
- **relaimpo** fournit plusieurs mesures permettant d'évaluer l'importance relative dans les modèles linéaires.

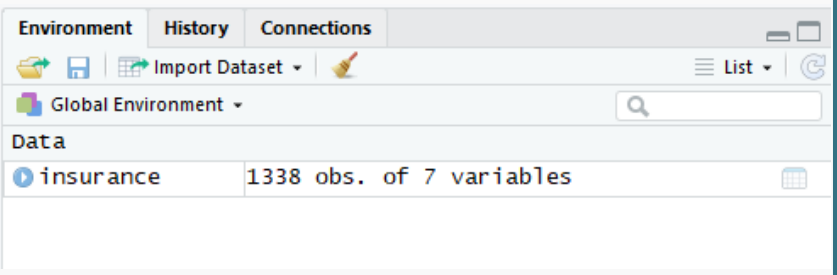
Institut supérieur des études technologiques de Radès | majdi_hannachi@outlook.fr

2- CHARGEMENT DES DONNÉES

```
Console  Terminal
~/
> insurance <- read.csv("d:/insurance.csv")
```

Le chargement du données se fait grace à la fonction **'read.csv'**

L'interface de l'environnement nous indique que les données sont chargés et que le résultat obtenu est de 1338 lignes et 7 colonnes.




```
Console  Terminal
~/
> view(insurance)
```

On peut visualiser tout les données en double cliquant sur résultat obtenu par l'environnement ou à l'aide du commande '**view()**'.

insurance

Filter

	age	sex	bmi	children	smoker	region	charges
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31	female	25.740	0	no	southeast	3756.622
7	46	female	33.440	1	no	southeast	8240.590
8	37	female	27.740	3	no	northwest	7281.506
9	37	male	29.830	2	no	northeast	6406.411
10	60	female	25.840	0	no	northwest	28923.137
11	25	male	26.220	0	no	northeast	2721.321
12	62	female	26.290	0	yes	southeast	27808.725

Showing 1 to 12 of 1,338 entries

3- ANALYSE DESCRIPTIVE:

Pour mener une étude descriptive sur les données on doit avoir une idée sur les différentes variables utilisées par la fonction '**summary()**'.

```
Console  Terminal
~/
> summary(insurance)
      age            sex           bmi           children     smoker          region       charges
 Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064   northeast:324   Min.   : 1122
 1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274   northwest:325   1st Qu.: 4740
 Median :39.00                Median :30.40   Median :1.000              southeast:364   Median : 9382
 Mean   :39.21                Mean   :30.66   Mean   :1.095              southwest:325   Mean   :13270
 3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000                              3rd Qu.:16640
 Max.   :64.00                Max.   :53.13   Max.   :5.000                              Max.   :63770
> |
```

Le sexe et la région d'origine des répondants sont équitablement répartis et ont un âge compris entre 18 et 64 ans. Les non-fumeurs sont 4 fois plus nombreux que les fumeurs. Le coût moyen des soins médicaux est de 13 270 $ et la valeur médiane de 9 382 $.

Maintenant on va comparer le coût dans les différents population des variable, ceci est accordé par la fonction '**describeby()**' et **la boîte à moustaches** qui résume seulement quelques caractéristiques de position du caractère étudié (médiane, quartiles, minimum, maximum ou déciles).

