TP STATISTIQUE

RAMDÉ Ismaïl

13, octobre, 2020

Analyse de données du parc de Monteshinho au Portugal

Exercice 1

1) Lecture des données

```
setwd("~/Bureau/Statistique/DM_statistique")
data <- read.csv("6414_forestfires.csv")
```

Visualisation des données

```
str(data)
```

2) Affichage des données avec print

```
print(data)
```

- On remarque dans les données que certains jours (variable `day`) de la semaine sont écrites en toutes lettres pendant que d'autres ne le sont pas. On observe la même chose pour la variable mois (`month`). Nous remarquons aussi qu'il y a des valeurs Manquantes (NA). Ces erreurs pourraient biaisées nos différents calculs et analyses.

La solution dans cette situation est de recoder les modalités des variables respectives de telle sorte qu'elles soit uniformes. Voir le code ci-dessous :

```
data <- data %>% mutate(month = replace(month, month == "September","sep"))
data <- data %>% mutate(month = replace(month, month == "April", "apr"))
data <- data %>% mutate(month = replace(month, month == "July", "jul"))

data <- data %>% mutate(day = replace(day, day == "Wednesday", "wed"))
data <- data %>% mutate(day = replace(day, day == "Friday", "fri"))
data <- data %>% mutate(day = replace(day, day == "Thursday", "thu"))
```

- L'humidité relative (`RH`) est comprise de 15.0 à 100 or les valeurs des lignes 127 et 288 ne le sont pas. Pour ce faire nous allons les remplacer par la moyenne des mois respectifs.

```
data$RH[[127]] <- mean(subset(data, data$month == "sep")$RH)
data$RH[[288]] <- mean(subset(data, data$month == "aug")$RH)
```

- On observe une discordance entre les valeurs de la vitesse de vent (`wind`) et celle de la ligne 187 qui est très élevée par rapport aux autres. Nous la remplacerons donc par la moyenne du mois.

```
data$wind[[187]] <- mean(na.omit(subset(data, data$month=="sep"))$wind)
test <-mean(subset(data, data$month=="sep")$wind)
```

- On observe également des températures très élévées pour le mois de Août (lignes 33, 64 et 289) par rapport aux autres. De sucroit elles ne reflètent pas la réalité. Nous allons les remplacer par la moyenne

du mois tout en ne tenant pas compte des des trois lignes.

```
data$temp[[33]] <- mean(subset(data,c(data$month == "aug" & data$temp < 40))$temp)
data$temp[[64]] <- mean(subset(data, c(data$month == "aug" & data$temp < 40))$temp)
data$temp[[289]] <- mean(subset(data, c(data$month == "aug" & data$temp < 40))$temp)
```

3) Resumé des données

```
summary(data)
```

On voit tout de suite qu'il y a des valeurs manquantes pour les variables `temp` (température) et `wind` (vitesse du vent).

Proposition de correction :

Pour palier à ce problème, nous avons deux possibilités:

a) On peut décider de laisser la valeurs manquantes (NA) et en tenir compte dans tout nos calculs (lignes de commandes) pour ne pas erroner nos résultats.

b) Ou alors les remplacer par les moyennes respectives. Nous choisirons cette option car la taille de notre data frame nous le permet.

```
data$temp[[20]] <- mean(na.omit(subset(data, data$month=="sep")$temp))
data$wind[[24]] <- mean(na.omit(subset(data, data$month=="sep")$wind))
```

```
summary(data)
```

4) Création des matrices corrélation, de scatter plot et analysons

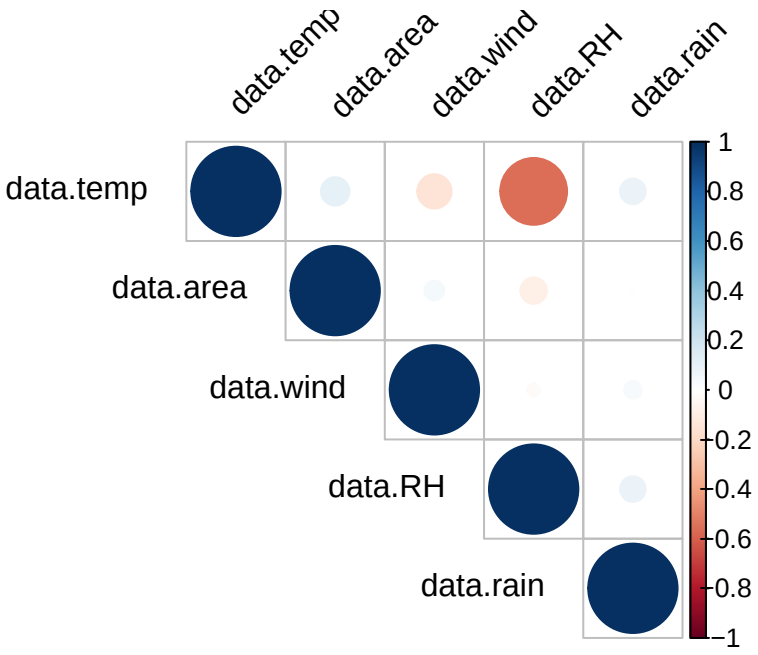
Pour représenter la matrice de scatter plot nous avons décidés de ne pas inclure les cordonnées(`X` et `Y`), le jour(`day`) et le mois(`month`) car elles n'ont pas vraiment d'impacts majeurs sur les autres variables. Pour cela nous avons crées une data frame avec les variables a représenter.

