Initiation à R et Dataviz

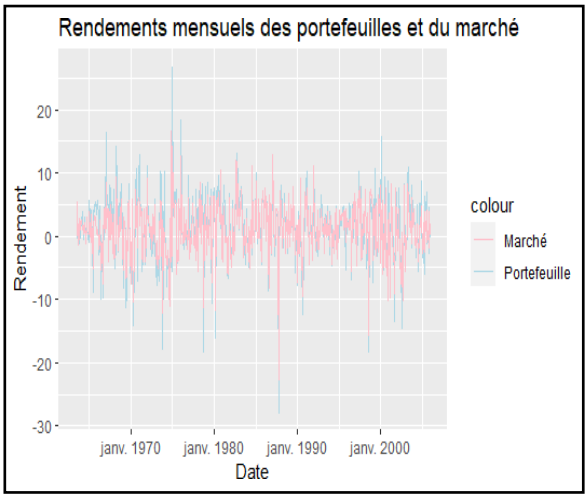
Projet Final

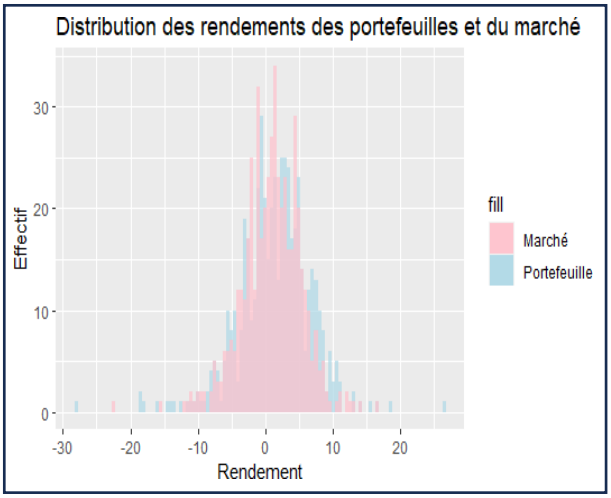
Exercice 1 : Estimation modèle MEDAF sur plusieurs périodes

Période 01 : De juillet 1963 à décembre 2005 :

Statistiques descriptives, graphiques et estimation du modèle linéaire du MEDAF :

```
> summary(donnees_periode_1)
 rend_pf_ME1_BM2    rend_pf_marche           RF         
 Min.   :-27.892   Min.   :-22.6400   Min.   :0.0600  
 1st Qu.: -1.348   1st Qu.: -1.7425   1st Qu.:0.3200  
 Median :  1.571   Median :  1.2150   Median :0.4300  
 Mean   :  1.353   Mean   :  0.9451   Mean   :0.4716  
 3rd Qu.:  4.702   3rd Qu.:  3.9500   3rd Qu.:0.5800  
 Max.   : 26.742   Max.   : 16.6100   Max.   :1.3500  
> ecart_type_ME1_BM2     >  ecart_type_pf_marche
[1] 4.430562              [1] 5.369225
```





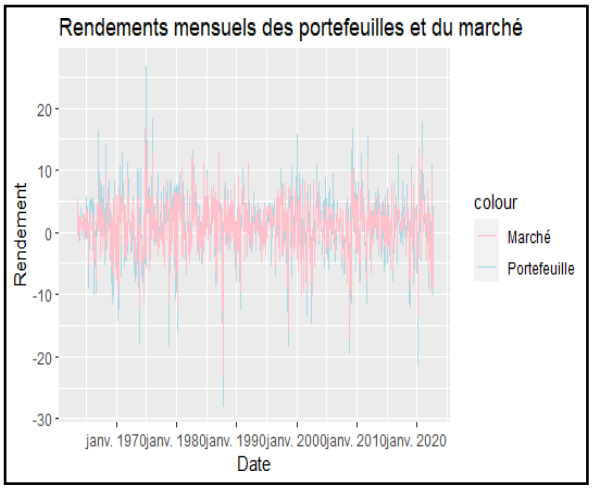

```
> summary(modele_medaf_1)

Call:
lm(formula = rend_pf_ME1_BM2 ~ rend_pf_marche + RF, data = donnees_periode_1)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.1111  -1.5388   0.0219   1.4240  12.2451 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.63240    0.27701   2.283   0.0228 *  
rend_pf_marche  1.04802    0.02693  38.921   <2e-16 ***
RF           -0.57219    0.52552  -1.089   0.2768    
---
```