a- Région et coûts

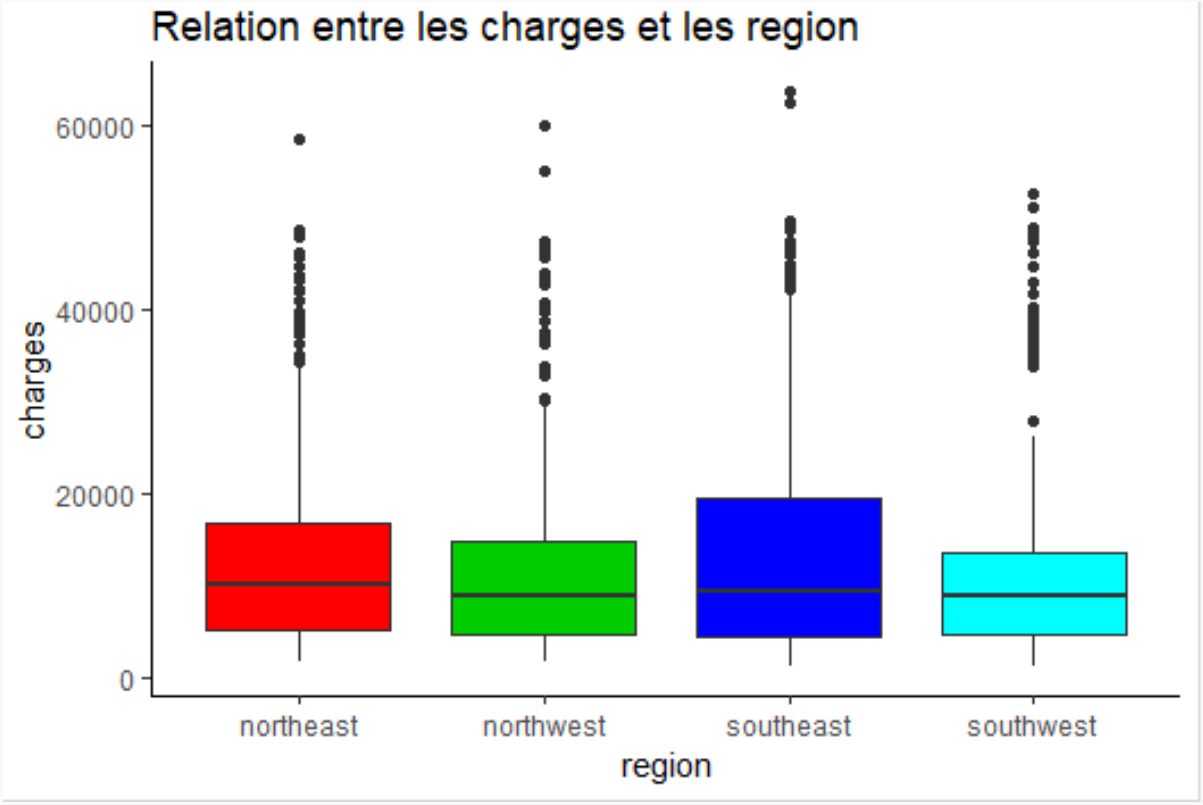
```
Console  Terminal
~/
> describeBy(insurance$charges,insurance$region)

 Descriptive statistics by group
group: northeast
   vars   n     mean      sd   median  trimmed     mad    min      max    range skew kurtosis     se
X1    1 324 13406.38 11255.8 10057.65 11444.31 7806.78 1694.8 58571.07 56876.28 1.48     1.68 625.32
-------------------------------------------------------------------------------------------------
group: northwest
   vars   n     mean       sd median  trimmed     mad     min      max    range skew kurtosis     se
X1    1 325 12417.58 11072.28 8965.8 10414.54 7001.14 1621.34 60021.4 58400.06 1.67     2.53 614.18
-------------------------------------------------------------------------------------------------
group: southeast
   vars   n     mean      sd  median  trimmed     mad     min      max    range skew kurtosis     se
X1    1 364 14735.41 13971.1 9294.13 12563.65 8749.51 1121.87 63770.43 62648.55 1.24     0.48 732.28
-------------------------------------------------------------------------------------------------
group: southwest
   vars   n     mean       sd  median  trimmed     mad     min      max    range skew kurtosis     se
X1    1 325 12346.94 11557.18 8798.59 10120.52 6329.39 1241.57 52590.83 51349.26 1.67     2.03 641.08
```

La variable région ait 4 modalités; northest, nothwest, southeast, southwest. Les populations de l'Est payent des charges plus que 1 000$ que l'Ouest, ceci est montré par la différence entre les moyennes.

Institut supérieur des études technologiques de Radès | majdi_hannachi@outlook.fr

```
Console  Terminal
~/
> ggplot(data = insurance,aes(region,charges)) + geom_boxplot(fill = c(2:5)) +
      theme_classic() + ggtitle("Relation entre les charges et les region")
```



Sur la base du graphique ci-dessus, nous pouvons indiquer que la région d'origine n'a pas beaucoup d'impact sur le montant des frais médicaux.