- Matrice de corrélation:

```
# création de la data frame
fr <- data.frame(data$temp, data$RH, data$wind, data$rain, data$area)

# matrice de corrélation
mcor <- cor(fr)
# visualisation dela matrice de corrélation
corrplot(mcor, type="upper", order="hclust", tl.col="black", tl.srt=45, title="Matrice de corrélation (
```

Matrice de corrélation (corrélogramme)



Ce diagramme met en corrélation les variables deux à deux. La couleur bleue est utilisée pour illustrer la corrélation positive et la couleur rouge pour illustrer la corrélation négative. La taille et l'intensité des cercles sont proportionnelles aux coefficients de corrélations. En effet il y a une:

a) forte corrélation négative entre les variables `temp` (température) et `RH` (humidité relative)

b) faible corrélation négative entre les variables `temp` (température) et `wind` (vitesse du vent)

c) faible corrélation positive entre les variables `temp` (température) et `area` (surface de foret brûlée)

d) faible corrélation positive entre les variables `temp` (température) et `rain` (pluviométrie)

e) faible corrélation négative entre les `area` (surface de foret brûlée) et `RH` (humidité relative)

f) faible corrélation positive entre les `RH` (humidité relative) et `rain` (pluviométrie)

- Matrice de scatter plot:

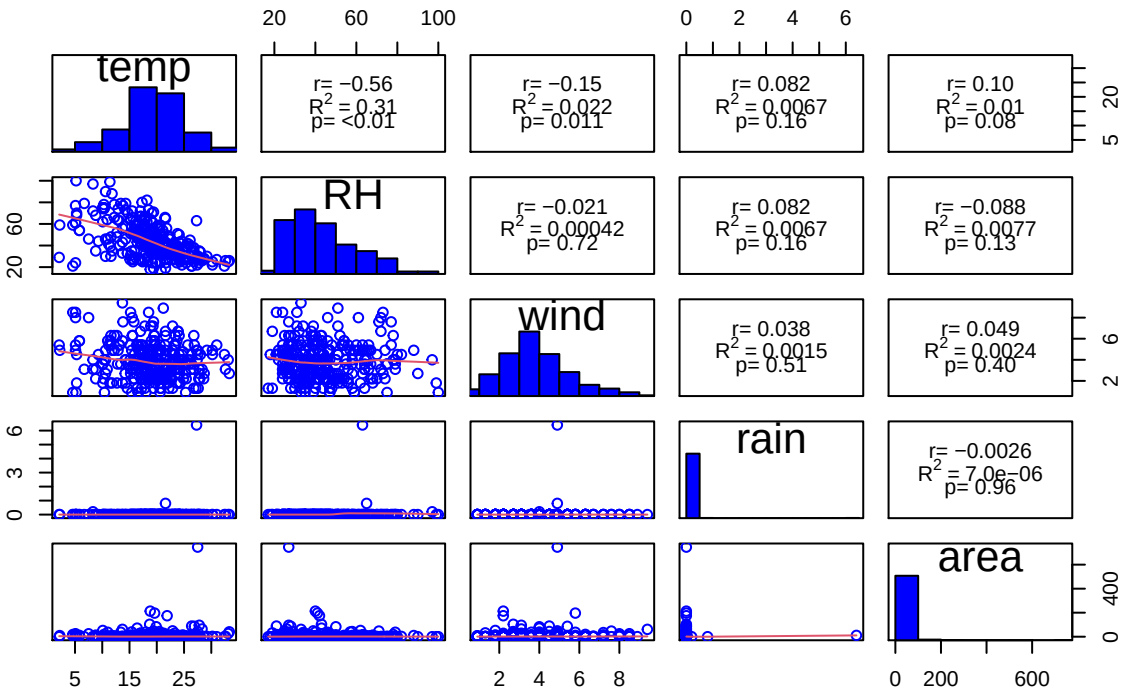
Afin d'affiner nos résultats, nous allons construire la matrice de scatter plot sur laquelle nous allons :

