Thème : Modèles de régression linéaires TP 2

Prenez soin de bien comprendre les fonctions et commandes de **R** utilisées dans ce travail pratique et ne vous contentez pas d'effectuer une simple copie dans votre session de **R** des commandes se trouvant dans l'énoncé.

Ce travail pratique peut être réalisé par groupes de deux étudiant·e·s. Dans vos recherches bibliographiques, indiquez clairement et précisément vos sources et ne vous limitez pas à Wikipédia, l'encyclopédie libre.

Par rapport au rendu du travail pratique précédent, étoffez celui de ce travail pratique par de nouvelles possibilités que vous offre **rmarkdown**.

Exercice 1

Les données contenues dans le tableau ci-dessous indiquent la taille moyenne en centimètres (cm) d'un groupe d'enfants de Kalama, village en Egypte, mesurée chaque mois entre le 18-ième et le 29-ième mois.

âge (mois)	18	19	20	21	22	23	24	25	26	27	28	29
taille (cm)	76.1	77	78.1	78.2	78.8	79.7	79.9	81.1	81.2	81.8	82.8	83.5

Les données se trouvent dans le fichier `kalama.txt`.

Pour décrire la relation existant entre la taille moyenne des enfants et l'âge, on se propose d'utiliser un modèle de régression linéaire simple

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i, \quad i = 1, \ldots, 12,$$

où les x_i représentent les âges (**age**) et les y_i les tailles moyennes (**taille**) correspondantes. Les ε_i sont les réalisations de variables aléatoires, appelées erreurs, supposées indépendantes et issues de distributions normales d'espérance 0 et de variance σ^2 constante.

a) Calculer le coefficient de corrélation entre la taille et l'age.

b) Tracer le nuage de points **taille** versus **age**.

c) Estimer les coefficients β_0 et β_1 par la méthode des moindres carrés. Enregistrer le résultat de l'ajustement du modèle dans l'objet **kalama.lm** de **R**.

d) Ajuster sur le graphique la droite des moindres carrés en utilisant la fonction **abline()**.

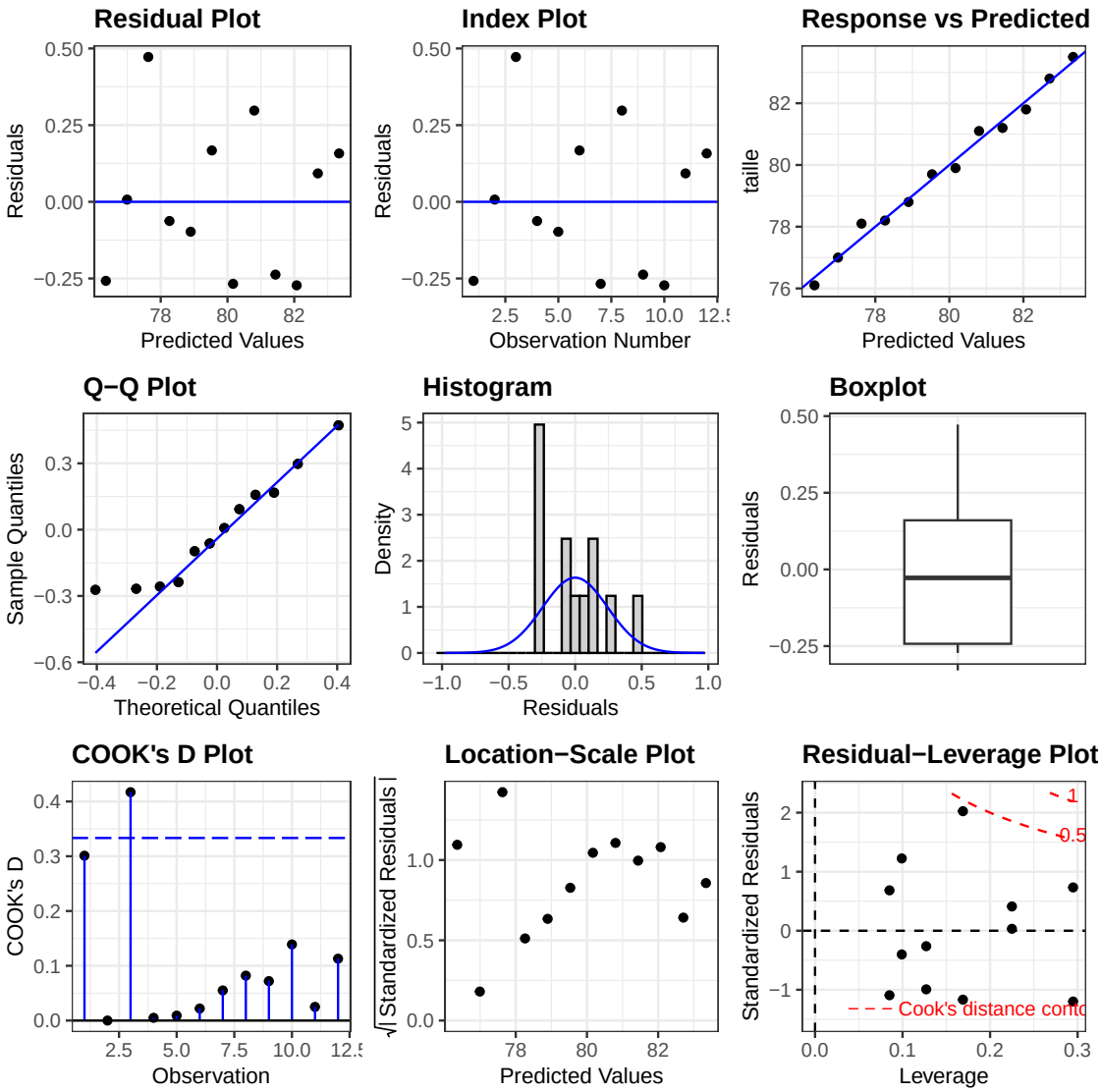
e) Estimer la variance des erreurs σ^2.

f) Déterminer la valeur du coefficient de détermination R^2 de la régression.

g) Effectuer un diagnostic du modèle ajusté à l'aide des graphiques appropriés.

Le diagnostic peut également être réalisé à l'aide de la librairie **ggResidpanel** en utilisant les commandes

```
library(ggResidpanel)
resid_panel(kalama.lm, plots="all")
```



et

```
resid_interact(kalama.lm)
```

h) En combinant les résultats obtenus en d), f) et g), qualifier l'ajustement du modèle aux données observées.

i) Tester à un niveau de signification de 5 % si la pente de la droite de régression est significativement différente de zéro.

Exercice 2

Le Produit National Brut (PNB) par habitant (x) et la mortalité infantile (y) ont été relevés il y a quelques années dans 14 pays européens et figurent dans le tableau ci-dessous. On se demande si une relation linéaire existe entre les deux variables.

Pays	PNB	Mortalité
Allemagne RFA	190	24
Autriche	128	28
Belgique	180	24
Danemark	212	19
Espagne	56	37
France	192	22
Grèce	68	34
Irlande	98	25
Italie	110	36
Luxembourg	197	24
Pays-Bas	155	14
Portugal	40	65
Royaume–Uni	181	20
Suisse	233	18