b- Statut de fumeur et coûts

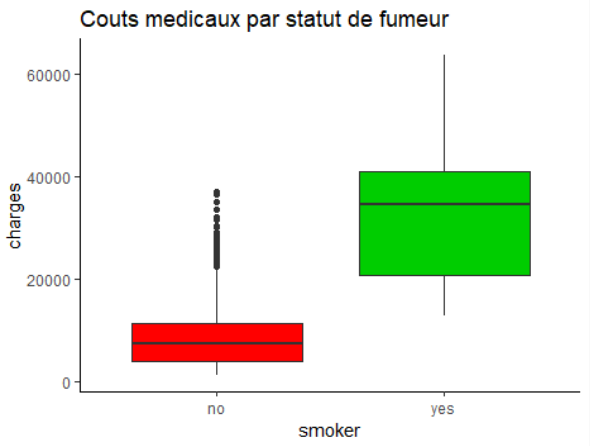
```
Console  Terminal
~/
> describeBy(insurance$charges,insurance$smoker)

 Descriptive statistics by group
group: no
   vars    n    mean      sd  median trimmed     mad     min      max    range skew kurtosis
X1    1 1064 8434.27 5993.78 7345.41 7599.76 5477.15 1121.87 36910.61 35788.73 1.53     3.12
      se
X1 183.75
------------------------------------------------------------------------
group: yes
   vars   n     mean       sd   median  trimmed      mad      min      max    range skew
X1    1 274 32050.23 11541.55 34456.35 31782.89 15167.19 12829.46 63770.43 50940.97 0.13
   kurtosis     se
X1    -1.05 697.25
```


Institut supérieur des études technologiques de Radès I majdi_hannachi@outlook.fr

```
Console  Terminal
~/
> ggplot(data = insurance,aes(smoker,charges)) + geom_boxplot(fill = c(2:3)) +
+     theme_classic() + ggtitle("Couts medicaux par statut de fumeur")
```

Par contre, on ne peut pas en dire autant du statut de fumeur. On peut clairement tromper que les fumeurs dépensent près de 4 fois plus en frais médicaux que les non-fumeurs.



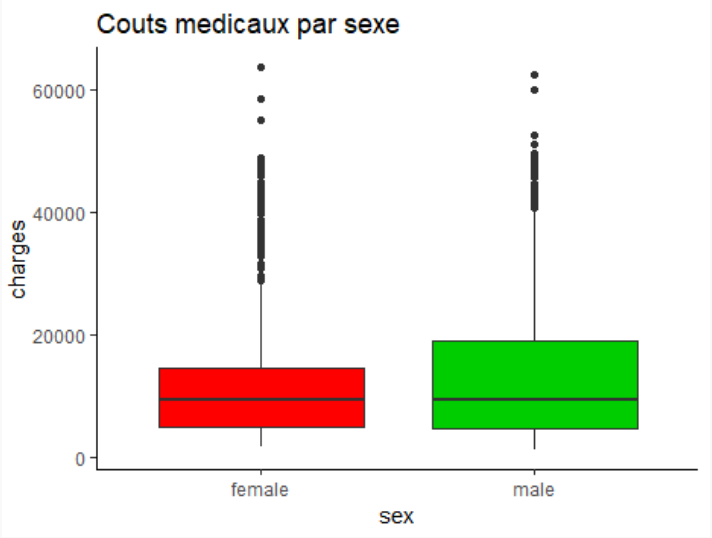
c- Sexe et Coûts

```
Console  Terminal
~/
> describeBy(insurance$charges,insurance$sex)

 Descriptive statistics by group
group: female
   vars   n     mean      sd  median  trimmed     mad     min      max    range skew kurtosis     se
X1    1 662 12569.58 11128.7 9412.96 10455.16 7129.08 1607.51 63770.43 62162.92 1.72     2.71 432.53
------------------------------------------------------------------------------------------------
group: male
   vars   n     mean      sd  median trimmed     mad     min      max range skew kurtosis     se
X1    1 676 13956.75 12971.03 9369.62 11825.4 8121.53 1121.87 62592.87 61471 1.33     0.79 498.89
```

```
Console  Terminal
~/
> ggplot(data = insurance,aes(sex,charges)) + geom_boxplot(fill = c(2:3)) +
+     theme_classic() + ggtitle("Couts medicaux par sexe")
```

Les frais médicaux ne semblent pas non plus être affectés par le sexe. Au moyenne les hommes payent plus de 1 400 $ que les femmes.