-afficher la distribution de chacune des variables en diagonale,

-afficher les scatter plots avec la courbe de tendance en bas de la diagonale

-afficher en haut de la diagonale le résultat du test sur le coefficient de Pearson de la corrélation (`r`), le niveau de significativité (`p-value`) et l'indice de corrélation multiple (R^2).

```
# matrice scatterplot et test
pairs(fr, pch = 1, lower.panel=panel.smooth, upper.panel=panel.cor,diag.panel=panel.hist, col = "blue",
```



- Analyse :

Cette matrice confirme les différentes corrélations évoquées plus haut. Cela se laisse voir au travers de la forme allongée de certains nuages de points.

a) Les variables `temp` (température) et `RH` (humidité relative) sont négativement corrélées (r = -0.56) avec une très forte significativité (p = 0.01). Cela voudrait dire que quand la température baisse, l'humidité relative augmente et diminue quand la température s'élève.

b) Les variables `temp` (température) et `wind` (vitesse du vent) sont négativement corrélées (r = -0.15) avec une forte significativité (p = 0.011). En effet la température s'abaisse ("ressentie") quand la vitesse du vent augmente et inversement.

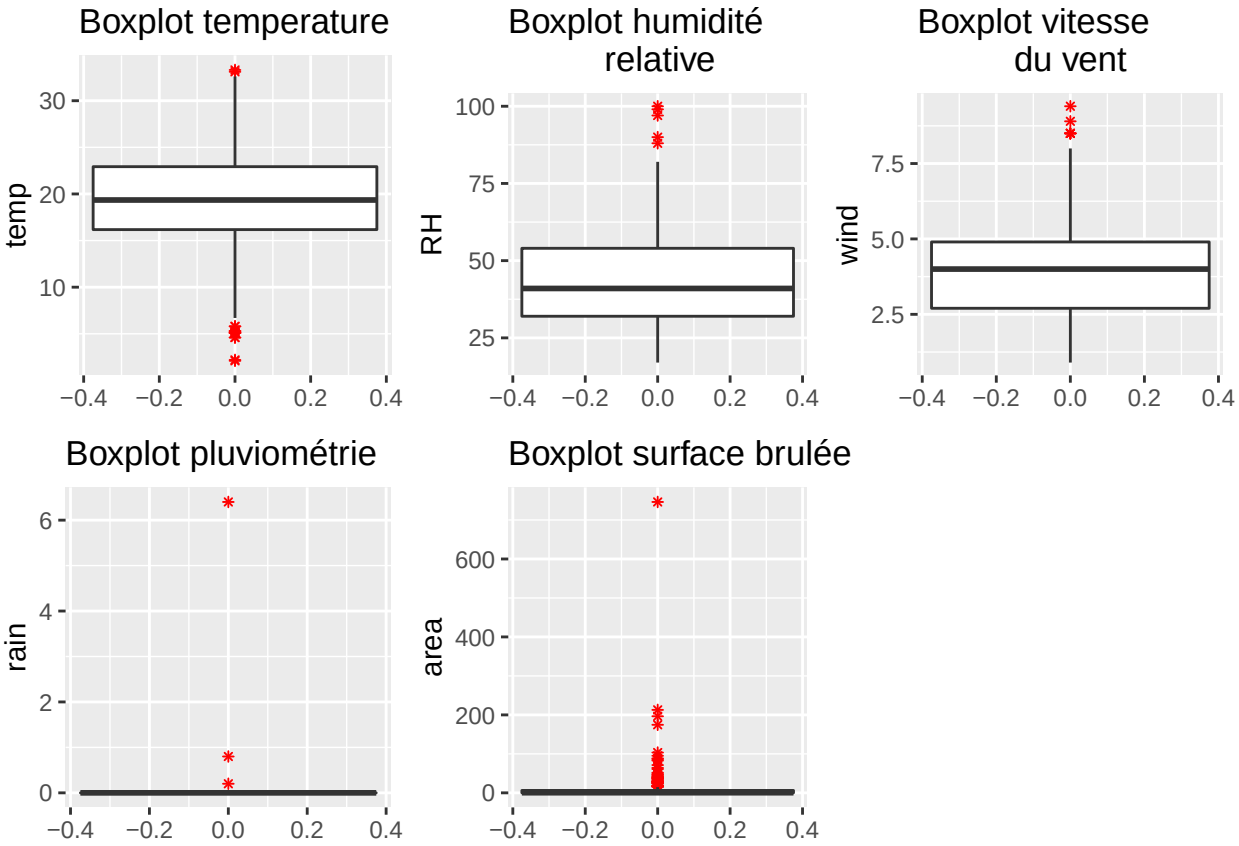
c) Les variables `temp` (température) et `area` (surface de foret brûlée) sont positivement corrélées (r = 0.10) avec une faible significativité (p = 0.10). Cela voudrait dire que les feux de cette forêt augmentent (surface) lors des hausses de températures et diminue quand la température diminue.