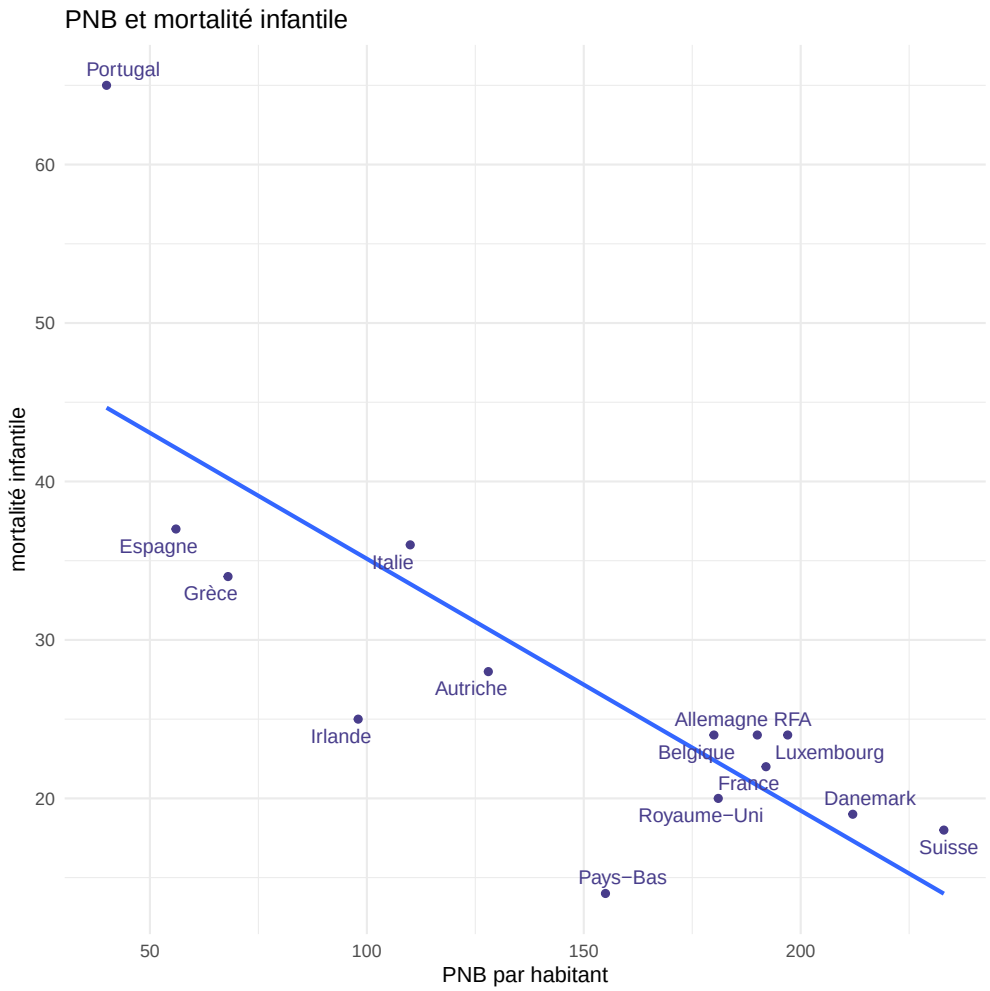
a) Introduire les données observées dans votre session actuelle de **R** en complétant les commandes ci-dessous.

```
mortalite.pnb <- data.frame(pays=c("Allemagne RFA","Autriche","Belgique","Danemark","Espagne",
                                   "France","Grèce","Irlande","Italie","Luxembourg","Pays-Bas",
                                   "Portugal","Royaume-Uni","Suisse"),
                  PNB=c(...),
                  mortalite=c(...))

print(mortalite.pnb)
```

```
##             pays PNB mortalite
## 1  Allemagne RFA 190        24
## 2       Autriche 128        28
## 3       Belgique 180        24
## 4       Danemark 212        19
## 5        Espagne  56        37
## 6         France 192        22
## 7          Grèce  68        34
## 8        Irlande  98        25
## 9         Italie 110        36
## 10    Luxembourg 197        24
## 11      Pays-Bas 155        14
## 12      Portugal  40        65
## 13   Royaume-Uni 181        20
## 14        Suisse 233        18
```

b) Déterminer le résumé des variables en utilisant la fonction **summary()**.

c) Reproduire le graphique de nuage de points de la page suivante représentant la mortalité infantile (y) en fonction du produit national brut (x). Pour y parvenir, il convient d'utiliser les librairies **tidyverse** et **ggrepel**.



Le graphique met-il en évidence des pays qui ne suivent pas la tendance générale dessinée par les points ? Dans l'affirmative, de quel(s) pays s'agit-il ? Dans la négative, expliquer pourquoi aucun pays atypique ne se dégage du graphique.

d) Déterminer l'équation de la droite de régression linéaire ajustée sur le nuage de points à l'aide de la méthode des moindres carrés.

e) Donner le coefficient de corrélation linéaire r.

f) En déduire le coefficient de détermination R^2.

g) Effectuer une vérification des hypothèses inhérentes au modèle à l'aide des graphiques appropriés.

h) Donner à l'aide du modèle ajusté une prédiction de la mortalité infantile pour un PNB de 100. Pour y parvenir, il convient d'utiliser la fonction **predict()**.

```
xnew <- matrix(c(100), nrow=1)
colnames(xnew) <- c("PNB")
xnew <- as.data.frame(xnew)
predict(mortalite.pnb.lm, xnew, interval="pred")
```

mortalite.pnb.lm contient le résultat de l'ajustement du modèle linéaire.

i) Un modèle de régression linéaire a été ajusté en considérant la mortalité comme variable de réponse et le logarithme naturel du PNB comme variable explicative. Le coefficient de détermination associé à l'ajustement de ce modèle vaut 0.74.

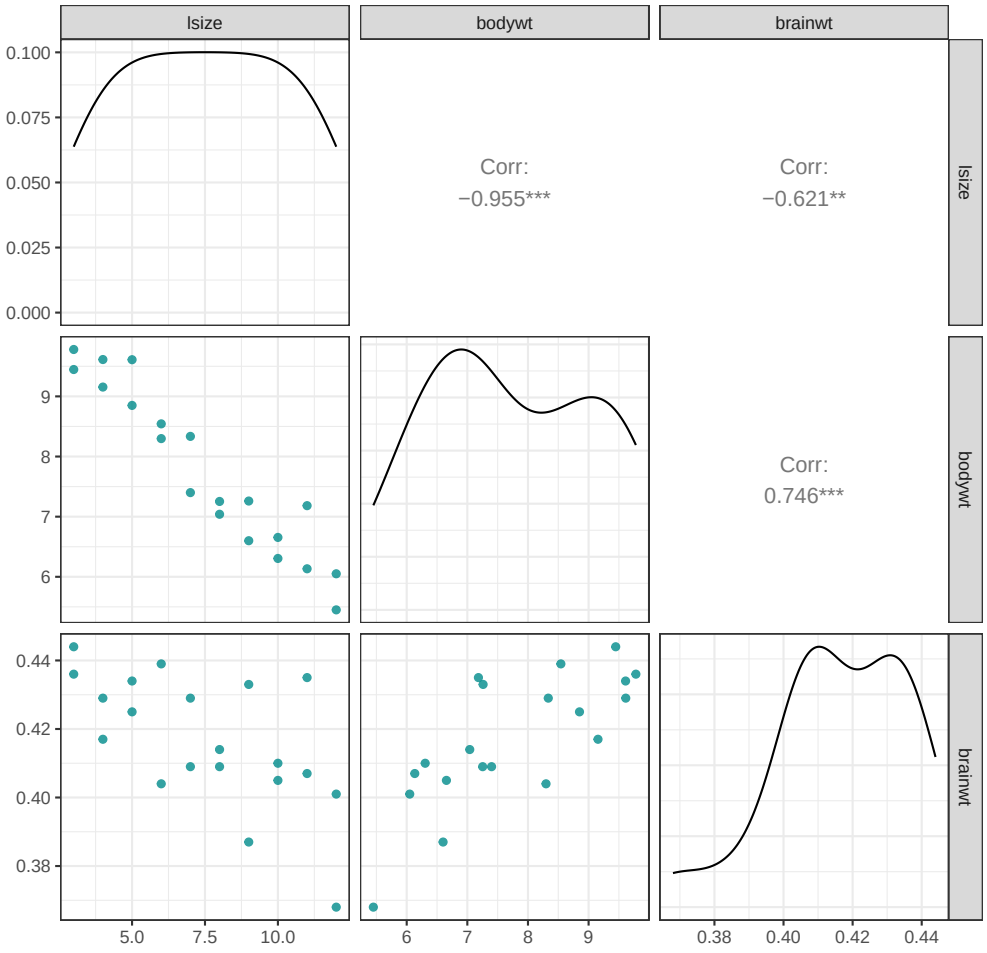
Pensez-vous que ce nouveau modèle de régression est de meilleure qualité que celui utilisé en d) ? Justifiez clairement votre réponse.

Exercice 3

La taille de la portée (**lsize**), le poids du corps (**bodywt**) et le poids du cerveau (**brainwt**) ont été relevés dans un échantillon formé de 20 souris. On se propose d'étudier la relation qui peut exister entre le poids du cerveau (variable de réponse) et la taille de la portée et le poids du corps (variables explicatives).