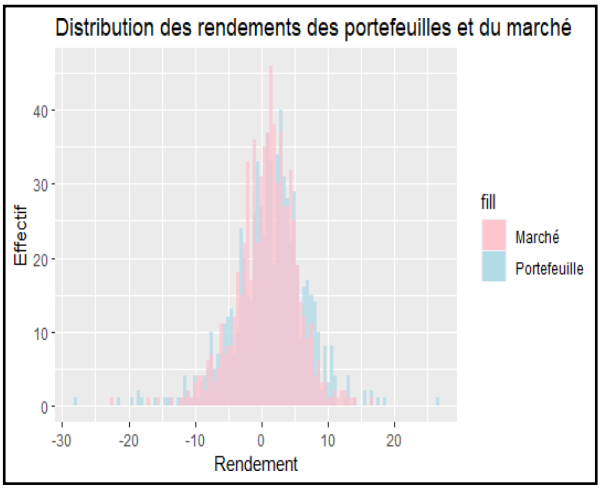
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.69 on 507 degrees of freedom
Multiple R-squared:   0.75,       Adjusted R-squared: 0.7491
F-statistic: 760.7 on 2 and 507 DF,  p-value: < 2.2e-16
```

Période 02 : De juillet 1963 à septembre 2022 :

Statistiques descriptives, graphiques et estimation du modèle linéaire du MEDAF :

```
>summary(donnees_periode_2)
 rend_pf_ME1_BM2    rend_pf_marche           RF
 Min.   :-27.892   Min.   :-22.6400   Min.   :0.0000
 1st Qu.: -1.696   1st Qu.: -1.6950   1st Qu.:0.1400
 Median :  1.485   Median :  1.2600   Median :0.3800
 Mean   :  1.214   Mean   :  0.9057   Mean   :0.3624
 3rd Qu.:  4.532   3rd Qu.:  3.7550   3rd Qu.:0.5100
 Max.   : 26.742   Max.   : 16.6100   Max.   :1.3500
> ecart_type_ME1_BM2      > ecart_type_pf_marche
[1] 4.474923              [1] 5.493661
```



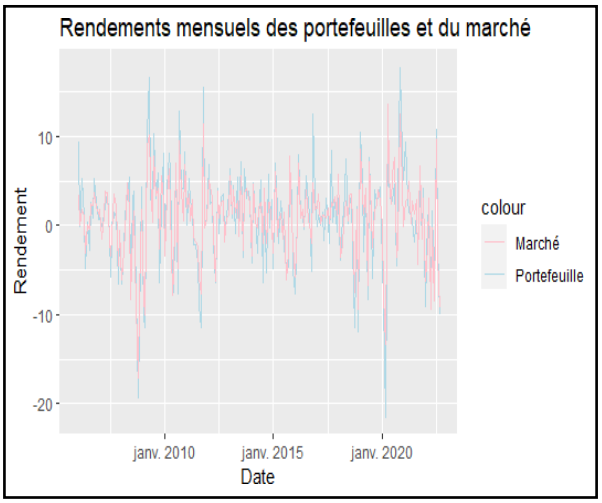


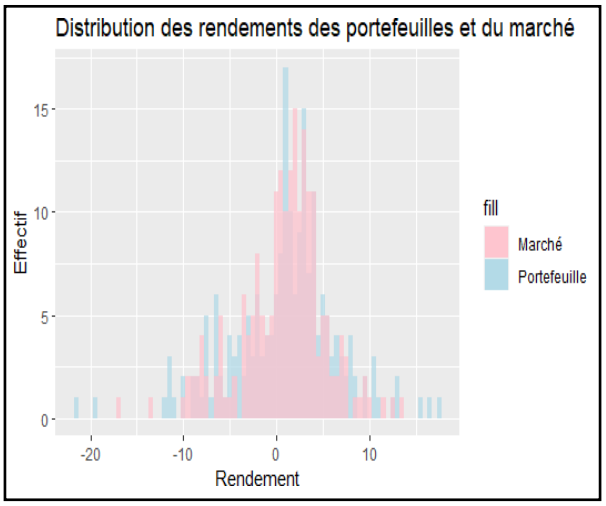

```
> summary(modele_medaf_2)