d- Nombre d'enfants et Coûts

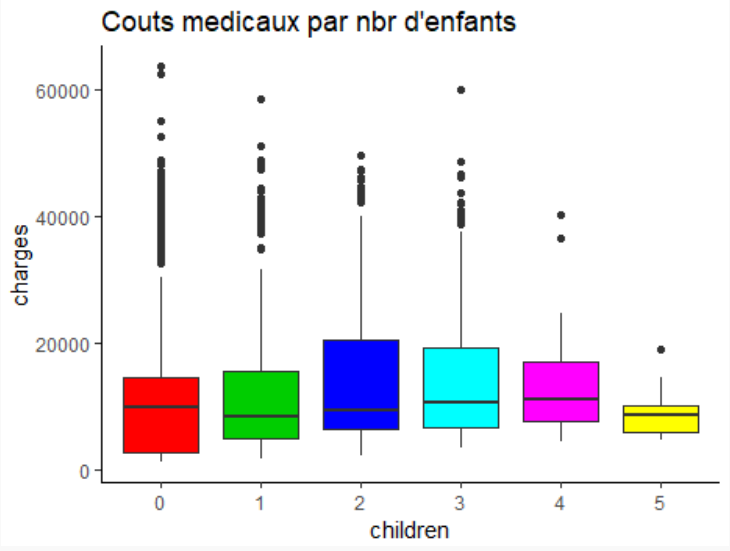
Comme déjà mentionné dans l'analyse descriptive, le nombre des enfants est compris entre 0 et 5, donc on obtient 6 population de la façon suivante:

```
Console  Terminal
~/
> describeBy(insurance$charges,insurance$children)

 Descriptive statistics by group
group: 0
   vars   n     mean       sd  median  trimmed      mad     min      max    range skew kurtosis
X1    1 574 12365.98 12023.29 9856.95 10155.21 10067.29 1121.87 63770.43 62648.55 1.53     1.95
       se
X1 501.84
------------------------------------------------------------------------------
group: 1
   vars   n     mean       sd  median trimmed     mad     min      max    range skew kurtosis     se
X1    1 324 12731.17 11823.63 8483.87 10364.8 5859.46 1711.03 58571.07 56860.05 1.66     1.97 656.87
------------------------------------------------------------------------------
group: 2
   vars   n     mean       sd  median  trimmed     mad  min      max    range skew kurtosis     se
X1    1 240 15073.56 12891.37 9264.98 12895.82 6587.43 2304 49577.66 47273.66 1.28     0.35 832.13
------------------------------------------------------------------------------
group: 3
   vars   n     mean       sd   median  trimmed     mad     min     max    range skew kurtosis     se
X1    1 157 15355.32 12330.87 10600.55 13220.71 6918.06 3443.06 60021.4 56578.33 1.45     1.21 984.11
------------------------------------------------------------------------------
group: 4
   vars  n     mean      sd   median  trimmed    mad     min      max    range skew kurtosis      se
X1    1 25 13850.66 9139.22 11033.66 12401.81 7109.3 4504.66 40182.25 35677.58 1.45     1.59 1827.84
------------------------------------------------------------------------------
group: 5
   vars  n    mean      sd  median trimmed     mad    min      max    range skew kurtosis     se
X1    1 18 8786.04 3808.44 8589.57 8402.35 3631.71 4687.8 19023.26 14335.46 1.04     0.54 897.66
```

```
Console  Terminal
~/
> ggplot(data = insurance,aes(as.factor(children),charges)) + geom_boxplot(fill = c(2:7)) +
+     theme_classic() +  xlab("children") +
+     ggtitle("Couts medicaux par nbr d'enfants")
```

Les personnes avec 5 enfants ont en moyenne moins de dépenses médicales par rapport aux autres groupes. Hors la valeur maximale est accordée par un individu qui n'a pas d'enfants.