Les données se trouvent dans le fichier `litters.rds` qu'il faut lire dans votre session actuelle de **R** à l'aide de la fonction **readRDS()**.

a) Tracer le graphique des corrélations et des nuages de points se trouvant ci-dessous à l'aide de la fonction **ggpairs()** de la librairie **GGally**.



b) Existe-t-il une relation entre les deux variables explicatives **lsize** et **bodywt** ?

c) Qu'en est-il de la relation entre la variable de réponse **brainwt** et les deux variables explicatives à considérer l'une après l'autre ?

d) Répondre aux mêmes questions en utilisant la librairie **rgl** de **R** et en complétant le code ci-dessous.

```
library(rgl)
plotids <- with(litters, plot3d(..., ..., ..., type="s", col="blue"))
rglwidget(elementId = "plot3drgl")
```

e) Sera-t-il commode de distinguer clairement les effects des deux variables explicatives sur la variable de réponse ?

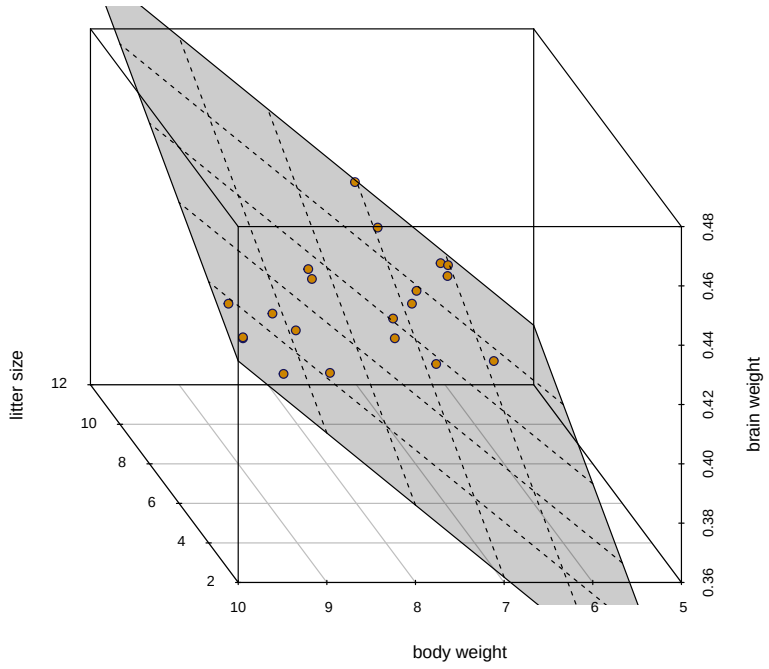
f) Déterminer l'équation du modèle de régression linéaire multiple obtenue par la méthode des moindres carrés.

g) Dresser la table qui résume l'ajustement du modèle à l'aide de la fonction **summary()**. Tester séparément la significativité de chaque variable explicative par rapport au modèle complet.

Quelles variables explicatives sont significatives à un niveau de signification de 5 % ? Que peut-on conclure en considérant également les graphiques des corrélations et des nuages de points tracés ci-dessus ?

h) Tracer le graphique du nuage de points ci-dessous en complétant les commandes

```
library(scatterplot3d)
s3d <- scatterplot3d(litters$..., litters$..., litters$..., main = "",
                     color="midnightblue", xlab="litter size", ylab="body weight",
                     zlab="brain weight", angle = -60, pch=21, bg="orange")
s3d$plane3d(litters.fit, draw_polygon=TRUE, lty.box="solid")
```



litters.fit contient le résultat de l'ajustement du modèle linéaire.

i) Déterminer le coefficient de détermination R^2 et le coefficient de détermination ajusté R^2_{adj} associés au modèle. Que peut-on conclure ?

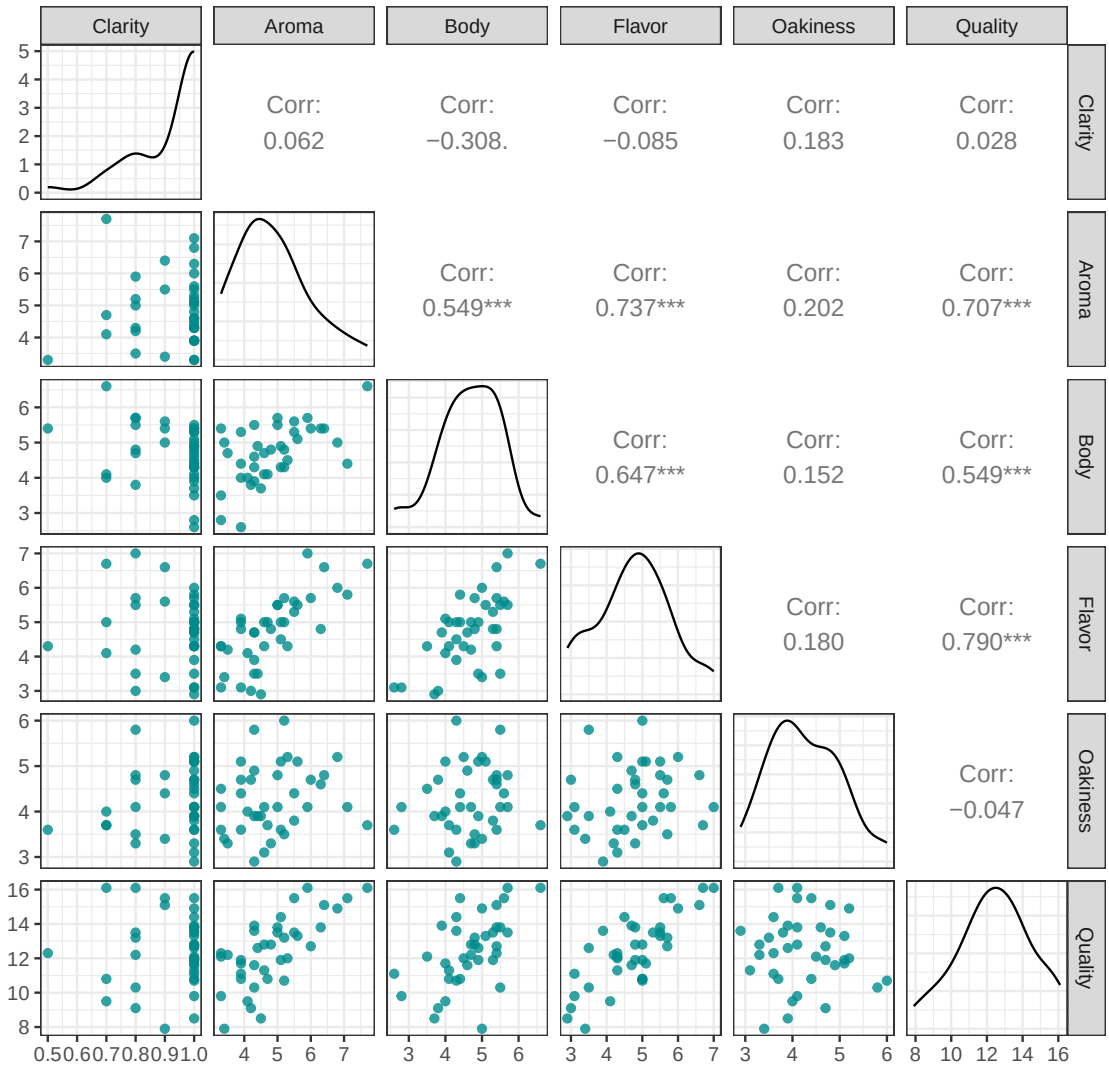
j) Effectuer une vérification des hypothèses inhérentes au modèle à l'aide des graphiques appropriés. À votre avis, que peut-on faire pour améliorer la qualité du modèle ajusté ?

Exercice 4

Dans une dégustation sensorielle de vin, on s'est intéressé à la qualité de 38 pinots noirs. Cinq variables explicatives ont été considérées, la transparence (**Clarity**), l'arôme (**Aroma**), le corps (**Body**), la saveur (**Flavor**) et le goût boisé (**Oakiness**) ; la variable de réponse y étant la qualité du vin dégusté. Les données proviennent de l'article écrit par Kwan, Kowalski et Skogenboe dans la revue *Journal of Agricultural and Food Chemistry, Vol. 27, 1979*. Elles se trouvent dans le fichier `Wine.csv`.

