

Analyse exploratoire des données

De la théorie à la pratique (TP 1)

Jacques Zuber

This version: March 3, 2024

Préambule

Le logiciel de statistique qui sera utilisé dans les travaux pratiques est **R**, logiciel libre distribué sous les termes de la *GNU, General Public Licence*, au site web du **CRAN** (*Comprehensive R Archive Network*). 

(a) **R**.

Ce logiciel est disponible pour les systèmes d'exploitation Linux, Windows et Mac OS X. Des exécutables précompilés de la version actuelle **R-4.3.3** (Angel Food Cake) sont disponibles sur l'un des miroirs du CRAN. Les instructions à suivre pour les installer s'y trouvent.

Pour faciliter votre apprentissage du logiciel, Emmanuel Paradis et Julien Barnier ont écrit de bonnes documentations françaises pour **R**, “**R** pour les débutants” et “Introduction à **R**”, qui se trouvent dans la page [Moodle du cours](#). Un aide-mémoire des principales commandes de **R** figure dans le fichier “aide_memoire.pdf” qui se trouve dans la même page.

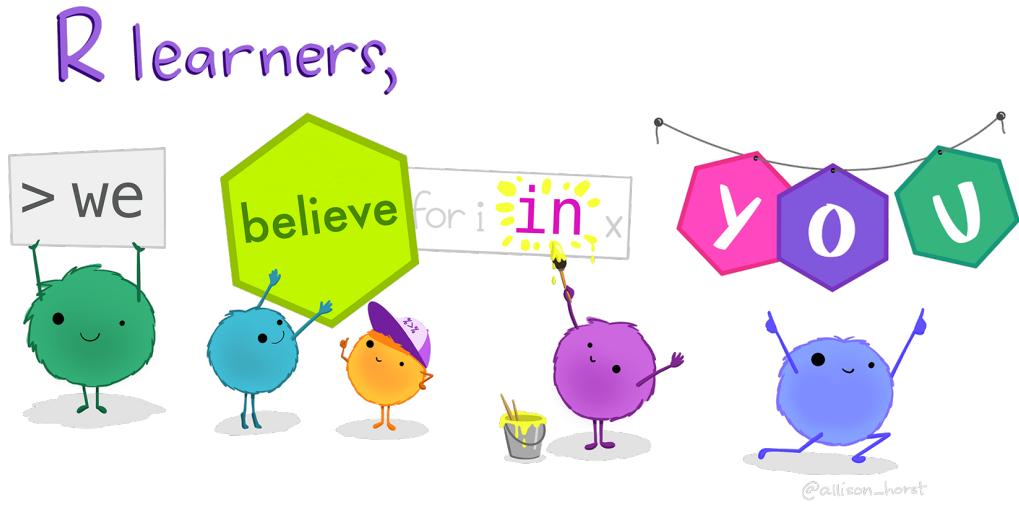


Figure 2: Dessin de Allison Horst.

Une aide en ligne existe directement dans **R**. Elle est très utile pour connaître l'utilisation des fonctions du logiciel. Plusieurs méthodes existent pour y accéder : la première en utilisant la commande

```
help("mean")
```

Une deuxième possibilité consiste à utiliser l'alias de la commande `help()`, un point d'interrogation suivi du nom de la commande, `?mean`, et finalement, une dernière variante revient à utiliser simplement le menu de l'interface **R-GUI** de **R**.

Un moteur de recherche pratique pour obtenir une aide supplémentaire et complète sur **R**, ses fonctions, ses librairies complémentaires et la programmation dans **R** est Rseek.org. L'utilisation de **R** peut aussi être facilitée en utilisant le [Quick-R](#).

Pour les utilisateurs de Linux, Windows et Mac OS X, il existe des éditeurs très pratiques comme **RStudio IDE** (*Integrated Development Environment*) et **VS Code** (*Visual Studio Code*) encore plus conviviaux que celui que vous propose **R** par défaut. Les scripts de commandes peuvent être archivés et accessibles à tout moment. Il est également possible d'afficher simultanément plusieurs fichiers contenant différents scripts et passer aisément de l'un à l'autre. La possibilité d'écrire des scripts, de les archiver, de les exécuter plusieurs fois en des temps différents est indéniablement un avantage par rapport à ce que vous proposent les logiciels à menus déroulants.

Il est conseillé d'installer **RStudio** ou **VS Code** pour être plus efficace dans sa programmation et pour travailler plus agréablement.

Il est aussi possible d'écrire de manière simple ses propres fonctions. Sans entrer dans les détails, une fonction **R** est écrite dans un fichier sauvé au format ASCII avec extension **.R**. Comme les autres langages, **R** possède des structures de contrôle qui ne sont pas sans rappeler celles du langage **C**.

The R-Files

The truth is in the data

Lorsque vous terminez votre session **R**, n'oubliez pas d'en sauver une image. Elle vous permettra de conserver les objets et de récupérer les dernières commandes utilisées.

Les étudiants *doivent rendre un rapport* du travail pratique dans lequel figurent les réponses aux questions posées ainsi que les graphiques tracés. Ils devront les rendre par groupes de deux dans la page **Moodle du cours** avant la date butoir. Les étudiants seront aussi interrogés sur les travaux pratiques aux travaux écrits. Les rapports doivent contenir une introduction présentant le sujet du travail pratique ainsi que ses objectifs et une conclusion formée d'une synthèse du travail pratique. Les travaux pratiques seront notés.

Les rapports doivent être rédigés à l'aide de **quarto®**, système calligraphique libre conçu notamment pour écrire des documents scientifiques. Il se base sur **Pandoc**, logiciel libre permettant la conversion de documents en ligne de commande.

Le fonctionnement de **quarto®**¹ pour créer un document est résumé dans la figure ci-dessous.

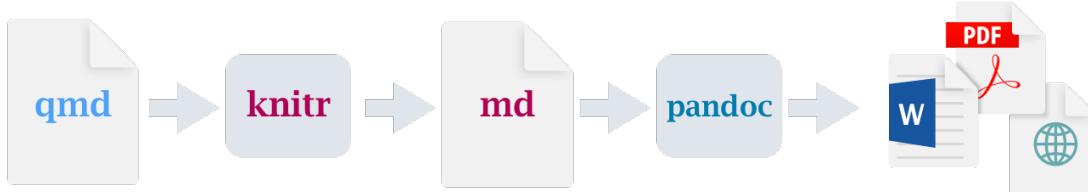


Figure 3: Source : quarto®.

Sans entrer dans les détails, **quarto®**² permet de générer des documents de manière dynamique en mélangeant texte et résultats obtenus à l'aide du code **R**. Les documents créés peuvent être notamment au format HTML, PDF, Word. Ainsi, **quarto®** est un outil très pratique pour exporter, communiquer et diffuser des résultats d'analyses statistiques. La librairie sous-jacente et fort élégante de **R** pour compiler, ou plutôt "tricoter" (*knit* en anglais), un document

¹Mickaël Canouil a réuni de nombreuses contributions, notamment des modèles, à l'adresse [quarto](#). Elles nous permettent de se familiariser avec le système calligraphique.