Call:
lm(formula = rend_pf_ME1_BM2 ~ rend_pf_marche + RF, data = donnees_periode_2)

Residuals:
     Min       1Q   Median       3Q      Max
-11.1811  -1.5718  -0.0706   1.4229  12.2957

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.21711    0.16543   1.312    0.190
rend_pf_marche 1.08232    0.02179  49.678   <2e-16 ***
RF             0.04635    0.36363   0.127    0.899
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.597 on 708 degrees of freedom
Multiple R-squared:  0.7772,      Adjusted R-squared: 0.7765
F-statistic: 1235 on 2 and 708 DF,  p-value: < 2.2e-16
```

Période 03 : De janvier 2006 à septembre 2022 :

Statistique descriptive, graphiques et estimation du modèle linéaire du MEDAF :

```
> summary(donnees_periode_3)
 rend_pf_ME1_BM2    rend_pf_marche           RF          
 Min.   :-21.4683   Min.   :-17.1500   Min.   :0.00000  
 1st Qu.: -2.1295   1st Qu.: -1.5600   1st Qu.:0.00000  
 Median :  1.3336   Median :  1.3400   Median :0.01000  
 Mean   :  0.8617   Mean   :  0.8056   Mean   :0.08527  
 3rd Qu.:  4.0535   3rd Qu.:  3.4000   3rd Qu.:0.14000  
 Max.   : 17.7379   Max.   : 13.6500   Max.   :0.44000  
> ecart_type_ME1_BM2        > ecart_type_pf_marche
[1] 4.595261                [1] 5.79632
```






```
> summary(modele_medaf_3)