Un modèle de régression linéaire standard a été ajusté aux observations relevées.

a) Tracer le graphique des corrélations et des nuages de points se trouvant ci-dessous à l'aide de la fonction **ggpairs()** de la librairie **GGally**.



Quelles sont les variables explicatives qui semblent au mieux expliquer la variable de réponse ?

b) Déterminer l'équation du modèle de régression linéaire multiple obtenue par la méthode des moindres carrés.

c) Dresser la table qui résume l'ajustement du modèle. Tester séparément la significativité de chaque variable explicative par rapport au modèle complet.

Quelles variables explicatives sont significatives à un niveau de signfication de 5 % ? Que peut-on conclure en considérant également les graphiques des corrélations et des nuages de points tracés en a) ?

d) Déterminer le coefficient de détermination R^2 et le coefficient de détermination ajusté R^2_{adj} associés au modèle.

e) Le coefficient de détermination R^2 augmente de manière monotone avec l'introduction de nouvelles variables même si celles-ci sont peu corrélées avec la variable de réponse. Ainsi, il ne permet pas de sélectionner de manière efficace les variables explicatives. Pour pallier cet inconvénient, différentes méthodes existent pour sélectionner les covariables qui expliquent le mieux la variable de réponse. Parmi elles figurent, outre le coefficient de détermination ajusté R^2_{adj}, le critère d'Akaike (*AIC*), le critère d'Information Bayésien (*BIC*) ainsi que le critère C_p de Mallows. Il faut noter que ces critères ne sont pas des critères d'ajustement des modèles. La qualité de l'ajustement est un autre domaine. En fait, ces critères permettent de comparer des modèles et sont fréquemment utilisés pour sélectionner le "meilleur" modèle.

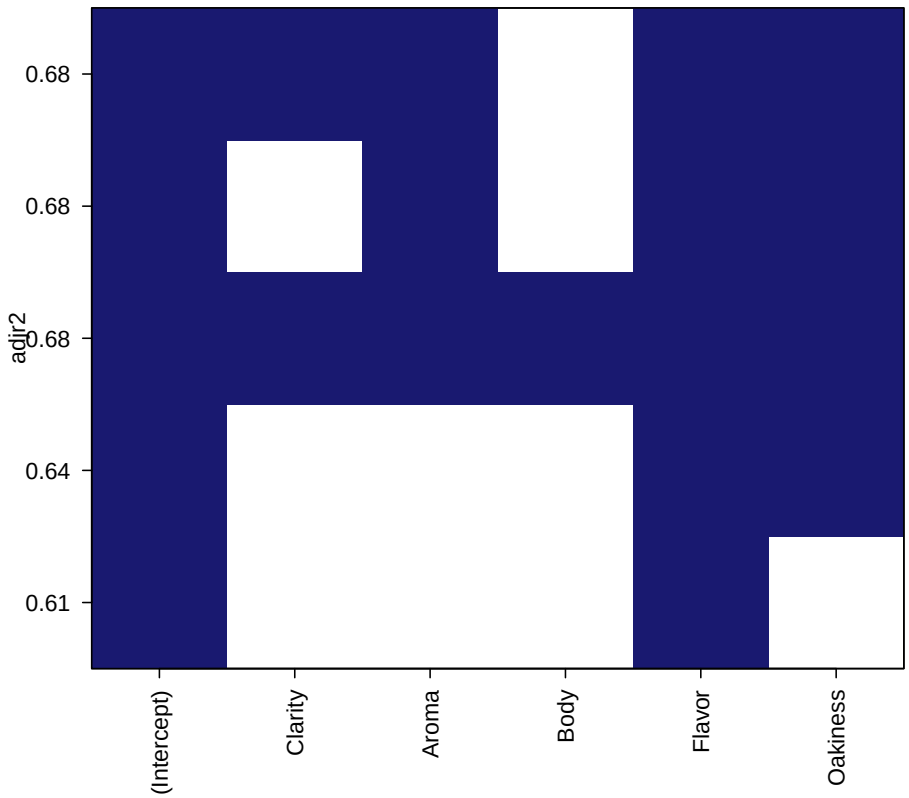
Faire une recherche bibliographique pour expliquer en utilisant vos propres mots ce que représentent ces critères, leur qualité et leur défaut.

f) Utiliser la librairie **leaps** et appliquer les critères R^2_{adj}, *BIC* et C_p en adaptant le code ci-dessous.

```
library(leaps)
choix <- regsubsets(Quality~., data=wine, nbest=1, nvmax=11)
plot(choix, scale="adjr2", col="midnightblue")
leaps <- regsubsets(Quality~., data=wine, nbest=10)
summary(leaps)
```

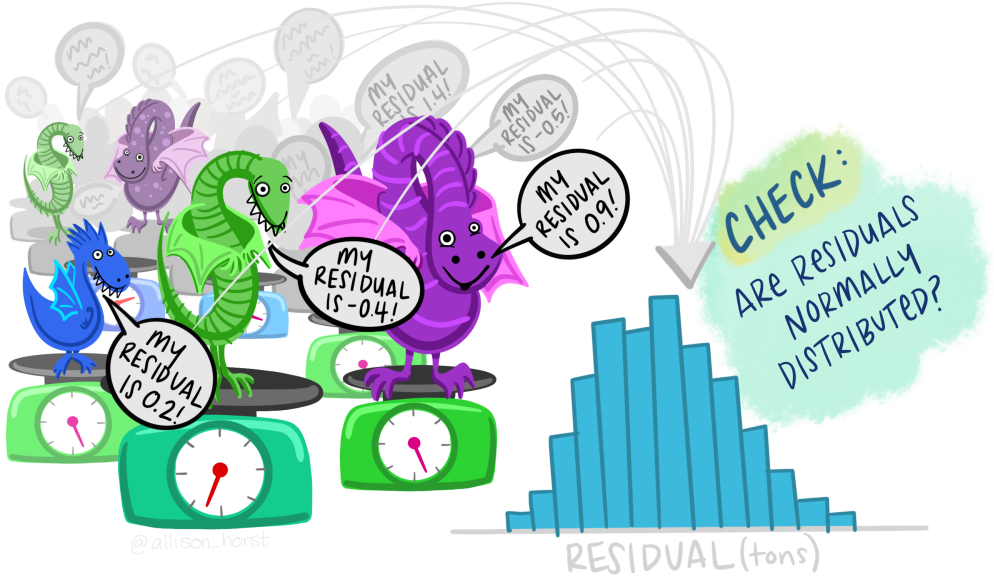
```
## Subset selection object
## Call: regsubsets.formula(Quality ~ ., data = wine, nbest = 10)
## 5 Variables  (and intercept)
##          Forced in Forced out
## Clarity      FALSE      FALSE
## Aroma        FALSE      FALSE
## Body         FALSE      FALSE
## Flavor       FALSE      FALSE
## Oakiness     FALSE      FALSE
## 10 subsets of each size up to 5
## Selection Algorithm: exhaustive
##           Clarity Aroma Body Flavor Oakiness
## 1  ( 1 )  " "     " "   " "  "*"    " "
## 1  ( 2 )  " "     "*"   " "  " "    " "
## 1  ( 3 )  " "     " "   "*"  " "    " "
## 1  ( 4 )  " "     " "   " "  " "    "*"
## 1  ( 5 )  "*"     " "   " "  " "    " "
## 2  ( 1 )  " "     " "   " "  "*"    "*"
## 2  ( 2 )  " "     "*"   " "  "*"    " "
## 2  ( 3 )  "*"     " "   " "  "*"    " "
```

```
## 2  ( 4 )  " "     " "   "*"  "*"    " "
## 2  ( 5 )  " "     "*"   " "  " "    "*"
## 2  ( 6 )  " "     "*"   "*"  " "    " "
## 2  ( 7 )  "*"     "*"   " "  " "    " "
## 2  ( 8 )  "*"     " "   "*"  " "    " "
## 2  ( 9 )  " "     " "   "*"  " "    "*"
## 2  ( 10 ) "*"     " "   " "  " "    "*"
## 3  ( 1 )  " "     "*"   " "  "*"    "*"
## 3  ( 2 )  "*"     " "   " "  "*"    "*"
## 3  ( 3 )  " "     " "   "*"  "*"    "*"
## 3  ( 4 )  "*"     "*"   " "  "*"    " "
## 3  ( 5 )  " "     "*"   "*"  "*"    " "
## 3  ( 6 )  "*"     " "   "*"  "*"    " "
## 3  ( 7 )  " "     "*"   "*"  " "    "*"
## 3  ( 8 )  "*"     "*"   "*"  " "    " "
## 3  ( 9 )  "*"     "*"   " "  " "    "*"
## 3  ( 10 ) "*"     " "   "*"  " "    "*"
## 4  ( 1 )  "*"     "*"   " "  "*"    "*"
## 4  ( 2 )  " "     "*"   "*"  "*"    "*"
## 4  ( 3 )  "*"     " "   "*"  "*"    "*"
## 4  ( 4 )  "*"     "*"   "*"  "*"    " "
## 4  ( 5 )  "*"     "*"   "*"  " "    "*"
## 5  ( 1 )  "*"     "*"   "*"  "*"    "*"
```



En regroupant les résultats obtenus par ces critères, quel est à votre avis le modèle à retenir ?