d) Les variables `temp` (température) et `rain` (pluviométrie) sont positivement corrélées (r = 0.082) avec une absence de significativité (p = 0.16). Ces deux variables semblent ne pas être corrélées.

e) Les variables `area` (surface de foret brûlée) et `RH` (humidité relative) sont négativement corrélées (r = 0.088) avec une absence de significativité (p = 0.13). Ces deux variables semblent ne pas être corrélées.

f) Les variables `RH` (humidité relative) et `rain` (pluviométrie) sont positivement corrélées (r = 0.082) avec une absence de significativité (p = 0.16). semblent ne pas être corrélées.

On remarque également la présence de valeurs aberrantes pour toutes les variables. C'est à dire ces valeurs qui s'écartent fortement de la moyenne. Pour mieux les voir illustrons les par le biais de boîtes à moustaches.



Les différents graphes (boxplot) illustrent encore mieux les point aberrants que nous avons représentés en rouge.

VOIR LA NOUVELLE DATA FRAME NETTOYEE DANS LE DOSSIER

```
save("data", file="New_Data_F.rda")
```

Exercice 2

1) Chargement des données et affichage

```
setwd("~/Bureau/Statistique/DM_statistique")
datat <- read.csv2("titanic.csv")
```

Visualisation des données

```
str(datat)
```

2) Résumé de l'ensemble des données et construction des graphes

```
summary(datat)
```

- On voit tout de suite dans le résumé que la variable âge (`age`) est considérée par R comme une class "Character" ce qui ne l'ai pas en vrai. Cela pourrait entraver nos calculs ou analyses. Pour ce faire, nous allons la convertir en "Numeric".

```
datat$age <-  as.numeric(datat$age)
```

Représentations graphiques

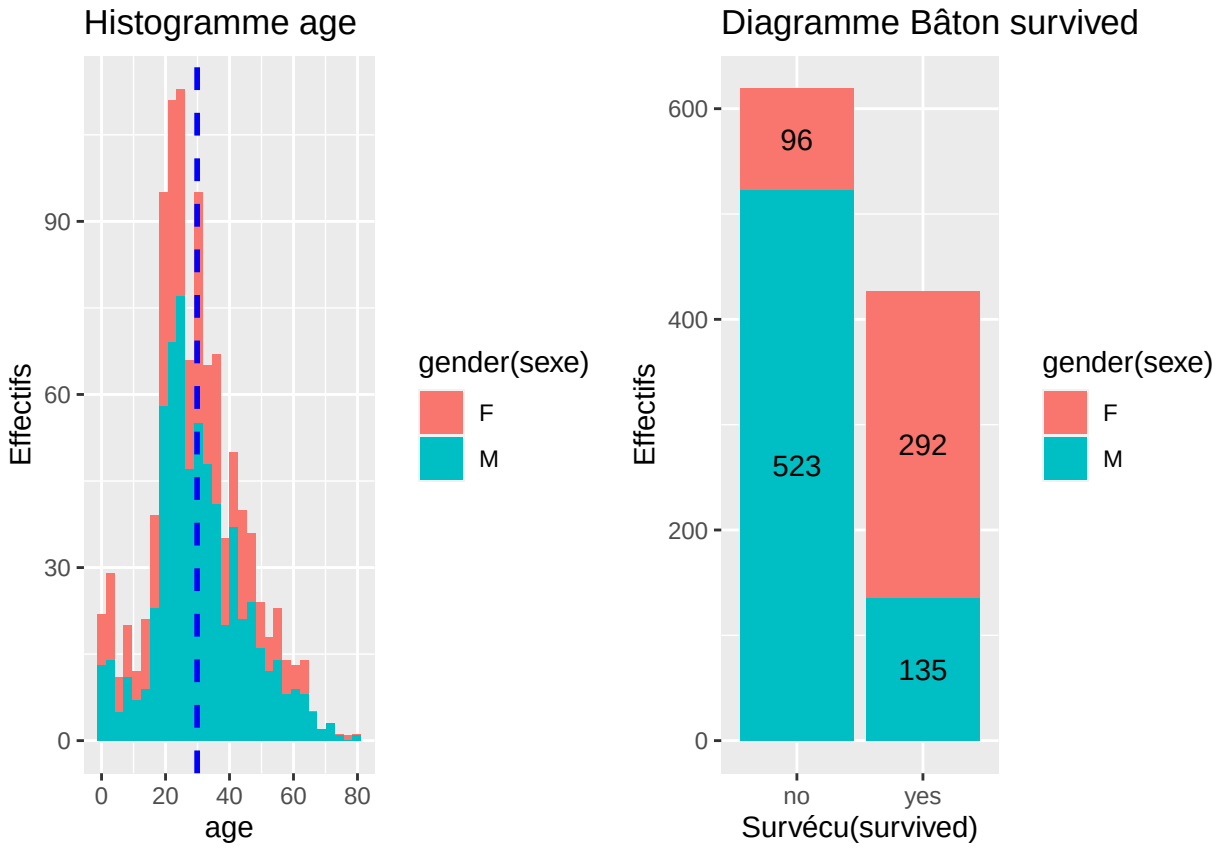
- Représentation de la variable `age` et de la variable survécu `survived`