Call:
lm(formula = rend_pf_ME1_BM2 ~ rend_pf_marche + RF, data = donnees_periode_3)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.0212 -1.6343 -0.0713  1.2101  6.9084 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -0.03087    0.19830  -0.156    0.876    
rend_pf_marche  1.15798    0.03547  32.648   <2e-16 ***
RF             -0.47280    1.27122  -0.372    0.710    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.298 on 198 degrees of freedom
Multiple R-squared:  0.8444,      Adjusted R-squared: 0.8428 
F-statistic: 537.2 on 2 and 198 DF,  p-value: < 2.2e-16
```

Comparaison entre les trois périodes :

	Rendement de portefeuille ME1 BM2	Rendement du portefeuille de marché
Période 01	Mean : 1,353 Median : 1,571 Max : 26,742 Sd : 4,430562	Mean : 0,9451 Median : 1,215 Max : 16,61 Sd : 5,369225
Période 02	Mean : 1,214 Median : 1,485 Max : 26,742 Sd : 4,474923	Mean : 0,9057 Median : 1,26 Max : 16,61 Sd : 5,493661
Période 03	Mean : 0,8617 Median : 1,3336 Max : 17,7379 Sd : 4,595261	Mean : 0,8056 Median : 1,34 Max : 13 ,65 Sd : 5,79632

On remarque que le rendement moyen du portefeuille ME1 BM2 est relativement supérieur sur la période de juillet 1963 à décembre 2005 par rapport aux autres périodes. De plus, le rendement moyen du portefeuille ME1 BM2 est relativement supérieur sur la période de juillet 1963 à septembre 2022 par rapport à la période qui s'étend de janvier 2006 à septembre 2022 => On peut conclure que le rendement moyen du portefeuille ME1 BM2 a significativement baissé à partir de janvier 2006.

On remarque aussi que l'écart-type sur la période de janvier 2006 à septembre 2022 est légèrement plus élevé que celui de la période de juillet 1963 à décembre 2005, qui peut indiquer une certaine augmentation de volatilité des investissements qui peut être vu comme un risque pour un certain type d'investisseur. Concernant le rendement du portefeuille de marché, on peut voir qu'il suit presque la même tendance que le rendement du portefeuille ME1 BM2 (chose qu'on peut remarquer sur les graphiques ci-dessus), avec une baisse significative du rendement moyen et une légère augmentation de volatilité à partir de janvier 2006.

	Estimation du modèle du MEDAF
Période 01	(Intercept) 0.63240 rend_pf_marche 1.04802 RF -0.57219 Multiple R-squared: 0.75 F-statistic: 760.7 p-value: < 2.2e-16
Période 02	(Intercept) 0.21711 rend_pf_marche 1.08232 RF 0.04635 Multiple R-squared: 0.7772

	F-statistic: 1235 p-value: < 2.2e-16
Période 03	(Intercept) -0.03087 rend_pf_marche 1.15798 RF -0.47280 Multiple R-squared: 0.8444 F-statistic: 537.2 p-value: < 2.2e-16

Sur les trois périodes, on remarque que le coefficient de détermination est suffisamment élevé, qui veut dire que nos variables explicatives expliquent une partie importante de notre modèle.

On remarque aussi que l'estimateur du rendement du portefeuille de marché est positif, c'est-à-dire qu'il y a une relation positive entre le rendement du portefeuille du marché et le rendement de portefeuille ME1 BM2 et aussi l'estimateur est statistiquement significatif qui veut dire qu'il y a une forte corrélation entre eux.

On peut voir qu'il y a une relation négative entre le rendement sans risque et le rendement de portefeuille ME1 BM2.

La p-value du modèle est inférieur à 5% pour les trois périodes => c'est-à-dire que l'estimation du modèle du MEDAF est statistiquement significatif.