Ajuster ensuite le modèle retenu, dresser la table qui résume son ajustement puis effectuer une vérification des hypothèses inhérentes au modèle à l'aide des graphiques usuels.



`https://github.com/allisonhorst/stats-illustrations`.

Utiliser la fonction **compare_performance()** de la librairie **performance** pour comparer le modèle retenu et le modèle de regression linéaire initial dans lequel figurent toutes les variables explicatives. Que peut-on conclure ?

g) En utilisant le modèle retenu, donner la valeur de la qualité du vin prédite par le modèle en attribuant les valeurs 7.7, 6.7 et 3.7 aux variables **Aroma**, **Flavor** et **Oakiness**. Pour y parvenir, il convient d'utiliser la fonction **predict()**.

Exercice 5

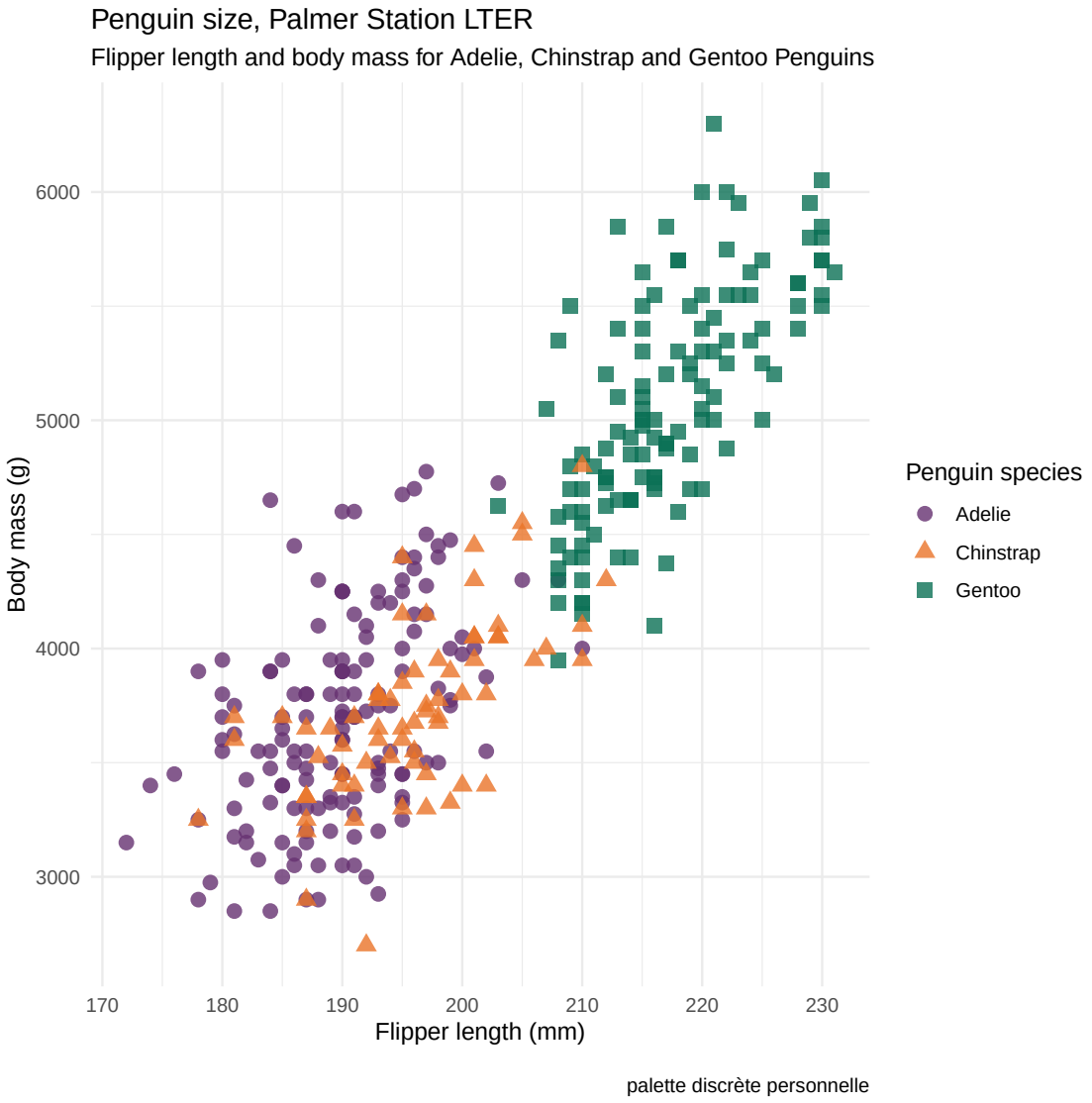
Effectuer une recherche bibliographique sur les arbres de régression puis résumer et synthétiser par vos propres mots cette méthode d'apprentissage supervisé de la science des données.

Exercice 6

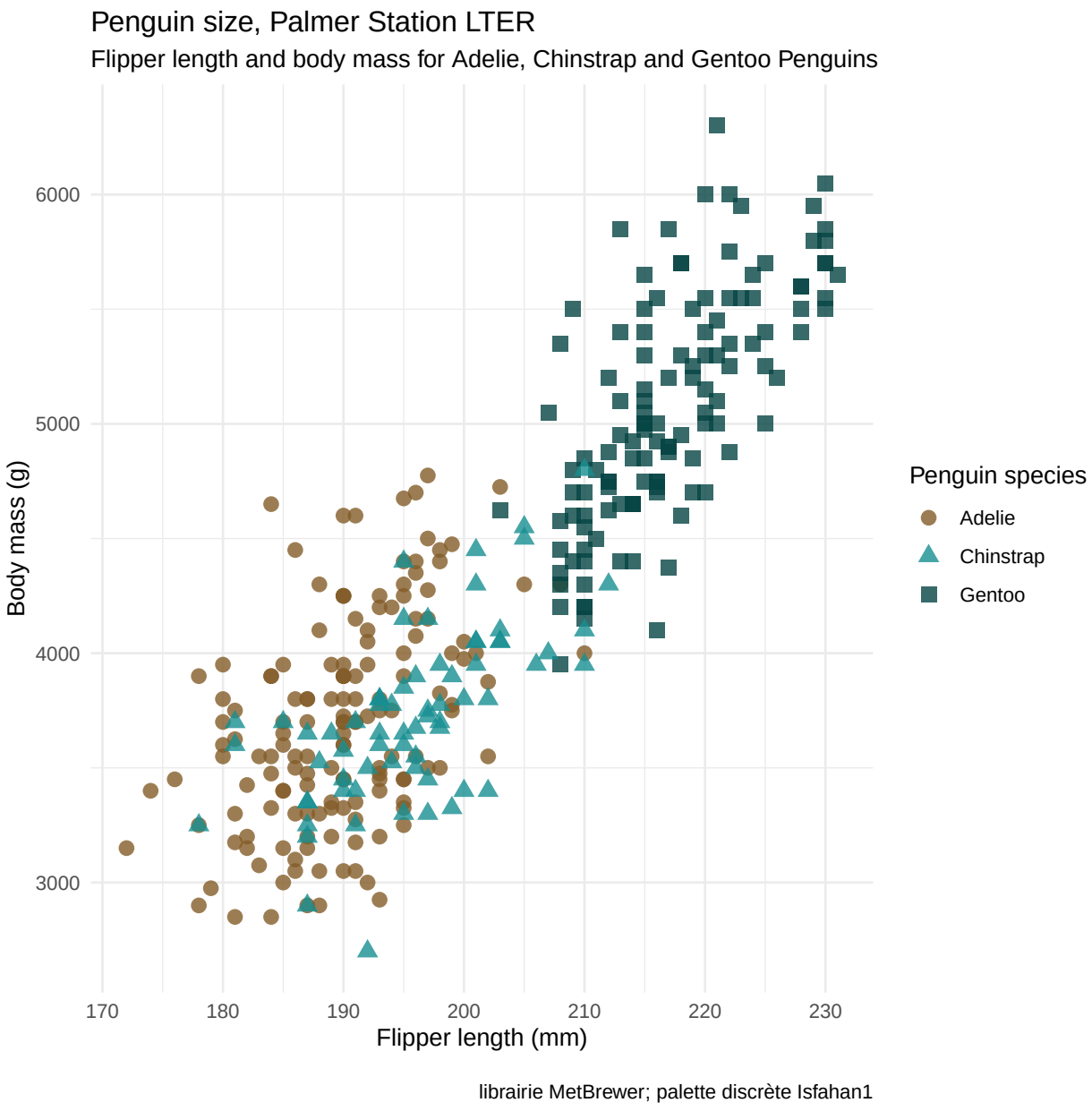
Plusieurs palettes de couleurs sont disponibles dans **R** ; certaines continues, d'autres discrètes. Dans cet exercice, on se propose de se familiariser avec l'utilisation de plusieurs palettes en partant du code permettant de tracer le graphique se trouvant en a) pour ensuite le modifier en utilisant différentes palettes de couleurs.