²Mickaël Canouil a réuni de nombreuses contributions, notamment des modèles, à l'adresse [quarto](#). Elles nous permettent de se familiariser avec le système calligraphique.

quarto® (.qmd) afin de visualiser le document généré est **knitr**. Elle crée entre autres un fichier .html dans lequel sont insérés les commandes, les sorties ainsi que les graphiques tracés. Vous avez entre vos mains ou à l'écran un document confectionné à l'aide de **knitr**³.

La librairie **knitr** exécute les morceaux de code de **R** contenus dans le document quarto® que vous avez écrit et crée un document **markdown** (.md) qui contient le code et sa sortie. Brièvement, **Markdown** est un langage de balisage très commode. Pandoc se charge ensuite de transformer le fichier *markdown* en un document de format désiré, comme par exemple en format .html, .pdf et Microsoft® Word.

L'objectif de ce travail pratique consiste à consolider les techniques d'analyse exploratoire et à améliorer la rédaction des rapports d'analyse de données. Différentes librairies seront utilisées dans ce travail pratique. Nous laissons le soin à l'étudiant·e de les installer en utilisant la fonction `install.packages()` de **R**.

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.

En route !

Exercice 1

Les étudiants suivant un cours de Mathématiques dans une école d'ingénierie ont passé l'examen de fin d'unité. Le cours était donné à 82 étudiants répartis en trois classes notées *A*, *B* et *F*. Les résultats obtenus figurent dans la table ci-dessous.

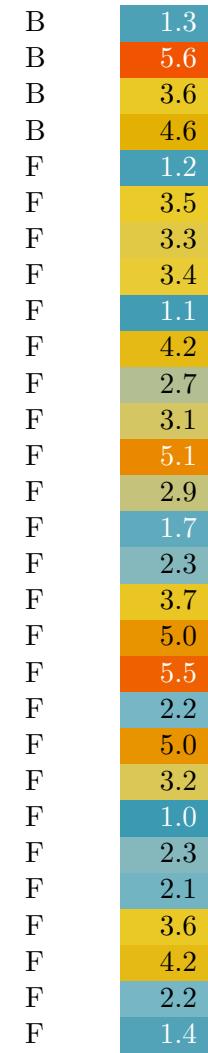
Table 1: Notes obtenues par les étudiants à l'examen de fin d'unité du cours de Mathématiques.

Table des notes selon les classes

Classe	Note
A	5.7
A	4.1
A	5.9
A	3.6
A	3.0
A	4.4
A	2.5
A	NA
A	5.3
A	3.0

³Des tutoriaux, démonstrations et exemples se trouvent à l'adresse [knitr](#).

A	3.7
A	4.7
A	4.0
A	4.6
A	4.1
A	4.2
A	3.2
A	3.1
A	4.7
A	4.9
A	3.7
A	3.6
A	4.1
A	NA
A	4.3
A	4.3
A	NA
A	3.3
B	4.6
B	4.1
B	NA
B	4.1
B	5.5
B	4.1
B	5.1
B	3.3
B	3.7
B	4.1
B	2.6
B	2.7
B	4.2
B	4.6
B	3.6
B	4.4
B	2.5
B	3.1
B	3.6
B	4.0
B	4.5
B	1.9
B	3.5
B	4.5
B	4.4



On se demande si une différence significative existe entre les trois classes à l'examen.

- Les données figurent dans le fichier `Notes.xlsx` qui se trouve dans la page Moodle du cours. Télécharger le fichier et enregistrer les données dans l'objet `examen` de **R** en utilisant la librairie `readxl`.
- Reconstituer les boîtes à moustaches en parallèle figurant ci-dessous.
- En se basant sur la Figure 4, existe-t-il une différence significative entre les trois classes à l'examen de fin d'unité ?
- Observe-t-on sur les boîtes à moustaches une différence entre les dispersions des trois classes ?

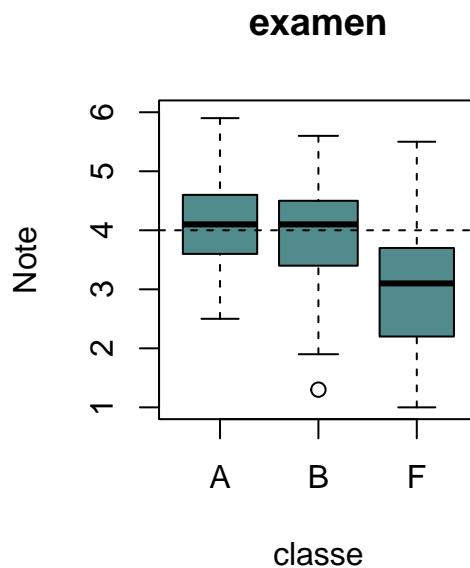


Figure 4: Boîtes à moustaches en parallèle.

e. Calculer les écarts-types des trois classes à l'aide des fonctions `by()` et `sd()`.

En se basant sur les écarts-types, existe-t-il une différence en dispersion entre les trois classes à l'examen de fin d'unité ?

f. Que peut-on déduire en comparant les conclusions établies en c., d. et e. ?

g. Calculer le résumé de la distribution des notes de la classe A à l'aide de la fonction `summary()`.

h. Déterminer l'asymétrie de la distribution des notes de la classe B à l'aide de la fonction `skewness()` de la librairie `e1071`.

i. Un autre graphique pour étudier les éventuelles différences entre les trois classes à l'examen de fin d'unité se trouve dans la Figure 5.

À votre avis, entre les boîtes à moustaches en parallèle et le graphique tracé ci-dessus, lequel est le plus approprié ?