En conclusion, on peut dire que les résultats de notre estimation viennent confirmer notre premier raisonnement qui est que le rendement du portefeuille ME1 BM2 et le rendement du portefeuille du marché suivent la même tendance et presque la même distribution et la preuve est qu'il y a une relation positive entre les deux (c-à-d que les rendements augmentent et baissent ensemble, chose qu'on peut remarquer sur les graphiques) et aussi la forte corrélation qui existe entre les deux rendements.

Exercice 2 : Estimation du modèle Fama-French à 3 facteurs

Voir script R.

Exercice 3 : Tabagisme et âge

Question 1 :

Nous avons à notre disposition un ensemble de données médicales sur les maladies pulmonaires et le tabac. Le fichier comprend 7 variables qui nous donnent des informations sur les individus étudiés (chacun a un numéro ID affecté). Les informations données sont l'âge, le genre, la situation conjugale, le niveau de consommation de tabac, l'exposition au tabagisme passif et enfin la présence de problème pulmonaires. La variable ID n'est pas pertinente à étudier car elle donne simplement un numéro à chaque individu.

Parmi ces variables 4 sont qualitatives :

_Le sexe qui renvoi « homme » ou « femme »

_La situation conjugale de l'individu qui peut être « marie », « en couple », « célibataire » ou « veuf »

_Le tabagisme passif qui nous dit si l'individu y est exposé ou non avec « TRUE » ou « FALSE »

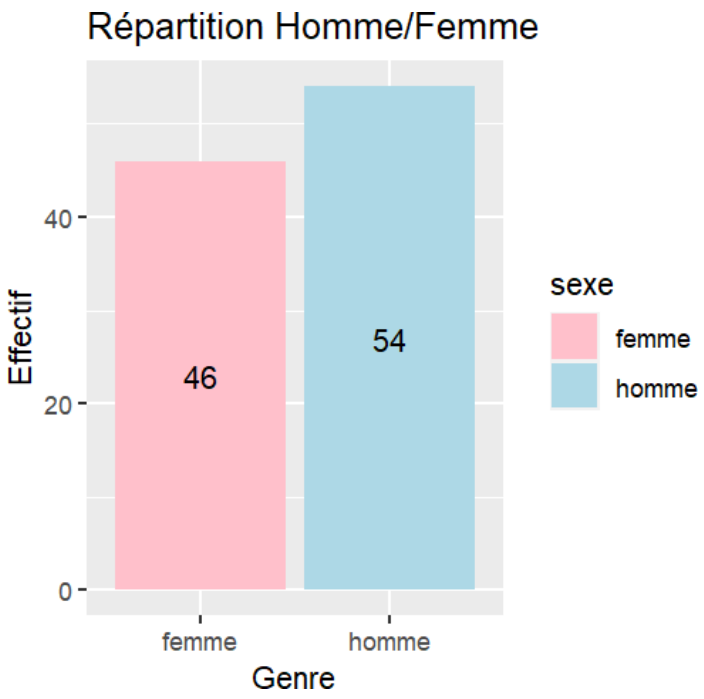
_La présence éventuelle de problème pulmonaires avec « TRUE » ou « FALSE »

Par conséquent 2 variables quantitatives :

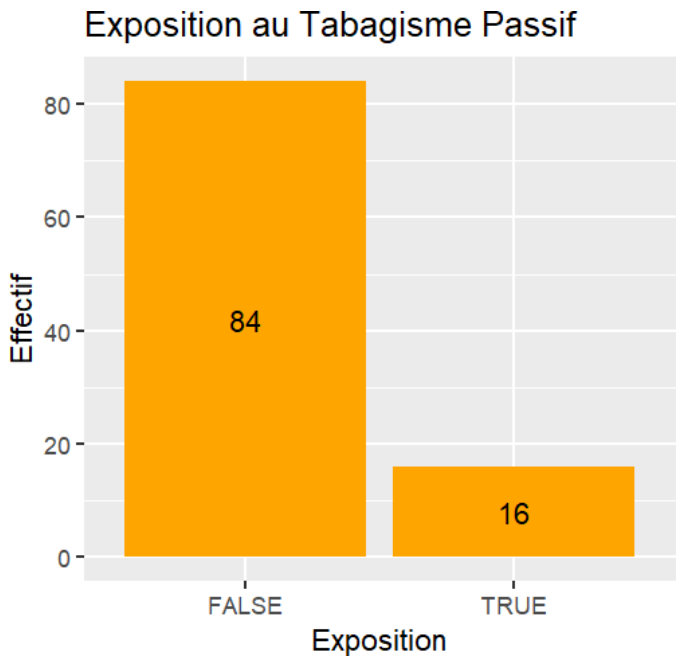
_L'âge de l'individu qui renvoi un endroit un nombre entier

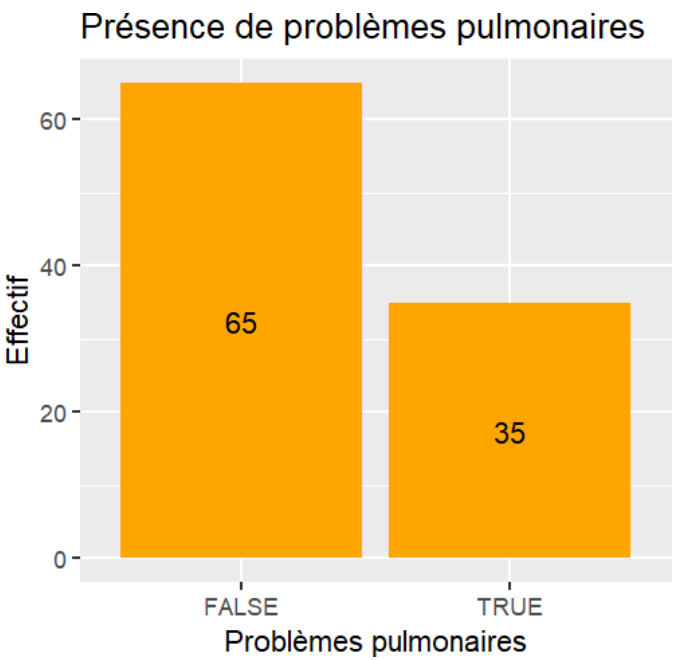
_Le niveau de consommation de tabac de chaque individu traduit par un nombre entier allant de 0 à 14

Nous pouvons désormais effectuer une analyse statistique des données :

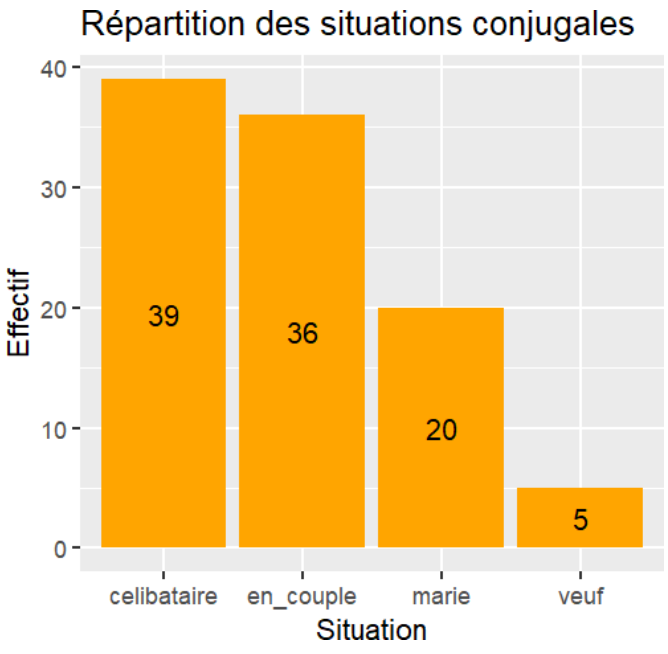


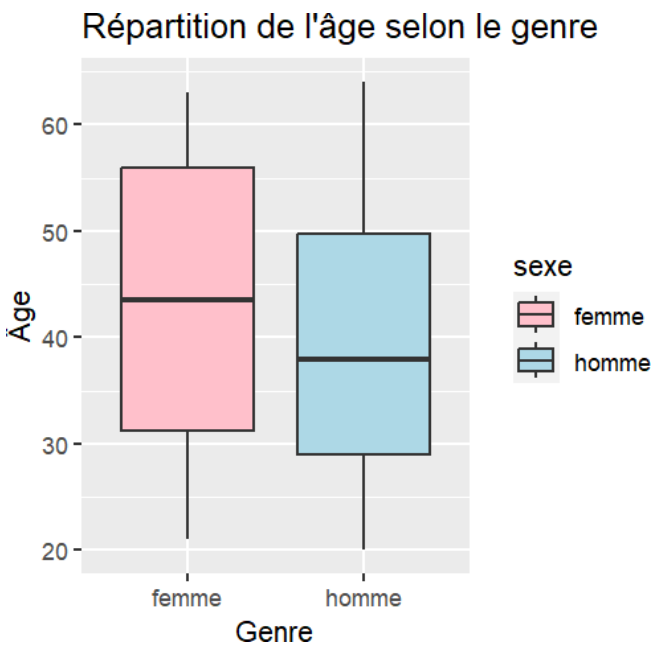
On remarque qu'il y a 100 individus dans notre étude, ainsi chaque valeur peut être interprétée comme un pourcentage. Il y a 46 femmes et 54 hommes, la parité est relativement respectée.





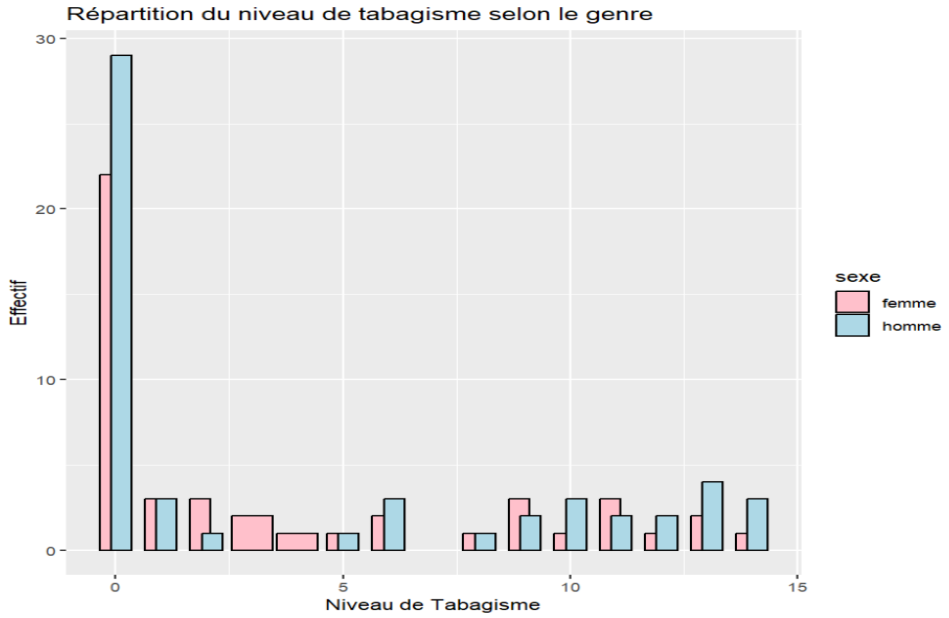
Ci-dessus, deux graphiques nous donnant des informations sur l'exposition au tabagisme passif et sur la présence de problèmes pulmonaires chez les individus de l'étude. On remarque que 84 personnes ne sont pas exposées au tabagisme passif et par conséquent 16 y sont exposés. Ensuite, 65 personnes n'ont pas de problèmes pulmonaires et 35 en ont.





Ci-dessus à gauche, une représentation de la répartition des différentes situations conjugales des individus, 39 sont célibataires, 36 sont en couple, 20 mariés et enfin 5 sont veufs.

Enfin, en haut à droite un graphique « en boîtes à moustache » nous donnant des informations sur l'âge des individus selon leur genre. On remarque que l'étendu est plus élevée chez les hommes mais la médiane et de manière générale les femmes sont plus âgées que les hommes dans notre étude.



Ici, un histogramme en barre représentant la répartition du niveau de tabagisme de 0 à 14 par genre. On remarque que les hommes fument moins que les femmes dans notre étude. Ensuite, les effectifs sont relativement similaires selon le genre.