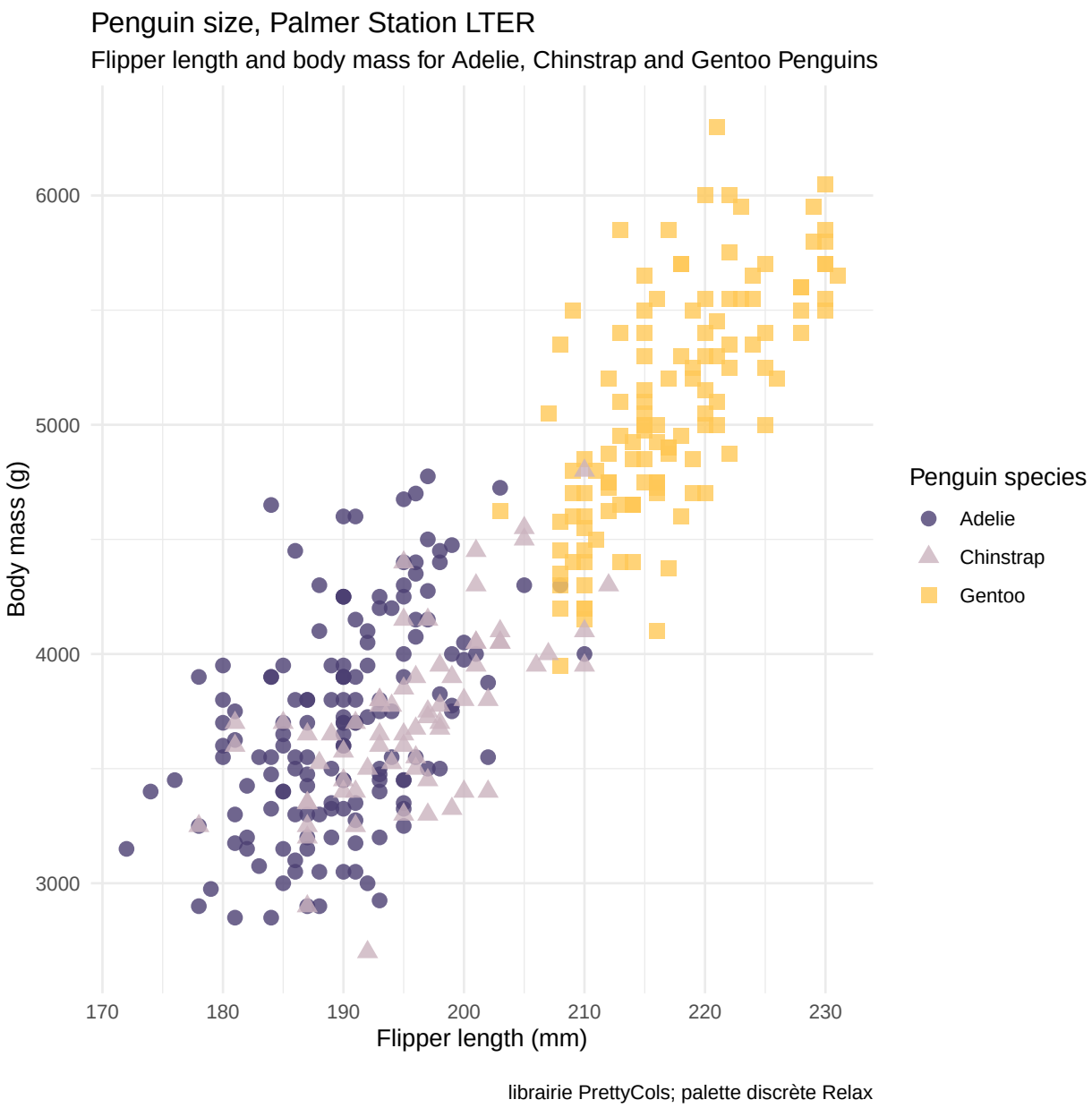
a) Tracer le graphique ci-dessous en utilisant le code suivant :

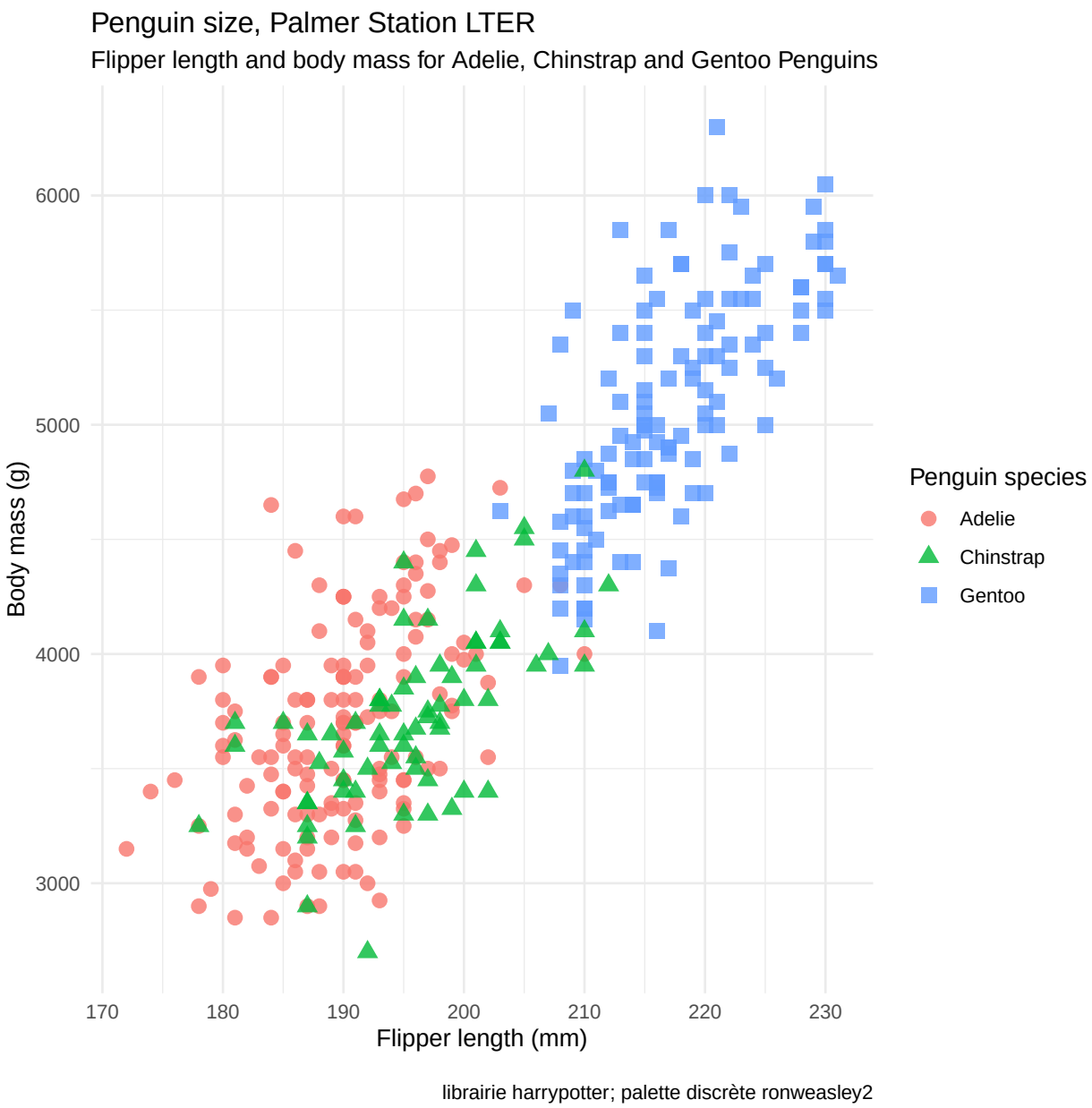
```
library(tidyverse)
library(palmerpenguins)
library(palettes)

discrete_pal <- pal_colour(c("#663171", "#EA7428", "#0C7156"))

ggplot(data=penguins, aes(x=flipper_length_mm, y=body_mass_g)) +
  geom_point(aes(color=species,
                 shape=species),
             size=3,
             alpha=0.8) +
  scale_colour_palette_d(discrete_pal) +
  labs(title = "Penguin size, Palmer Station LTER",
       subtitle = "Flipper length and body mass for Adelie, Chinstrap and Gentoo Penguins",
       x = "Flipper length (mm)",
       y = "Body mass (g)",
       color = "Penguin species",
       shape = "Penguin species",
       caption = "\n palette discrète personnelle") +
  theme_minimal()
```