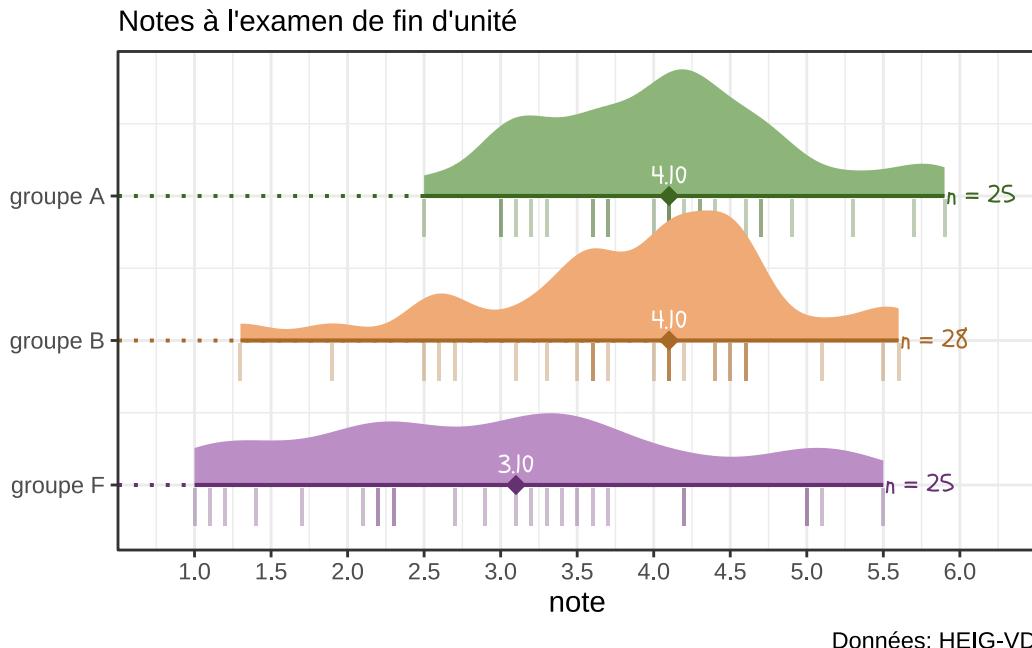


Figure 5: Diagramme de densité en parallèle.

Exercice 2

En 1912 sombra le Titanic, l'un des plus prestigieux paquebots transatlantiques de tous les temps. Le navire comprenait 2201 passagers parmi lesquels on compte environ 700 rescapés. Un tableau récapitulatif des survivants se trouve dans l'objet `Titanic` de la librairie `lattice` de **R**.

- a. Pour enregistrer puis visualiser les données, on peut utiliser les commandes

```
library(lattice)
data(Titanic)
Titanic
```

- b. En complétant le code ci-dessous, tracer un diagramme en barres des survivants et victimes du naufrage du Titanic selon le genre, l'âge et la classe.

```
titanic.bar<-barchart(Class~Freq|... + ... , data=as.data.frame(Titanic),
                        groups=..., stack=FALSE, layout=c(4,1),
                        auto.key=list(title="Survived", columns=2))
print(titanic.bar)
```

Que se passe-t-il si l'argument `stack` est fixé à `TRUE` ?

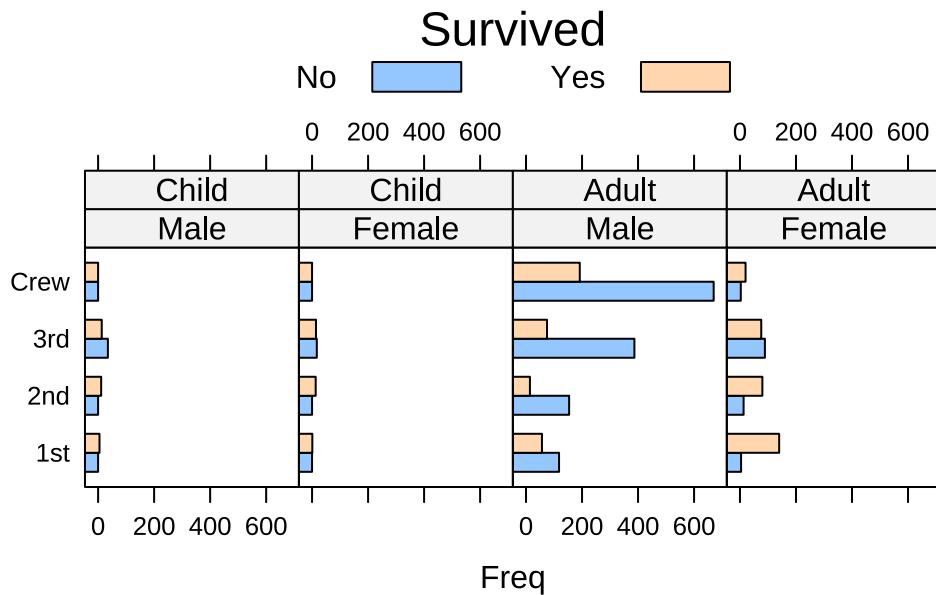


Figure 6: Diagramme en barres des survivants et victimes du naufrage du Titanic.

c. En termes de survivants du naufrage du Titanic, existe-t-il une différence

1. entre les genres ?
2. entre adultes et enfants ?
3. entre les types de classe ?

Justifiez clairement et précisément votre réponse.

d. Que proposeriez-vous pour améliorer la comparaison des genres, des âges et des classes des passagers du Titanic ?

Exercice 3

Les niveaux de deux protéines, le fibrinogène et la globuline, que contient le plasma sanguin ont été relevés sur 32 individus. Les données se trouvent dans la librairie **HSAUR2** de **R**.

a. Reconstituer le graphique de nuage de points Y : **globulin** versus X : **fibrinogen** figurant ci-dessous en utilisant la fonction **plot()**. Pour éviter des distorsions, utiliser un cadre carré et pour le symbole associé aux points, l'argument **pch=20**

b. En se basant sur le graphique, existe-t-il une relation entre les deux variables ? Dans l'affirmative, quel type de relation est-ce ?