```
      age            sexe          situation      tabac       tabagisme_passif
 Min.   :20.00   femme:46   celibataire:39   Min.   : 0.0   FALSE:84
 1st Qu.:29.75   homme:54   en_couple  :36   1st Qu.: 0.0   TRUE :16
 Median :41.00              marie      :20   Median : 0.0
 Mean   :41.38              veuf       : 5   Mean   : 3.9
 3rd Qu.:52.25                               3rd Qu.: 9.0
 Max.   :64.00                               Max.   :14.0
 probleme_pulmonaire
 FALSE:65
 TRUE :35
```

Ci-dessus un tableau récapitulatif de toutes les statistiques descriptives de notre étude. On remarque que l'âge est compris entre 20 et 64ans avec une moyenne de 41ans. De plus, la moyenne du niveau de tabagisme est de 4 (échelle de 0 à 14).

Question 2 :

Nous estimerons le lien entre la variable de la consommation de tabac et la variable de l'âge avec d'autres variables explicatives. Compte tenu de l'inclusion de certaines caractéristiques qualitatives telles que le sexe de la personne, les complications pulmonaires et la situation matrimoniale, nous utiliserons des variables dichotomiques (notamment pour le sexe et les problèmes pulmonaires) et polytomiques (notamment pour la situation matrimonial).

L'inclusion de ces variables explicative n'est pas sans conséquences. Nous donnerons les hypothèses principales qui justifie la pertinence d'inclure chacune de ces caractéristiques qualitatives et nous justifierons par les estimations pourquoi ce choix :
Il est important de noter que ces caractéristiques qualitatives sont de simples hypothèses est que leurs non significativité dans le modèle (hormis le lien positif entre problèmes pulmonaire et consommation de tabac qui a était prouvé scientifiquement) n'est pas étonnant.

Problèmes pulmonaires : Nous savons que plus la consommation de tabac augmente plus les complications survienne (apparition de maladie). On suggère donc qu'il y a potentiellement un lien positif entre les complications et la consommation de tabac.

Situation matrimoniale : Une personne seule sera plus encline à fumer qu'une personne en couple.

Inclusion du tabagisme : On peut se dire qu'une exposition au tabagisme peut influer sur notre choix de consommer du tabac ou non. Si on est exposé, alors on peut être moins enclin à consommer peut-être à cause des problèmes de santé que cela peut nous procurer. Nous savons qu'un fumeur passif aura plus de chance d'avoir des problèmes de santé qu'un fumeur.

Etant donné notre modèle linéaire multiple, nous verrons la multi colinéarité entre les variables explicatives. Cela est un bon indicateur de la précision de nos estimateurs de notre modèle (l'hypothèse d'indépendance des variables explicatives des hypothèses de Gauss-Markov (hypothèse de non colinéarité ou multi colinéarité) doit être respecté pour avoir des estimateurs sans biais). Pour rappel, une valeur GVIF proche de 1 signifie qu'il n'y a aucune multi colinéarité. Ici, les valeurs sont proches de 1 ce qui est parfaitement acceptable est nous pouvons donc estimer notre modèle avec ces variables explicatives

	GVIF	Df	GVIF^(1/(2*Df))
data_tabac$age	1.122087	1	1.059286
data_tabac$probleme_pulmonaire	1.276438	1	1.129795
data_tabac$tabagisme_passif	1.218923	1	1.104048
data_tabac$sexe	1.138520	1	1.067014
data_tabac$situation	1.252831	3	1.038282