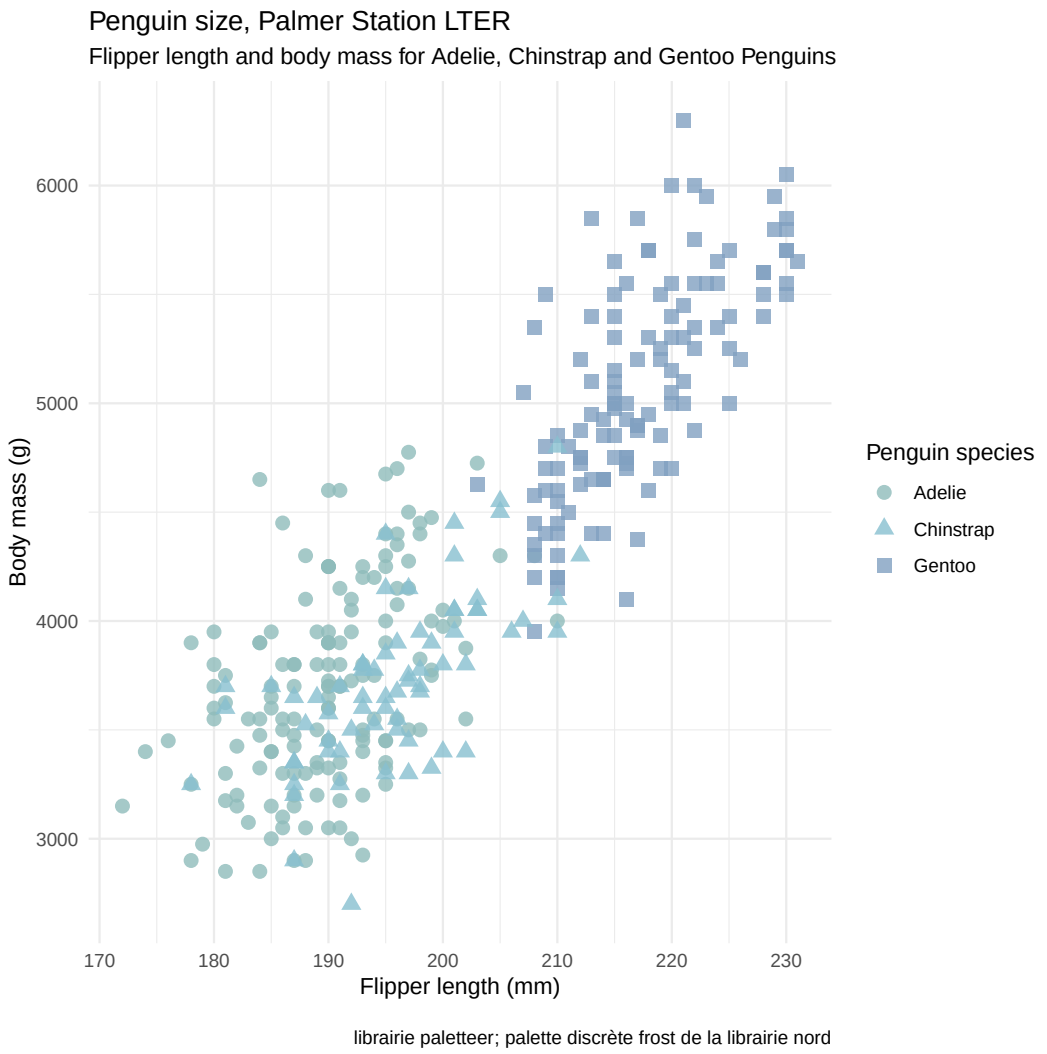


b) Reproduire les graphiques qui suivent en modifiant le code donné en a) en utilisant les palettes indiquées en légende.




Penguin size, Palmer Station LTER
Flipper length and body mass for Adelie, Chinstrap and Gentoo Penguins
Body mass (g)
6000
5000
4000
3000
170
180
190
200
210
220
230
Flipper length (mm)
Penguin species
Adelie
Chinstrap
Gentoo
librairie PrettyCols; palette discrète Relax


Penguin size, Palmer Station LTER
Flipper length and body mass for Adelie, Chinstrap and Gentoo Penguins
Body mass (g)
6000
5000
4000
3000
170
180
190
200
210
220
230
Flipper length (mm)
Penguin species
Adelie
Chinstrap
Gentoo
librairie harrypotter; palette discrète ronweasley2



c) À l'aide de la librairie **showtext** de **R**, on se propose dans cet exercice d'utiliser des polices inhérentes à Google.