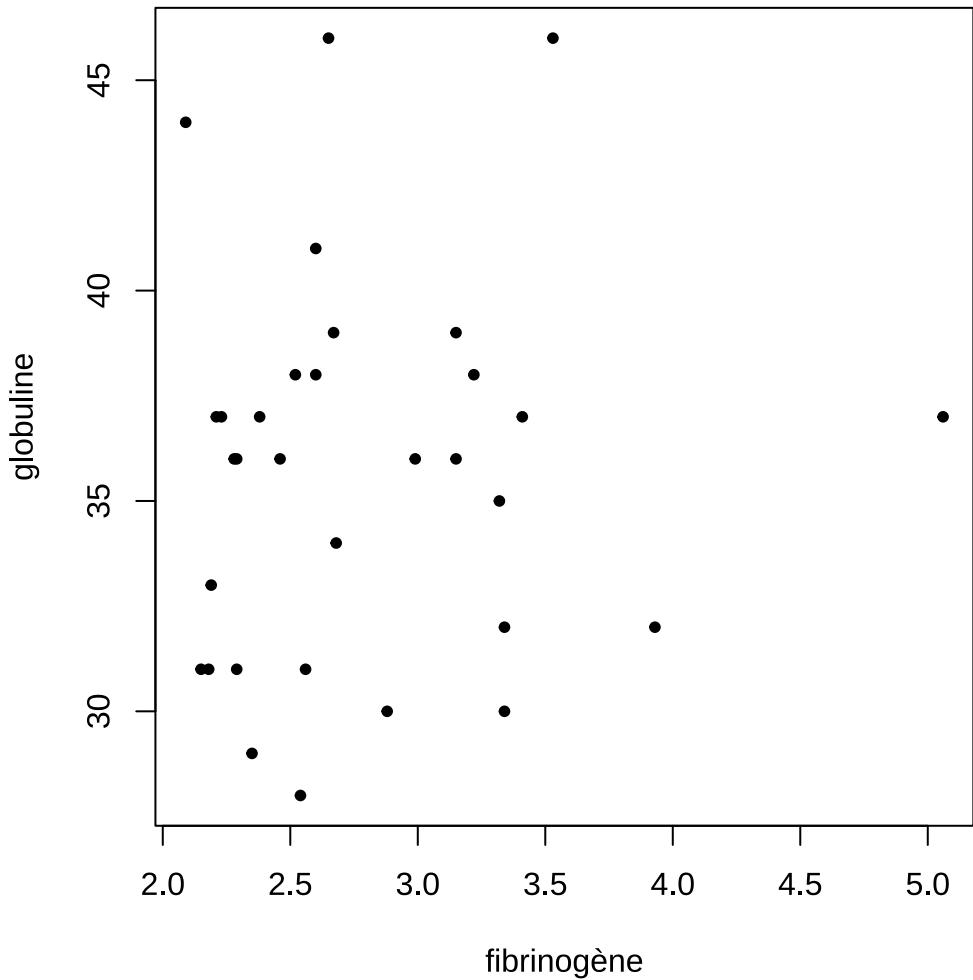


Figure 7: Nuage de points globuline versus fibrinogène.

c. Les variables explicatives (fibrinogène et globuline) sont utilisées pour prédire si la vitesse de sédimentation d'érythrocyte (ECR) est supérieure ou non à la valeur de référence de 20 mm par heure. À l'aide d'un modèle statistique, on estime la probabilité que la vitesse de sédimentation d'érythrocyte soit supérieure à la valeur de référence de 20. Pour l'illustrer graphiquement, on associe à chaque observation un cercle dont le rayon est proportionnel à la grandeur de la probabilité correspondante. Tracer le graphique à l'aide des commandes

```
plasma.glm<-glm(ESR~fibrinogen+globulin, data=plasma, family=binomial)
prob<-predict(plasma.glm, type="response")
par(pty="s")
plot(globulin~fibrinogen, data=plasma, xlim=c(2,6), ylim=c(25,55), pch=20,
      xlab="fibrinogène", ylab="globuline", main="")
symbols(plasma$fibrinogen, plasma$globulin, circles=prob, add=TRUE, fg="red",
        bg="orange")
```

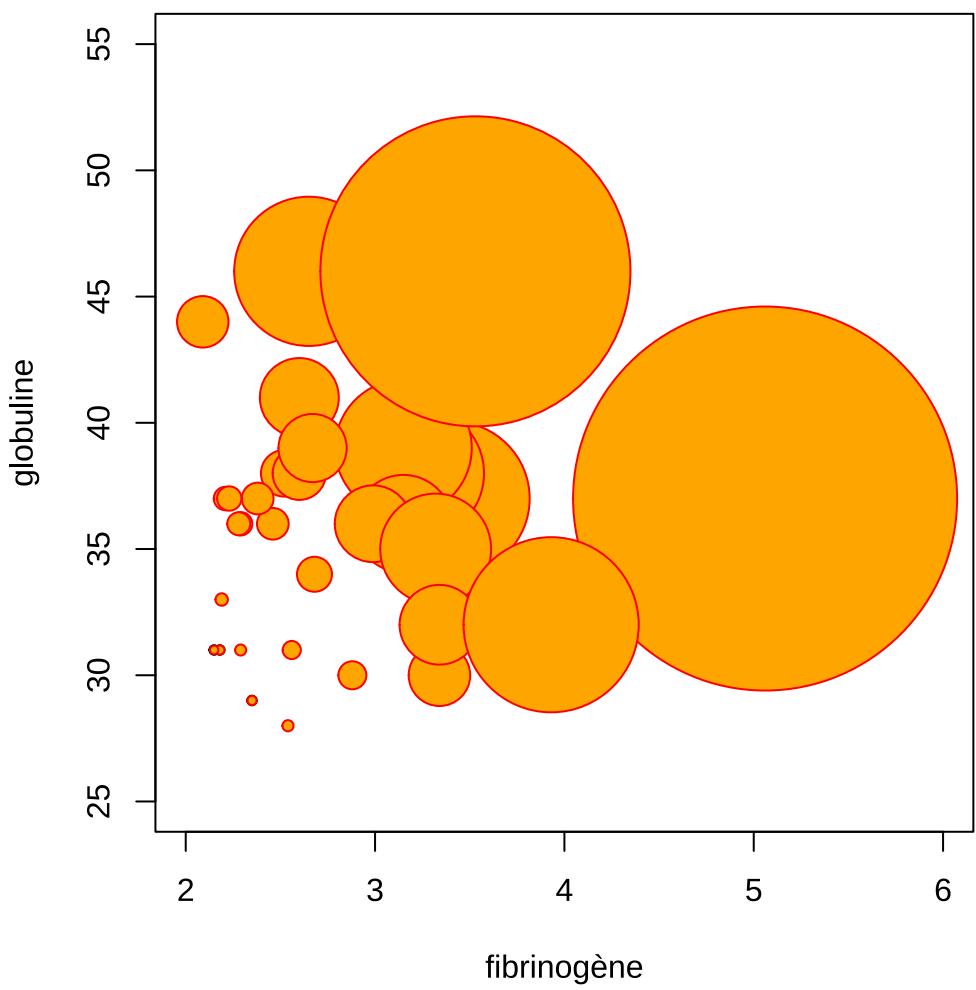


Figure 8: Nuage de points globuline versus fibrinogène.

En utilisant le graphique obtenu, que peut-on dire de l'influence des deux variables explicatives sur la probabilité que la vitesse de sédimentation d'érythrocyte soit supérieure ou non à 20

?

Exercice 4

Considérons les six séries de données :

x_1	10	8	13	9	11	14	6	4	12	7	5
y_1	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
y_2	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
y_3	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
x_4	8	8	8	8	8	8	8	19	8	8	8
y_4	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.50	5.56	7.91	6.89

- a. Les données se trouvent dans l'objet `anscombe` de **R**. Pour savoir ce qu'elles représentent, on peut utiliser la fonction `help()`.

`help("anscombe")`

- b. Calculer les coefficients de corrélation de (x_1, y_1) , (x_1, y_2) , (x_1, y_3) et (x_4, y_4) à l'aide de la fonction `cor()`.

Que remarque-t-on ?

- c. Créer la table de données `anscombe.1` en utilisant la commande

```
anscombe.1<-data.frame(x1=anscombe$x1, x4=anscombe$x4, y1=anscombe$y1,
                         y2=anscombe$y2, y3=anscombe$y3, y4=anscombe$y4)
```

- d. Reconstituer le tableau des corrélations ci-dessous à l'aide de la fonction `corrplot.mixed()` de la librairie `corrplot`.

- e. Créer le graphe des corrélations figurant ci-dessous en appliquant la fonction `correlate()` à l'objet `anscombe.1` puis la fonction `network_plot()` à la matrice des corrélations obtenue. Les fonctions `correlate()` et `network_plot()` se trouvent dans la librairie `corr` de **R**.