```
Call:
lm(formula = data_tabac$tabac ~ data_tabac$age + data_tabac$probleme_pulmonaire +
    data_tabac$tabagisme_passif + data_tabac$sexe + data_tabac$situation)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8352 -1.2654 -0.1267  1.4567  6.0999

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.63470    1.18128   3.923 0.000168 ***
data_tabac$age                  -0.08031    0.02481  -3.237 0.001682 **
data_tabac$probleme_pulmonaireTRUE  8.47318    0.72523  11.683  < 2e-16 ***
data_tabac$tabagisme_passifTRUE    -0.63450    0.92206  -0.688 0.493101
data_tabac$sexehomme                0.01987    0.65549   0.030 0.975886
data_tabac$situationen_couple      -0.40567    0.73334  -0.553 0.581479
data_tabac$situationmarie          -0.96444    0.87370  -1.104 0.272532
data_tabac$situationveuf            1.04993    1.50096   0.700 0.486003
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.062 on 92 degrees of freedom
Multiple R-squared:  0.6595,    Adjusted R-squared:  0.6336
F-statistic: 25.46 on 7 and 92 DF,  p-value: < 2.2e-16
```

On estime qu'une augmentation d'une année supplémentaire de l'âge fait baisser la consommation de tabac de -0,08311 unité, toute chose égale par ailleurs. De plus cela est significatif car la p-value est inférieure à 0,05. Il y a donc bien un lien entre l'âge et la consommation de tabac. L'explication la plus probable est l'augmentation des problèmes de santé (pulmonaire) au fur et à mesure de l'âge. Ce qui peut expliquer une certaine multi colinéarité dans notre modèle. Mais l'exclusion de ces variables n'est pas possible car elles ont un lien significatif avec le niveau de consommation de tabac.

Pour la situation matrimoniale, nous voyons qu'elle respecte bien notre hypothèse de départ qui nous dit qu'une personne isolée aura plus tendance a consommé plus de tabac. Nous voyons un lien positif entre la solitude (personne veuf) et la consommation de tabac. De plus, il y a un lien négatif entre la consommation de tabac et les personnes en couple ou marié. Le lien entre le sexe de la personne et la consommation de tabac n'est pas significatif. Enfin, on remarque qu'il y a bien un lien négatif entre le tabagisme passif et la consommation de tabac.

Enfin nous voyons une relation positive entre la consommation de tabac et les problèmes pulmonaire. Le coefficient de détermination est de 0,65 ce qui signifie que le modèle explique à 65,95% la variance de la consommation de tabac. Le modèle linéaire multiple devrait s'écrire comme suit car ce sont les 3 seules variables qui expliquent directement le niveau de consommation de tabac.