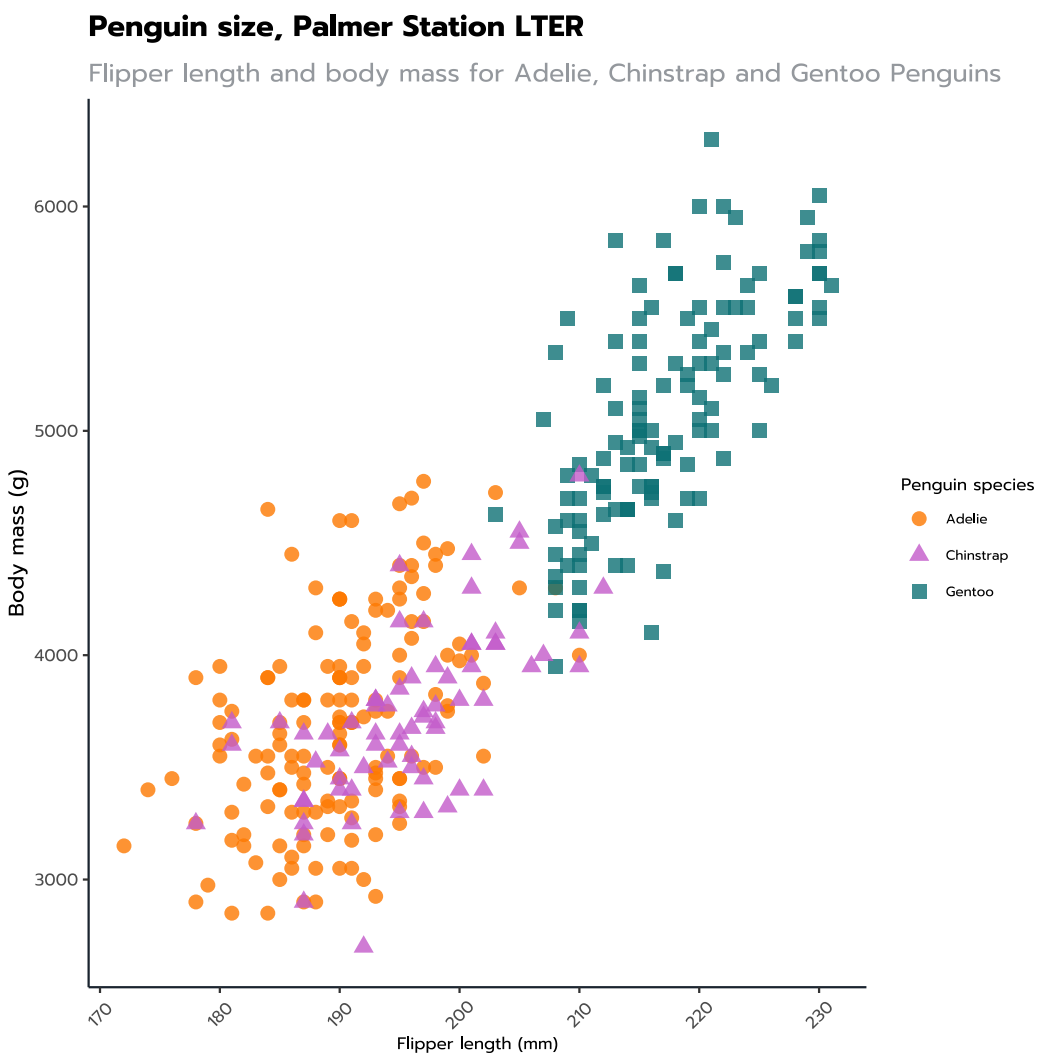
Dans le code permettant de tracer le graphique de la page suivante, on vous donne aussi un avant-goût de la fonctionnalité de la librairie **magrittr**[1] qui fait partie de l'ensemble des librairies **tidyverse**. La librairie **magrittr** repose sur un mécanisme grâce auquel des commandes peuvent être efficacement enchaînées, principalement à l'aide de l'opérateur %>% appelé pipe. Les pipes constituent un outil puissant pour exprimer clairement une séquence d'opérations multiples. Ainsi, l'utilisation des pipes nous aide à écrire des codes plus commodes à lire et à comprendre. À titre informatif, la librairie **magrittr** a été écrite à l'origine par Stephan Milton Bache.

Reproduire le graphique de la page suivante en utilisant le code ci-dessous :

```
library(showtext)

font_add_google(name = "Prompt", family = "prompt")
showtext_auto()

penguin_palette <- list("Adelie" = "#fd7901",
                        "Chinstrap" = "#c35bca",
                        "Gentoo" = "#0e7175",
                        "dark_text" = "#1A242F",
                        "light_text" = "#94989D")
```

1. Cette librairie porte le nom de **magrittr** en hommage au peintre surréaliste belge René Magritte (1898–1967) qui peignit un tableau dans lequel figure une pipe avec la légende "Ceci n'est pas une pipe".

```r
penguins %>%
  filter(!is.na(flipper_length_mm), !is.na(body_mass_g)) %>%
  ggplot(aes(x=flipper_length_mm, y=body_mass_g)) +
  geom_point(aes(color=species, shape=species), size=3, alpha=0.8) +
  scale_colour_manual(values=penguin_palette) +
  labs(title = "Penguin size, Palmer Station LTER",
       subtitle = "Flipper length and body mass for Adelie, Chinstrap and Gentoo Penguins",
       x = "Flipper length (mm)",
       y = "Body mass (g)",
       color = "Penguin species",
       shape = "Penguin species") +
  theme(
    panel.grid = element_blank(),
    panel.background = element_rect(fill="white"),
    text = element_text(family="prompt", size=8),
    plot.title = element_text(size=14, face="bold", margin=margin(b=10), hjust=0),
    plot.subtitle = element_text(size=12, colour = penguin_palette$light_text),
    axis.title.y = element_text(size = 10),
    axis.text = element_text(size = 8),
    axis.text.x = element_text(angle = 45, vjust = 0.5),
    axis.line.x = element_line(colour = penguin_palette$dark_text, linewidth = 0.5,
                               linetype = "solid"),
    axis.line.y = element_line(colour = penguin_palette$dark_text, linewidth = 0.5,
                               linetype = "solid"),
    panel.border = element_blank()
  )
```



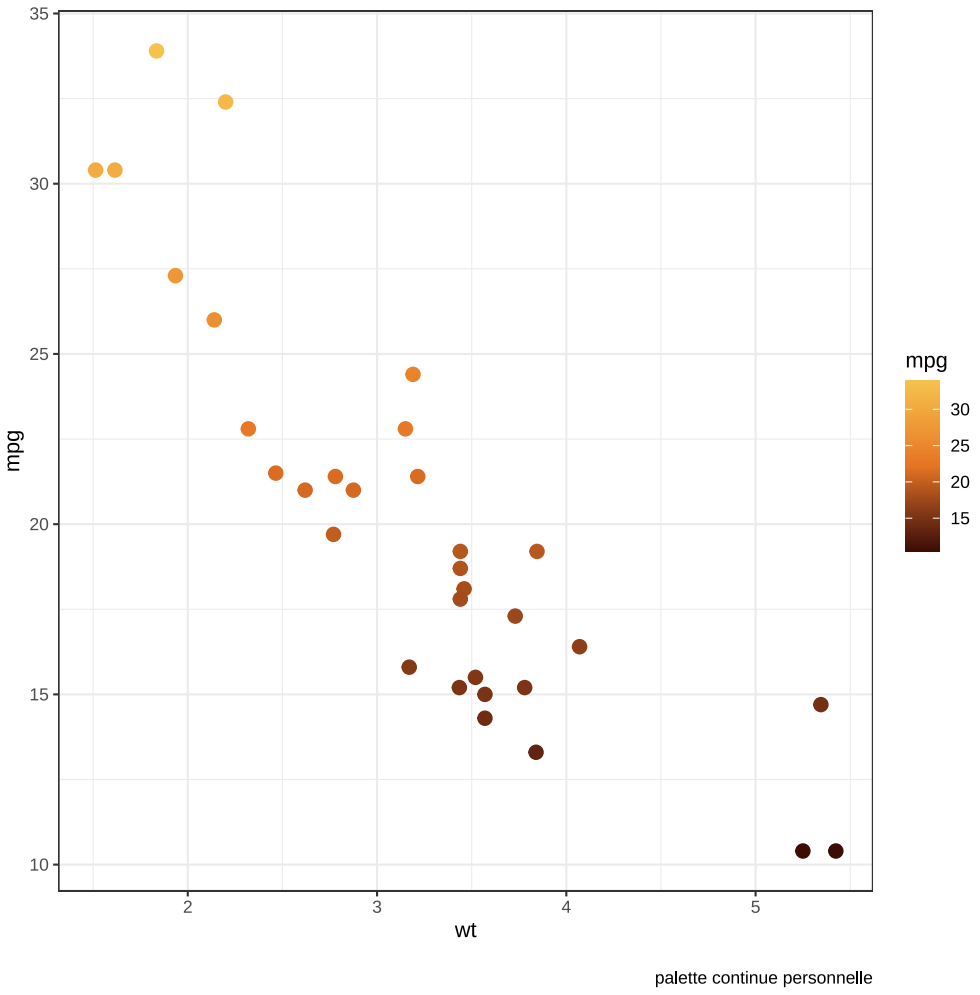
d) Tracer le même graphique en utilisant la police "Henny Penny" de la famille "henny", police disponible dans Google.

e) Que réalise la commande ci-dessous ?

```
... %>%
    filter(!is.na(flipper_length_mm), !is.na(body_mass_g)) %>%
    ...
```

f) Reproduire le graphique ci-dessous en utilisant une palette personnelle continue.

```
continuous_pal <- pal_colour(c("#3C0D03", "#E67424", "#F5C34D"))

ggplot(mtcars, aes(wt, mpg, colour = mpg)) +
  geom_point(size = 3) +
  scale_colour_palette_c(continuous_pal) +
  labs(
    caption = "\n palette continue personnelle"
    )
```



The R-Files

The truth is in the data