Nous savons que la variable `age` ici est quantitative continue ce qui nous oriente vers un histogramme.

```r
## Histogrammes des ages par groupes
a <- ggplot(datat, aes(x=age, fill=gender)) +
  geom_histogram() +
  ## Ajouter le trait de la moyenne
  geom_vline(aes(xintercept=mean(age)), color="blue", linetype="dashed", size=1) +
  ylab("Effectifs") +
  labs(fill = "gender(sexe)") +
  ggtitle("Histogramme age")

b <- ggplot(data = datat) +
  aes(x = survived, fill = gender, position = "dodge") +
  geom_bar() +
  geom_text(aes(label = after_stat(count)), stat = "count", position = position_stack(.5)) +
  ggtitle("Diagramme Bâton survived") +
  xlab("Survécu(survived)") +
  ylab("Effectifs") +
  labs(fill = "gender(sexe)")

plot_grid(a, b, ncol = 2, nrow = 1)
```



Sur l'histogramme on observe que l'effectif est le plus élevé pour les passagers ayant un age compris entre 19 ans et 40 ans. Plus l'âge augment ou diminue l'effectif diminue et ce pour les hommes ainsi que le femmes.

Sur le diagramme en bâton on observe que parmi les passagers 135 ont survécus contre 523 qui n'ont pas survécus et 292 femmes qui ont survécus contre 96 qui n'ont pas survécus. Les femmes ont un effectif élevé de survivant contrairement aux hommes qui ont un effectif de morts élevés