- f. Tracer le graphique des corrélations et nuages de points se trouvant ci-dessous à l'aide de la fonction `ggpairs()` de la librairie `GGally`. L'installation de la librairie `GGally` s'accompagne de celle de la librairie `ggplot2`.

Que proposeriez-vous pour améliorer ce graphiques ?

- g. Reconstituer la matrice de nuages de points ci-dessous.

- h. Commenter les résultats graphiques et numériques obtenus.



Figure 9: Tableau des corrélations.

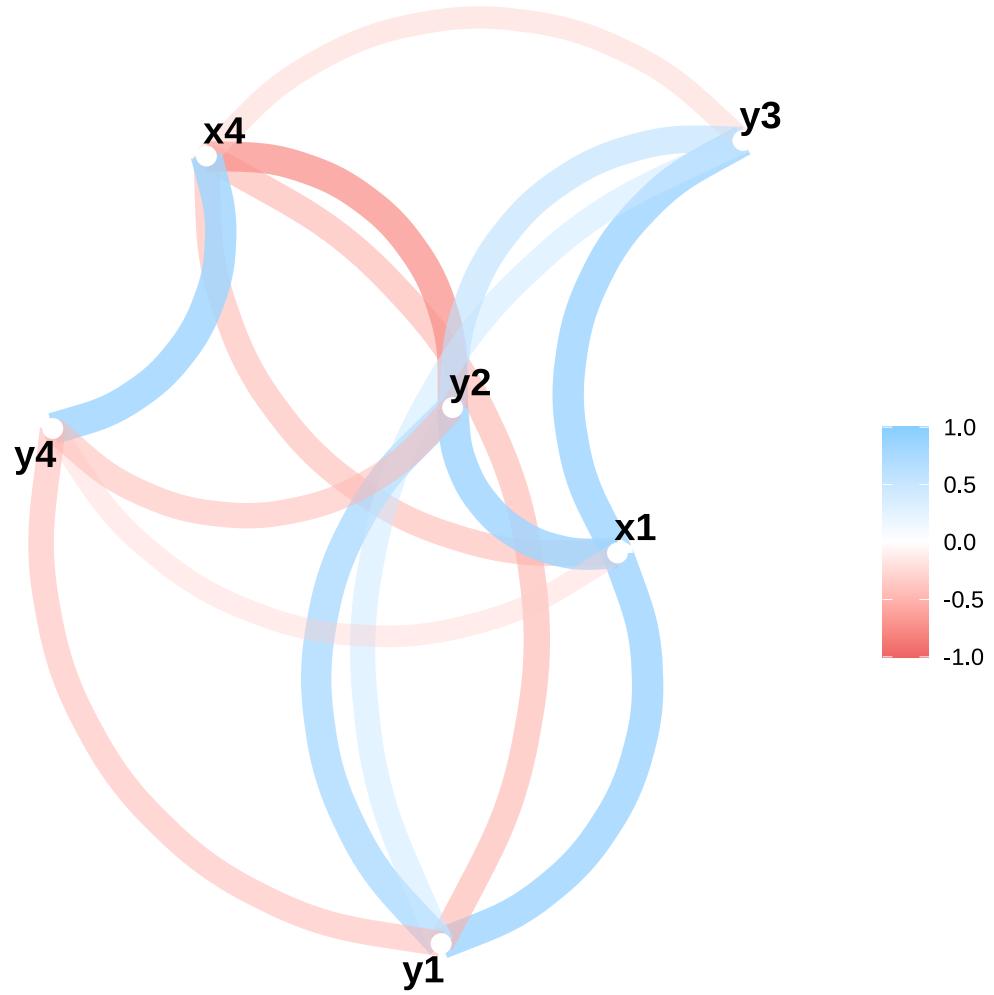


Figure 10: Graphe des corrélations.

Séries de Francis Anscombe

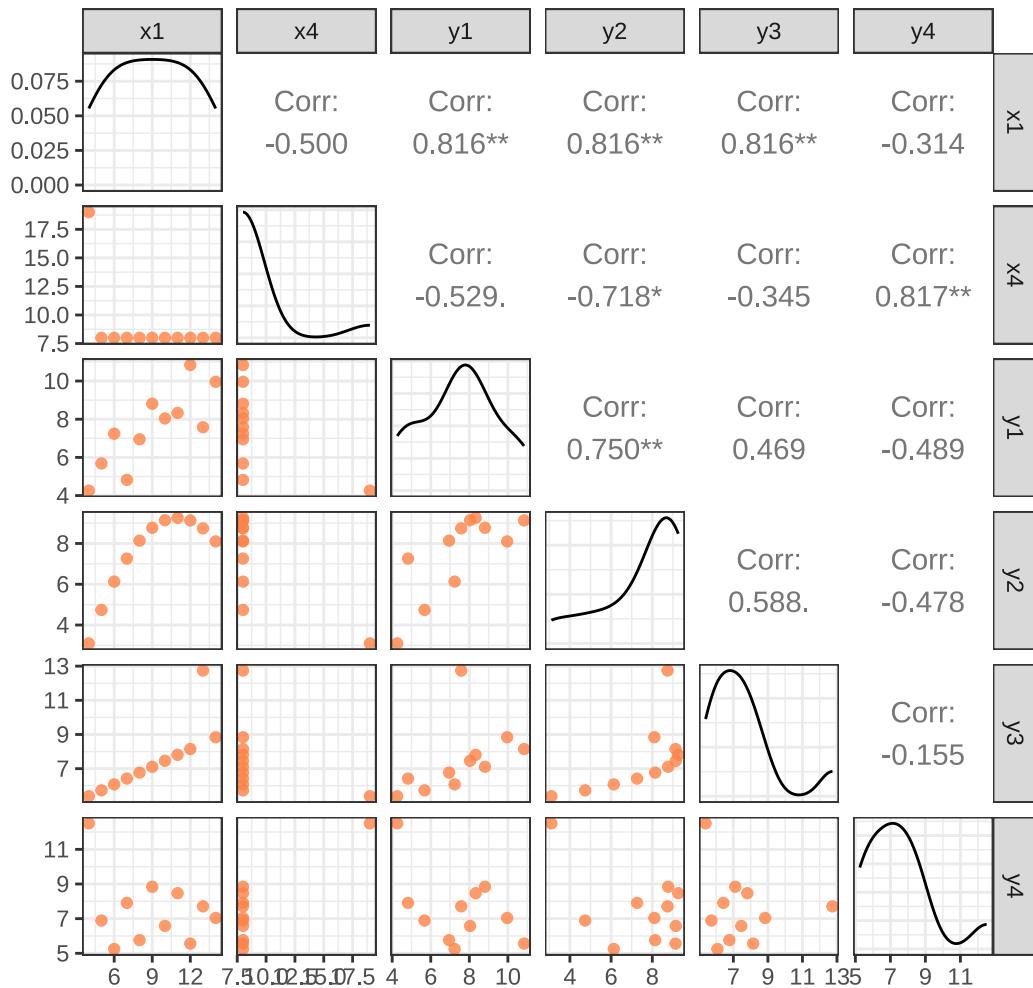


Figure 11: Graphique des corrélations et nuages de points.