```
Call:
lm(formula = data_tabac$tabac ~ data_tabac$age + data_tabac$probleme_pulmonaire +
    data_tabac$tabagisme_passif)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5423 -1.3705 -0.1218  1.3787  6.2175

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.35141    1.00995   4.309 3.97e-05 ***
data_tabac$age                  -0.07844    0.02398  -3.271  0.00149 **
data_tabac$probleme_pulmonaireTRUE  8.40713    0.69169  12.155  < 2e-16 ***
data_tabac$tabagisme_passifTRUE    -0.92440    0.88376  -1.046  0.29820
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.033 on 96 degrees of freedom
Multiple R-squared:  0.6513,    Adjusted R-squared:  0.6405
F-statistic: 59.78 on 3 and 96 DF,  p-value: < 2.2e-16
```

```
> vif(mod2)
              data_tabac$age data_tabac$probleme_pulmonaire    data_tabac$tabagisme_passif
                    1.068158                       1.183167                       1.141077
```

On remarque que cela a aussi baissé la multi colinéarité du modèle sans pour autant l'effacer Baisser cette colinéarité est très important car cela permet d'améliorer la précision de nos estimateurs (permet un estimateur sans-biais de par les hypothèses de gauss Markov). De plus, le coefficient de détermination est de 65,13% ce qui est moins élevé que l'autre modèle. Ce modèle n'a pas de capacité prédictive plus basse que l'autre modèle pour autant. Nous savons que le multiple R2 multiple est biaisé lorsqu'il y a un grand nombre de variables explicatives car il est toujours plus élevé lorsque on augmente le nombre de variables.

Pour comparer les deux modèles, nous devons plutôt étudier le R2 ajusté qui diminue lorsqu'une variable explicative ne contribue pas suffisamment à l'explication de la variable expliqué. Donc un R2 ajusté plus élevé signifie un modèle plus prédictif. Ici notre modèle à un R2 ajusté plus élevé (0,6405), il est donc préférable.