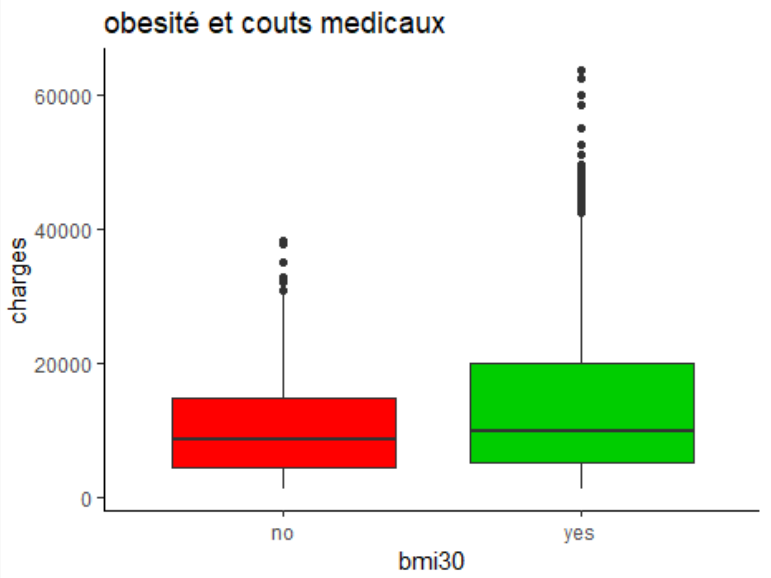
e- Obésité et Coûts

On a sous-titré le BMI en obésité car nous cherchons avec l'indice de masse corporelle d'identifier l'obésité, 30 est le seuil IMC pour l'obésité et nous savons tous que l'obésité joue un rôle important dans la santé d'une personne.
D'où, on va créer une nouvelle variable bmi30 qui a deux modalités (yes,no).

```
Console  Terminal
~/
> insurance$bmi30 <- ifelse(insurance$bmi>=30,"yes","no")
> describeBy(insurance$charges,insurance$bmi30)

 Descriptive statistics by group
group: no
   vars   n     mean       sd   median trimmed     mad     min      max    range skew kurtosis     se
X1    1 631 10713.67 7843.54 8604.48 9772.74 7024.01 1121.87 38245.59 37123.72 0.97     0.23 312.25
------------------------------------------------------------------------------------------------
group: yes
   vars   n     mean       sd   median  trimmed     mad     min      max    range skew kurtosis    se
X1    1 707 15552.34 14552.32 9964.06 13451.03 7883.43 1131.51 63770.43 62638.92 1.18     0.08 547.3
```

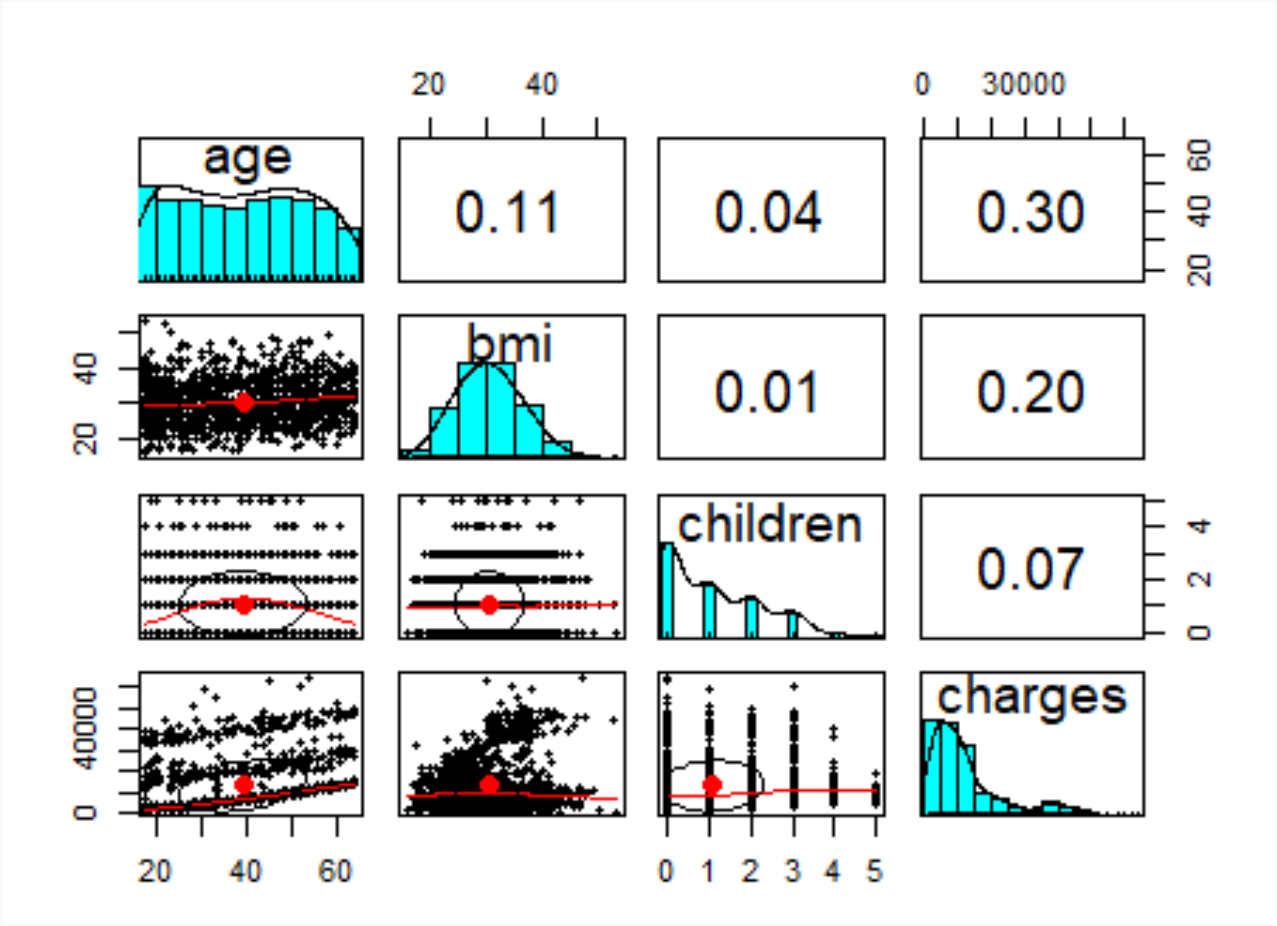
Comme nous pouvons le constater, bien que les dépenses médicales médianes soient identiques aux personnes obèses et non obèses, leurs dépenses moyennes diffèrent de près de 5 000 USD.