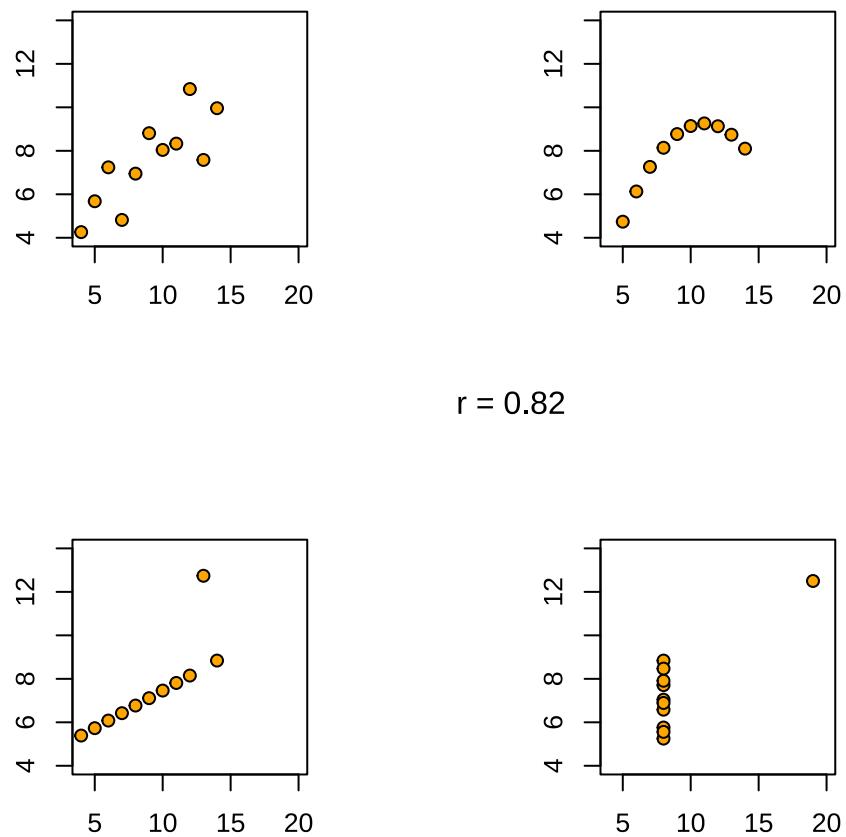
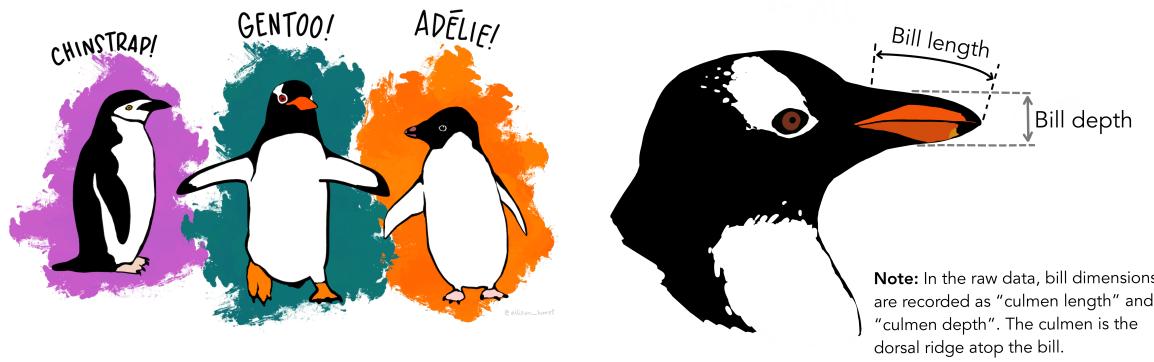


Figure 12: Graphique des nuages de points.

Exercice 5

Une étude a été réalisée sur des manchots vivant sur plusieurs îles de l'archipel Palmer en Antarctique. L'archipel fut nommé ainsi en hommage au capitaine Nathaniel Palmer qui navigua dans ces eaux en 1820. On a relevé différentes mesures et caractéristiques des manchots. Elles ont été mises à disposition par le Dr. Kristen Gorman et le programme de recherche écologique à long terme (LTER) de la station Palmer. Huit variables ont été relevées sur 344 manchots :

- l'espèce des manchots (Adélie, Chinstrap et Gentoo);
- l'île de l'archipel sur laquelle vivent les manchots (Biscoe, Dream et Torgersen);
- la longueur du bec (*bill length*) en millimètres;
- l'épaisseur du bec (*bill depth*) en millimètres;
- la longueur des nageoires (*flipper length*) en millimètres;
- la masse corporelle (*body mass*) en grammes;
- le sexe (male, femelle);
- l'année de l'observation.



(a) Dessin de Allison Horst.

Les données se trouvent dans l'objet `penguins` de la librairie `palmerpenguins`.

L'objectif de cet exercice consiste à se familiariser avec la librairie `ggplot2`. Elle a été développée par Hadley Wickham et permet de visualiser efficacement les données et de tracer des graphiques très sophistiqués. Plus précisément, elle explicite les liens conceptuels entre graphiques et analyses statistiques. Sa syntaxe est cohérente et puissante même si elle peut paraître de prime abord particulière. Elle se base sur un ensemble de composants indépendants qui peuvent être combinés de différentes manières.⁴

⁴De plus amples informations se trouvent dans [Tidyverse](#).

En fait, la librairie `ggplot2` met en oeuvre une “grammaire graphique” qui avait été introduite en théorie par Leland Wilkinson. Cette librairie dispose d'une souplesse remarquable dans son utilisation. Elle nécessite cependant l'apprentissage d'un “mini-langage” supplémentaire.

Pour de plus amples informations sur la librairie `ggplot2`, utilisez le lien [Create Elegant Data Visualisations Using the Grammar of Graphics](#) • `ggplot2 {ggplot2}` et pour la fonction `ggplot()` le lien [Create a new ggplot](#) — `ggplot` • `ggplot2 ggplot2::ggplot()` .

a. La librairie `skimr` nous permet d'avoir un aperçu global des caractéristiques des variables concernant les manchots. Installer puis activer cette librairie dans votre session actuelle de **R**. Les caractéristiques des variables s'obtiennent en utilisant la fonction `skim()`.

b. Quelle variable possède le plus de valeurs manquantes ?

c. Quelle espèce est la plus représentée ?