- Représentation de la variable classe du passager `pclass`

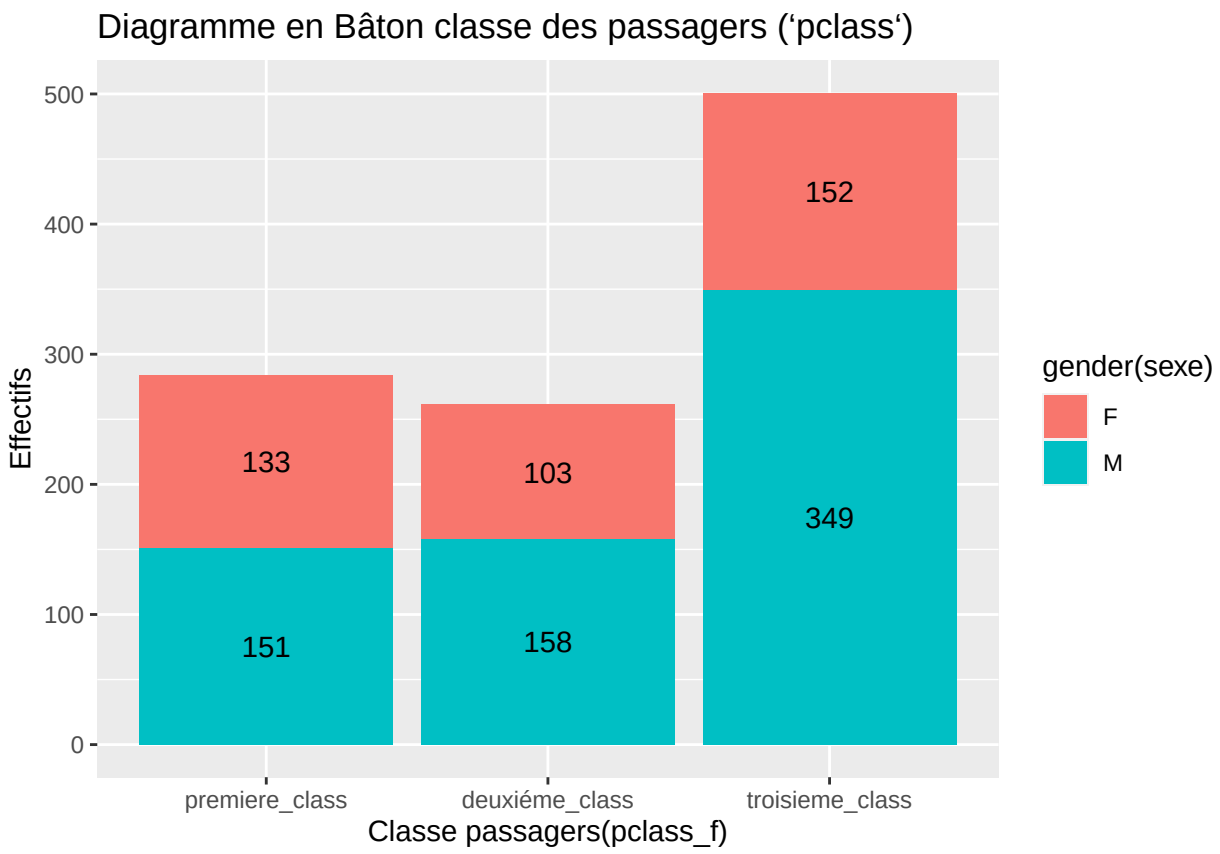
La variable `pclass` est considérée ici comme une variable “numeric”. Pour faciliter nos analyses, nous préférons qu’elle soit traitée comme une variable qualitative. Pour cela nous allons utiliser la commande “factor”.

```
## Associer des niveaux
head(factor(datat$pclass))
```

```
## Modification
datat$pclass_f <- factor(datat$pclass,
levels = c(1, 2, 3), labels = c("premiere_class","deuxiéme_class","troisieme_class"))
```

Maintenant nous pouvons utiliser un diagramme en bâton (“barplot”) pour la représenter au mieux.

```
## Tracé du diagramme en bâton
ggplot(data = datat) +
  aes(x = pclass_f, fill = gender, position = "dodge") +
  geom_bar() +
  geom_text(aes(label = after_stat(count)), stat = "count", position = position_stack(.5))+
  ggtitle("Diagramme en Bâton classe des passagers (`pclass`)") +
  xlab("Classe passagers(pclass_f)") +
  ylab("Effectifs") +
  labs(fill = "gender(sexe)")
```



Le diagramme en bâton nous montre qu’en première classe on a un effectif de 151 hommes et 133 femmes, en seconde classe 158 hommes et 103 femmes, et en troisième classe 349 hommes contre 152 femmes.

Étudions le lien entre les variables deux à deux :