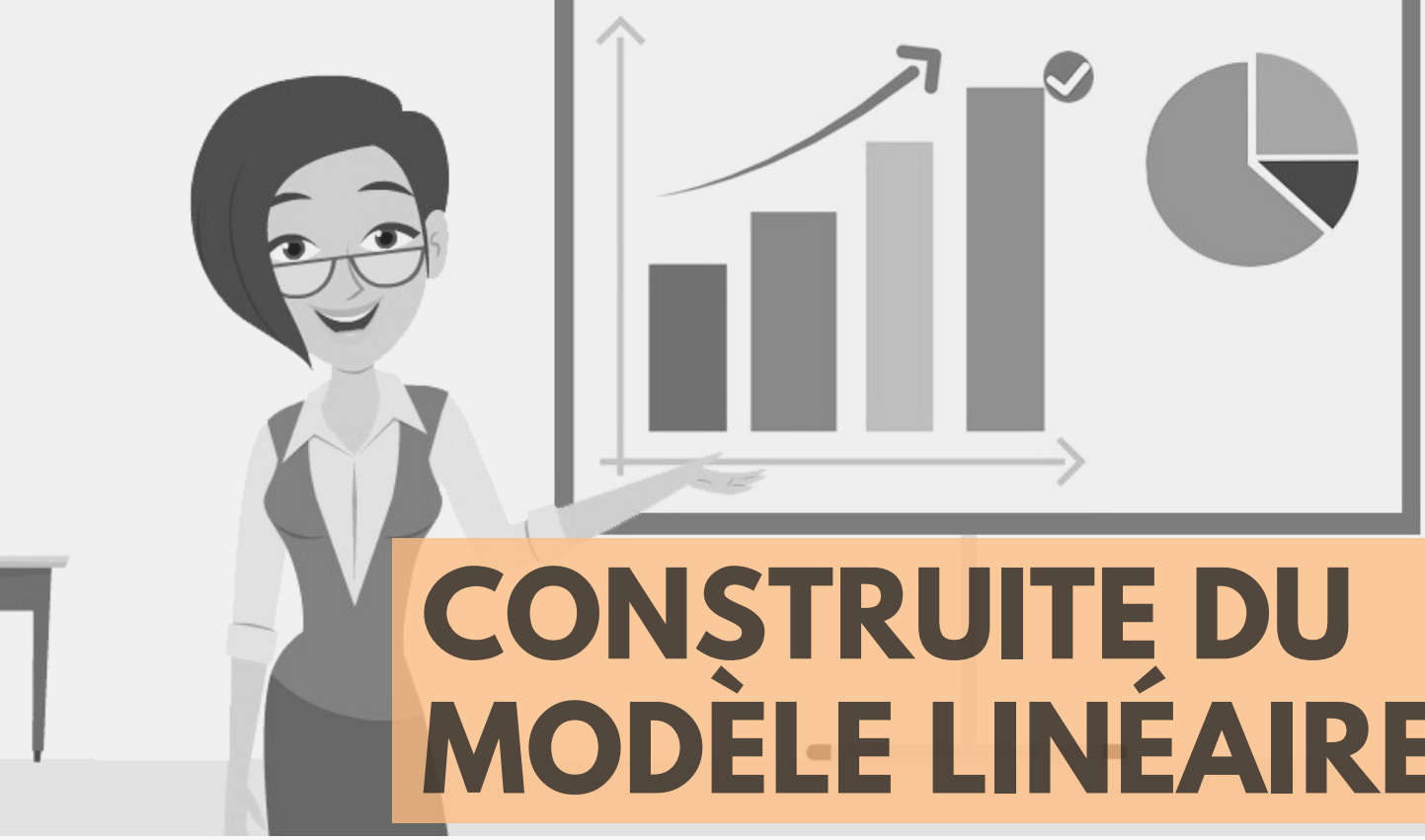
4- ETUDE DE CORRÉLATION

Etudier les corrélation entre les variables est un des étapes principales pour la prédiction, savoir les relations entre les variables est l'un des facteurs qui peut aider à mieux générer un modèle de prédiction plus compréhensif.
Dans R on va visualiser les résultats avec la fonction pairs.panels().

```
Console   Terminal
~/
> pairs.panels(insurance[c("age", "bmi", "children", "charges")])
```



Nous pouvons voir que l'âge a la plus forte corrélation avec les charges parmi nos variables numériques. Une autre observation que nous pouvons faire à partir de ce graphique est qu'aucune de nos valeurs numériques n'est hautement corrélée les unes aux autres, donc la multi-colinéarité ne serait pas un problème. Une autre chose à noter est que la relation entre l'âge et les charges peut ne pas être vraiment linéaire du tout. (nous entrerions dans cela plus tard)



CONSTRUITE DU MODÈLE LINÉAIRE

1- RÉGRESSION LINÉAIRE MULTIPLE

La régression linéaire multiple est une analyse statistique qui décrit les variations d'une variable endogène associée aux variations de plusieurs variables exogènes.

$$Y_i = a_0 + a_1 X_{i1} + a_2 X_{i2} + \ldots + a_p X_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n$$

Le codage factice permet de traiter une caractéristique nominale comme numérique en créant un variable binaire pour chaque catégorie de la fonctionnalité, qui est définie sur 1 si l'observation tombe dans cette catégorie ou 0 sinon. Par exemple, la variable de sexe a deux catégories, hommes et femmes. Ce sera divisé en deux valeurs binaires, qui R noms sexmale et sexfemale. Pour les observations où sexe = masculin, puis sexmale = 1 et sexfemale = 0; si sexe = femme, alors sexmale = 0 et sexfemale = 1. Le même codage s'applique aux variables comportant trois catégories ou plus. La région d'entité à quatre catégories peut être divisée en quatre variables: regionnorthwest, régions