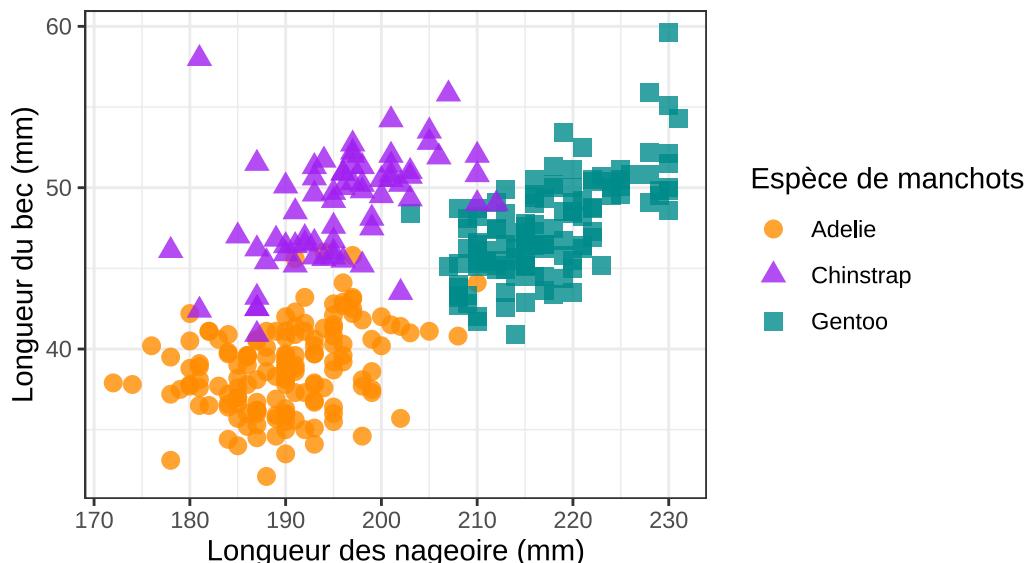
d. Nous nous proposons de tracer des graphiques de nuages de points à l'aide de la librairie `ggplot2` qu'il faut d'abord installer si nécessaire puis activer dans votre session.

Tracer le nuage de points de la longueur du bec (*bill length*) versus la longueur des nageoires (*flipper length*) en utilisant les commandes ci-dessous.

```
ggplot(data = penguins, aes(x = flipper_length_mm, y = bill_length_mm)) +
  geom_point(aes(color = species, shape = species), size = 3, alpha = 0.8) +
  scale_color_manual(values = c("darkorange","purple","cyan4")) +
  labs(title = "Taille des manchots, Palmer Station LTER",
       subtitle = "Longueur des nageoires et longueur du bec chez les manchots Adelie, Chinstrap et Gentoo",
       x = "Longueur des nageoires (mm)",
       y = "Longueur du bec (mm)",
       color = "Espèce de manchots",
       shape = "Espèce de manchots") +
  theme_bw()
```

Taille des manchots, Palmer Station LTER

Longueur des nageoires et longueur du bec chez les manchots Adelie, Chinstrap et Gentoo



Existe-t-il une relation entre la longueur du bec et la longueur des nageoires ? Dans l'affirmative, de quelle nature est-elle ?

- e. Remarque-t-on des observations inhabituelles dans le graphique de nuage de points ?
- f. Quelle valeur attribueriez-vous à la longueur des nageoires pour distinguer les manchots de Gentoo des deux autres espèces ?
- g. Déterminer la corrélation entre la longueur du bec et la longueur des nageoires.
- h. Modifier le code du nuage de points pour faire en sorte qu'on distingue également dans le graphique les îles où habitent les manchots.

Sur quelle île vivent exclusivement les manchots de Gentoo ?

- i. Pour répondre à cette question, on peut également construire le graphique de points tracé ci-dessous.

```
ggplot(penguins, aes(x = island, y = species, color = species)) +
  geom_jitter(size = 3) +
  scale_color_manual(values = c("darkorange","purple","cyan4")) +
  labs(x = "Îles",
       y = "Espèce de manchots",
       color = "Espèce de manchots")
```

Taille des manchots, Palmer Station LTER

Longueur des nageoires et longueur du bec chez les manchots Adelie, Chi

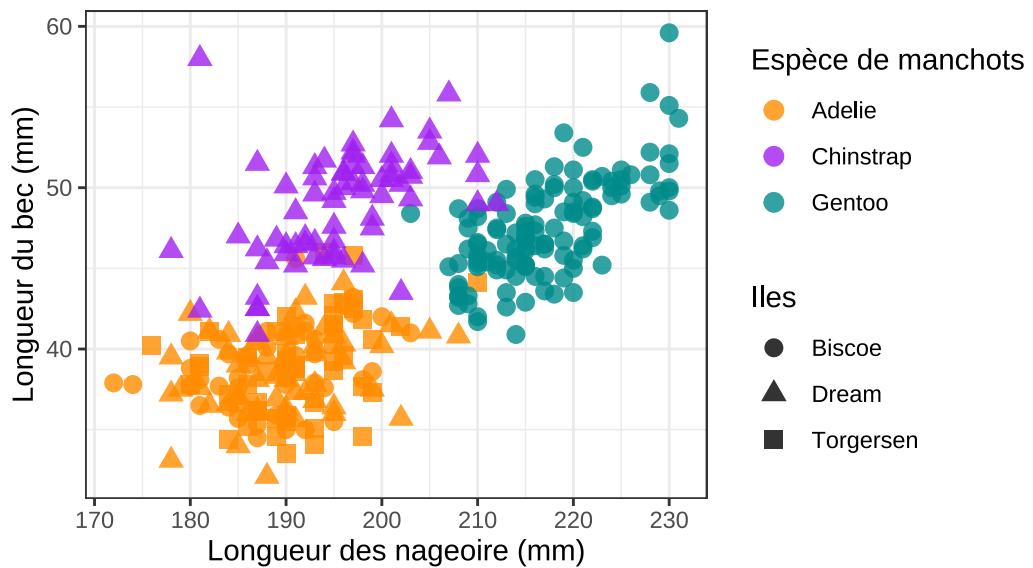


Figure 14: Graphique des nuages de points.