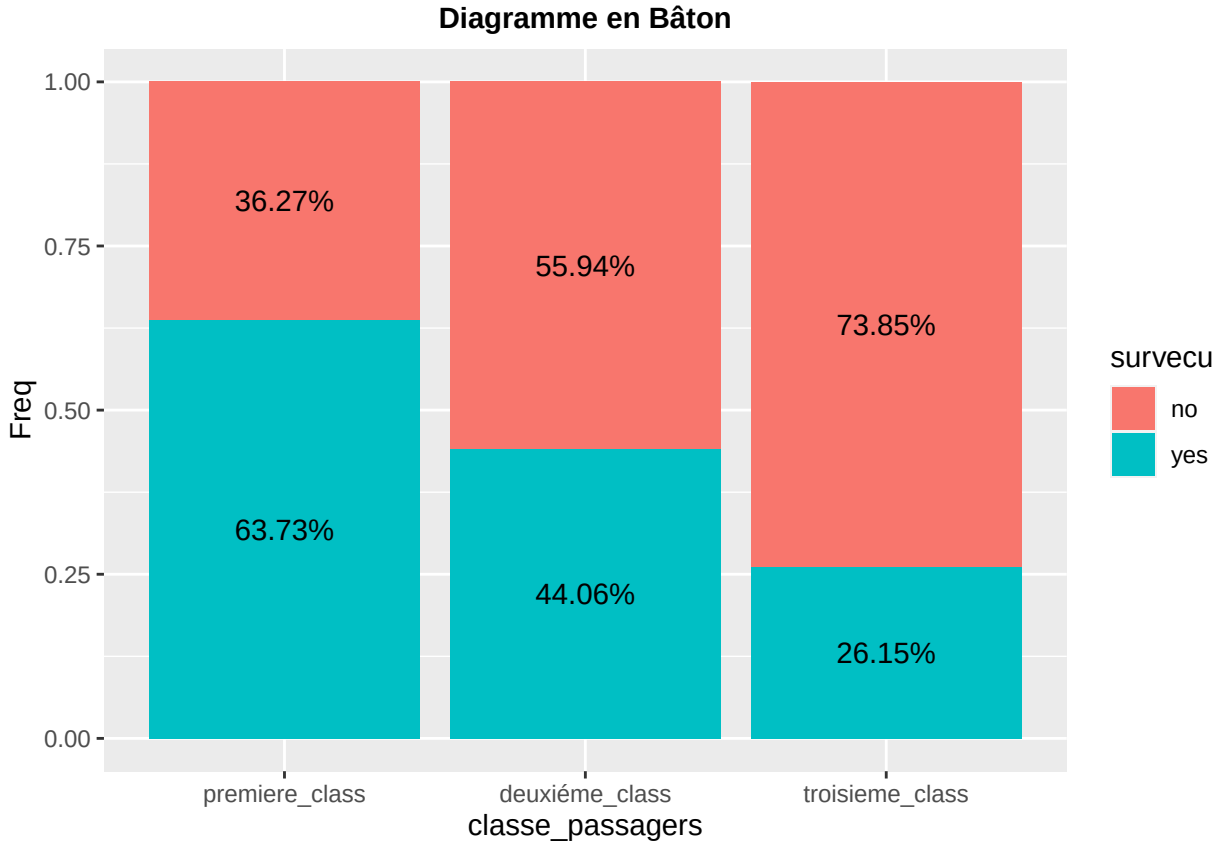
Pour ce faire nous commencerons par observer leur comportement au travers d’un diagramme en bâton (barplot) et d’une boîte à moustache. Et nous terminerons par la réalisation du test de khi-deux (test de pourcentage) et celui de Student afin de savoir si il y a correspondance entre la théorie et les repartirions

observées.

4) Étudions le lien entre la classe de passagers et leur statut de survie

```
classe_passagers <- datat$pclass_f
survecu <- datat$survived
t1 <- table(classe_passagers, survecu)
percent1 <- proportions(t1, "classe_passagers")
df <- data.frame(percent1)
df$labelpos <- ifelse(df$survecu=="yes", df$Freq/2,1 - df$Freq/2)

p <- ggplot(data = df) +
  geom_bar(stat="identity", mapping = aes(x=classe_passagers, y=Freq, fill=survecu)) +
  geom_text(aes(x=classe_passagers, y=labelpos, label=paste0(round(Freq*100,2),"%"))) +
  ggtitle("Diagramme en Bâton")
p + theme (plot.title = element_text(size=11,face="bold",hjust = 0.5))
```



A l'issu de ce tracé, on remarque que : en première classe 63.73% des passagers ont survécus contre 36,.27% qui n'ont pas survécus. En deuxième classe 44.06% ont survécus contre 55.94%. Et en troisième classe 26.15% ont survécus contre 73.85% qui n'ont pas survécus.

On peut voir que les passagers de première classe qui ont survécus sont plus nombreux que ceux de deuxième classe qui sont a leur tour plus nombreux que ceux de la troisième classe.

Les deux variables (`pclass` et `survived`) semblent être liées c'est à dire plus la classe du passager est élevée, plus ils a de la chance de survivre. Vérifions cela à travers un test de khi-deux.

```
# Recodage de pclass en variable binaire
datat$pclass_b <- ifelse(datat$pclass>2,1,0)
# Test khi-deux
```

```r
chisq.test(datat$pclass_b, datat$survived, correct = FALSE)
```