sud-est, régions sud-ouest et regionnortheast. Lors de l'ajout d'une variable factice à un modèle de régression, une catégorie est toujours laissé de côté pour servir de catégorie de référence. Les estimations sont ensuite interprétées relatives à la référence. Dans notre modèle, R a tenu automatiquement le sexfemale, smokerno, et les variables régionnortheast, rendant les femmes non-fumeurs dans la région nord-est le groupe de référence.

Institut supérieur des études technologiques de Radès | majdi_hannachi@outlook.fr

2- MODÈLE AVEC TOUS LES VARIABLES

Pour ce modèle, on va tester d'inclure tous les variables. On cherche à exprimer les frais médicaux en fonction des autres variables.
Noter que la regression lineaire dans R est accordée par la fonction '**lm()**' en enregistrant le modéle dans une variable et pour interpréter les resultat du modéle on doit utiliser la fonction '**summary**()'.

Console Terminal

```
> ins_model <- lm(charges ~ age + sex + bmi + children + smoker + region, data = insurance)
> summary(ins_model)

Call:
lm(formula = charges ~ age + sex + bmi + children + smoker +
    region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Sur le premier modèle, nous avons utilisé uniquement les variables d'origine incluses dans l'ensemble de données et obtenu un r-carré décent de 0,7509, ce qui implique que 75,09% de la variation des charges pourrait être expliquée par l'ensemble des variables indépendantes que nous avons incluses. Nous pourrions également observer que toutes les variables indépendantes que nous avons incluses, à l'exception du sexe, constituent un prédicteur statistiquement significatif des frais médicaux (valeur p inférieure à 0,05 <- niveau de signification).

Institut supérieur des études technologiques de Radès | majdi_hannachi@outlook.fr

3- SÉLECTION DES VARIABLES SIGNIFICATIVES SEULEMENT:

Dans ce modèle on va mener une régression multiple sur les variables significatives seulement.