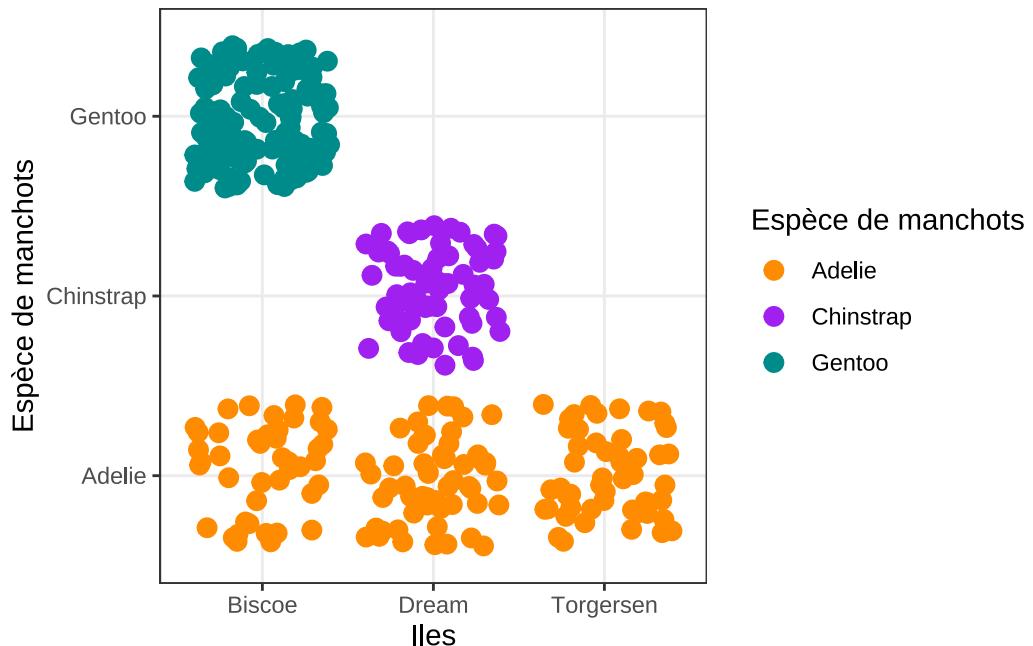
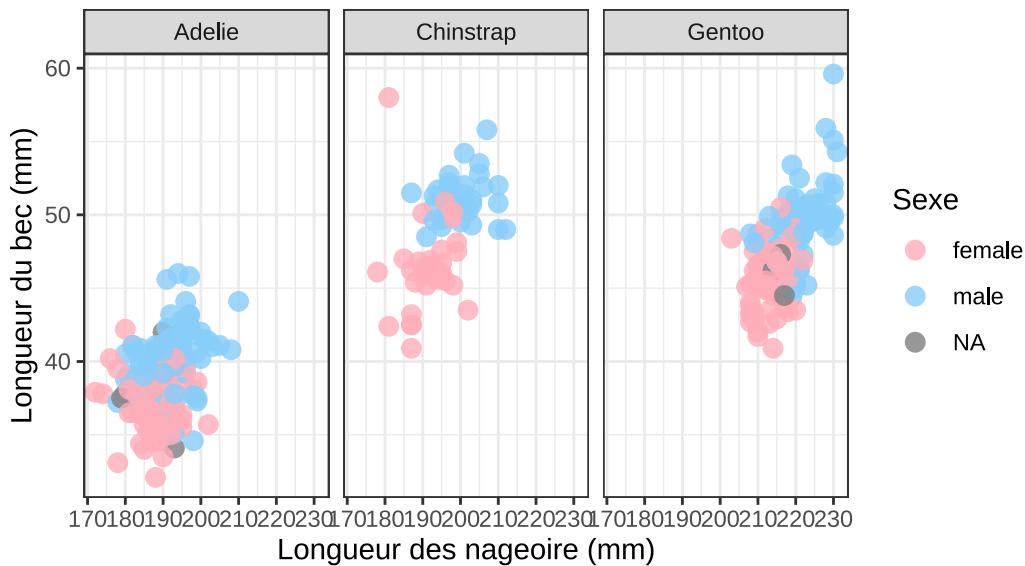


Figure 15: Graphique en points des manchots selon les îles.

- j. Tracer le nuage de points de la longueur du bec (*bill length*) versus la longueur des nageoires (*flipper length*) selon l'espèce et le sexe des manchots en utilisant le composant `facet_wrap()` de la fonction `ggplot()`.

Taille des manchots, Palmer Station LTER

Longueur des nageoires et masse corporelle chez les manchots Adelie, Cr



Exercice 6

- a. Les données que nous allons utiliser dans cet exercice contiennent des informations sur les dossiers de candidature à une bourse du Consortium⁵. Elles se trouvent dans le fichier `isc_grants.csv` figurant dans la page [Moodle du cours](#). Copiez-le dans votre répertoire de données en utilisant la fonction `read_csv()` de la librairie `readr`, l'une des composantes de `tidyverse`.
- b. Nous allons construire une carte à cases (*treemap*) des candidatures à l'obtention d'une bourse du Consortium en utilisant la librairie `treemapify` de *R*.

Pour sauvegarder la carte à cases, nous utilisons la librairie `camcorder`. Elle permet de sauvegarder tous les graphiques construits avec la librairie `ggplot2` non seulement au cours de cette session mais aussi au cours de plusieurs sessions. Dans cet exercice, la carte à cases sera enregistrée dans le répertoire *ISCgrant* qui sera automatiquement créé à l'aide de la commande

```
gg_record(dir="ISCgrant", device="png", width=9, height=8, units="in", dpi=320)
```

⁵Par intérêt, un coup d'oeil à l'adresse de [tidytuesday](#).



Figure 16: Les délices de la librairie ggplot2

Les codes des couleurs des cases de la carte figurent dans le vecteur `pal` donnée ci-dessous.

```
pal<-c("#002870", "#005A87", "#078788", "#A5A63C", "#DE9704", "#C45D27", "#AD3518", "#990C00")
```

Tracer la carte à cases ci-dessous en complétant les commandes données au-dessous du graphique.

```
ggplot(..., aes(area=..., fill=factor(...), subgroup=year)) +
  geom_treemap(radius=unit(0.2, "line"), color="white", size=2) +
  geom_treemap_text(aes(label=paste0(title, "\n", proposed_by, "\n\n", scales::dollar(funded),
  geom_treemap_subgroup_text(aes(label=year), color="white", grow=TRUE, alpha=0.25) + ①
  scale_fill_manual(values=pal) +
  labs(
    title="Bourses accordées selon les années",
    caption="Source: Comité de pilotage du consortium R",
    fill="year") +
```

Parmi les missions que remplit le Consortium **R** figure celle qui consiste à favoriser la maintenance et le développement du logiciel en accordant des bourses à la Fondation **R**, à la communauté des utilisateurs de **R** et à des organisations.



Bourses accordées selon le temps



Source: Comité de pilotage du consortium R

Figure 17: Carte à cases des bourses accordées par le consortium **R**.

```

theme_void() +
theme(
  legend.position="none",
  plot.background=element_rect(fill="grey99", color=NA),
  plot.title=element_text(size=30, face="bold", color="#0D2765", margin=margin(0, 0, 5, 0)),
  plot.caption=element_text(color="#0D2765"),
  plot.margin=margin(10, 10, 10, 10)
)

```

- ① Avez-vous aussi une mise en garde ?

Le rapport doit

- être rédigé avec soin en utilisant *quarto®*;
- contenir une introduction dans laquelle se trouvent les objectifs de l'analyse de données ainsi qu'une conclusion pour synthétiser le travail pratique;
- contenir les réponses aux questions posées;
- contenir les commandes de **R** utilisées, les résultats et graphiques obtenus;
- être rendu sur la page Moodle du cours en format .html ou .pdf avant la date butoir.

```

Sys.time()
sessionInfo()

```