```
> ins_model2<-lm(charges~age + bmi + smoker + children ,insurance)
> summary(ins_model2)

Call:
lm(formula = charges ~ age + bmi + smoker + children, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11897.9  -2920.8   -986.6   1392.2  29509.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
age            257.85      11.90  21.675  < 2e-16 ***
bmi            321.85      27.38  11.756  < 2e-16 ***
smokeryes    23811.40     411.22  57.904  < 2e-16 ***
children       473.50     137.79   3.436 0.000608 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6068 on 1333 degrees of freedom
Multiple R-squared:  0.7497,    Adjusted R-squared:  0.7489
F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

Nous avons maintenant un r-carré de 0,7497, ce qui implique que 74,97% de la variation des charges peut être expliquée par l'âge, l'indice de masse corporelle, le statut de fumeur et le nombre des personnes en charges;
Les résultats du modèle de régression linéaire ont un sens logique; La vieillesse, le tabagisme et l'obésité ont tendance à être liés à des problèmes de santé supplémentaires, tandis que d'autres personnes à la charge des membres de la famille peuvent entraîner une augmentation du nombre de visites chez le médecin et de soins préventifs tels que des vaccinations et des examens physiques annuels.

Institut supérieur des études technologiques de Radès | majdi_hannachi@outlook.fr

4- MODIFICATION DES VARIABLES EXOGÈNES

La première chose que on a fait dans ce bloc est de créer une nouvelle variable age2 qui est fondamentalement un âge carré. Comme on l'a dit plus tôt, la relation entre l'âge et les charges n'est peut-être pas totalement linéaire. L'idée sous-jacente est donc d'inclure la variable age2 pour traiter cette non-linéarité dans notre modèle.

Console Terminal

```
> insurance$age2 <- insurance$age^2
>
> ins_model2 <- lm(charges ~ age + age2 + children + bmi + sex + bmi30*smoker + region, data = insurance)
> summary(ins_model2)

Call:
lm(formula = charges ~ age + age2 + children + bmi + sex + bmi30 *
    smoker + region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-17296.4  -1656.0  -1263.3   -722.1  24160.2

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        134.2509  1362.7511   0.099 0.921539
age                -32.6851    59.8242  -0.546 0.584915
age2                 3.7316     0.7463   5.000 6.50e-07 ***
children           678.5612   105.8831   6.409 2.04e-10 ***
bmi                120.0196    34.2660   3.503 0.000476 ***
sexmale           -496.8245   244.3659  -2.033 0.042240 *
bmi30yes          -1000.1403   422.8402  -2.365 0.018159 *
smokeryes         13404.6866   439.9491  30.469  < 2e-16 ***
regionnorthwest    -279.2038   349.2746  -0.799 0.424212
regionsoutheast    -828.5467   351.6352  -2.356 0.018604 *
regionsouthwest   -1222.6437   350.5285  -3.488 0.000503 ***
bmi30yes:smokeryes 19810.7533   604.6567  32.764  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4445 on 1326 degrees of freedom
Multiple R-squared:  0.8664,    Adjusted R-squared:  0.8653
F-statistic: 781.7 on 11 and 1326 DF,  p-value: < 2.2e-16
```

Comme nous pouvons le constater, l'ajout de ces variables nous a permis d'améliorer considérablement notre modèle. Nous avons maintenant un r-carré de 0,8664, ce qui implique que 86,64% de la variation des charges peut être expliquée par nos variables indépendantes dans le modèle. Le R-carré ajusté du premier modèle est également bien proche que celui du deuxième, ce qui renforce encore notre revendication.

CONCLUSION

Notre modèle le plus ajusté maintenant explique 87% de la variation des coûts de traitement médical. De plus, nos théories sur la forme fonctionnelle du modèle semblent être validées. Le terme d'âge carré est statistiquement significatif, de même que l'indicateur d'obésité, bmi30. L'interaction entre l'obésité et le tabagisme suggère un effet massif; en plus des coûts accrus de plus de 13 404 dollars pour le tabagisme seul, les fumeurs obèses dépensent 19 810 dollars supplémentaires par an. Cela peut suggérer que le tabagisme exacerbe les maladies associées à l'obésité.Les résultats ont également que la variable fumeur est la variable la plus importante pour prédire les frais médicaux.

L'ajout des variables catégoriques des données numériques peut améliorer la performance du modèle: c'est une solution remarquable par l'amélioration de r-carré de 11 %.

Institut supérieur des études technologiques de Radès | majdi_hannachi@outlook.fr

RÉFÉRENCES

- https://edu.kpfu.ru/pluginfile.php/278552/mod_resource/content/1/MachineLearningR__Brett_Lantz.pdf
- https://github.com/stedy/Machine-Learning-with-R-datasets
- https://www.kaggle.com/mirichoi0218/insurance
- http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/
- https://www.rstudio.com/