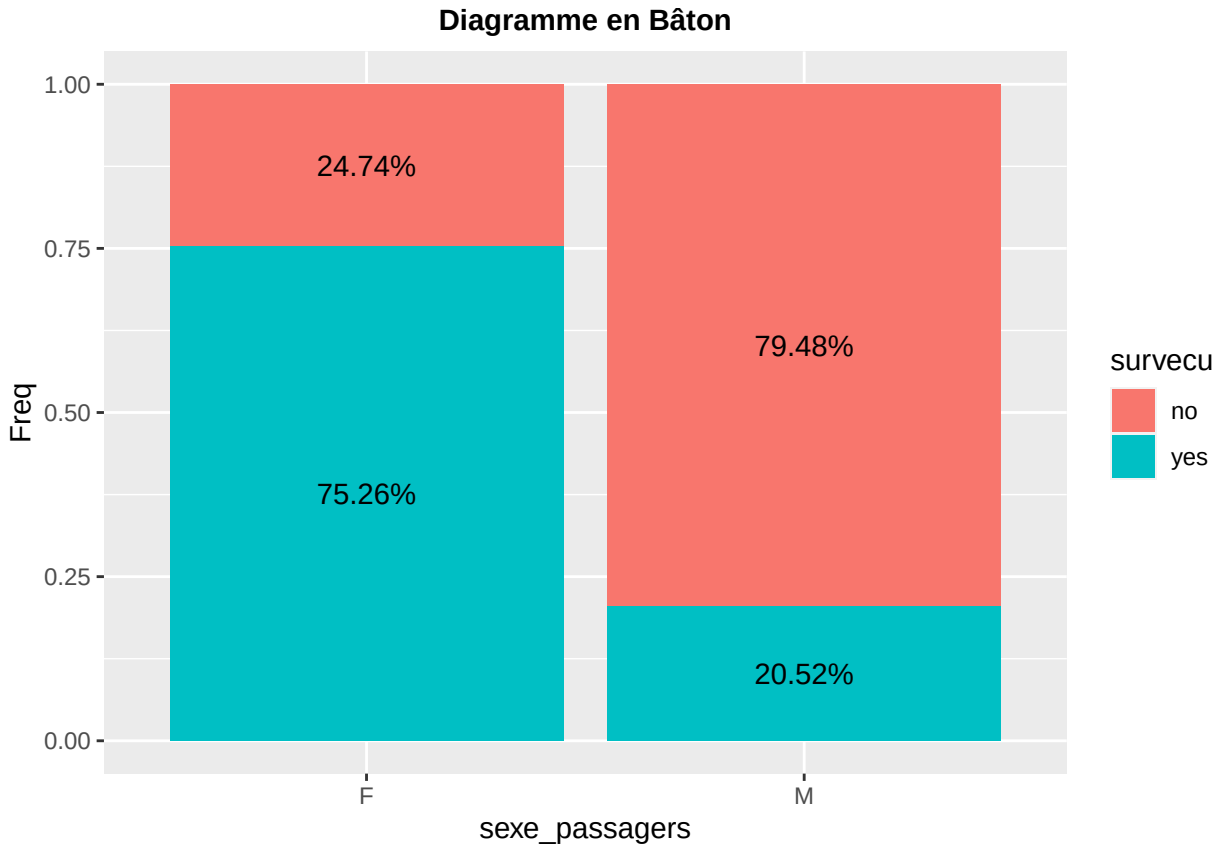
```
##
##  Pearson's Chi-squared test
##
## data:  datat$pclass_b and datat$survived
## X-squared = 85.712, df = 1, p-value < 2.2e-16
```

Ce résultat nous montre que la p-value = 2.2e-16 est très largement inférieur à 0.05. On peut affirmer avec un haut niveau de confiant que le hasard à lui seul ne pourrait pas expliquer le lien entre ces deux variables. Il y a donc un lien statistique entre La classe des passagers (`pclass`) et leur statut de survie (`survived`).

5) Étudions le lien entre le sexe des passagers et leur statut de survie

```r
sexe_passagers <- datat$gender
t1 <- table(sexe_passagers, survecu)
percent1 <- proportions(t1, "sexe_passagers")
df2 <- data.frame(percent1)
df2$labelpos <- ifelse(df2$survecu=="yes", df2$Freq/2,1 - df2$Freq/2)

p <- ggplot(data = df2) +
  geom_bar(stat="identity", mapping = aes(x=sexe_passagers, y=Freq, fill=survecu)) +
  geom_text(aes(x=sexe_passagers, y=labelpos, label=paste0(round(Freq*100,2),"%"))) +
  ggtitle("Diagramme en Bâton")
p + theme (plot.title = element_text(size=11,face="bold",hjust = 0.5))
```



On observe que 75.26 % des femmes ont survécus contre 24.74% qui n'ont pas survécus pendant que 20.52 % des hommes ont survécus contre 79.48 % qui n'ont pas survécus. On remarque donc que « être femme » donne plus de chance de « survie » . Il semble avoir un lien entre ces deux variables. Appliquons le test de

khi-deux :

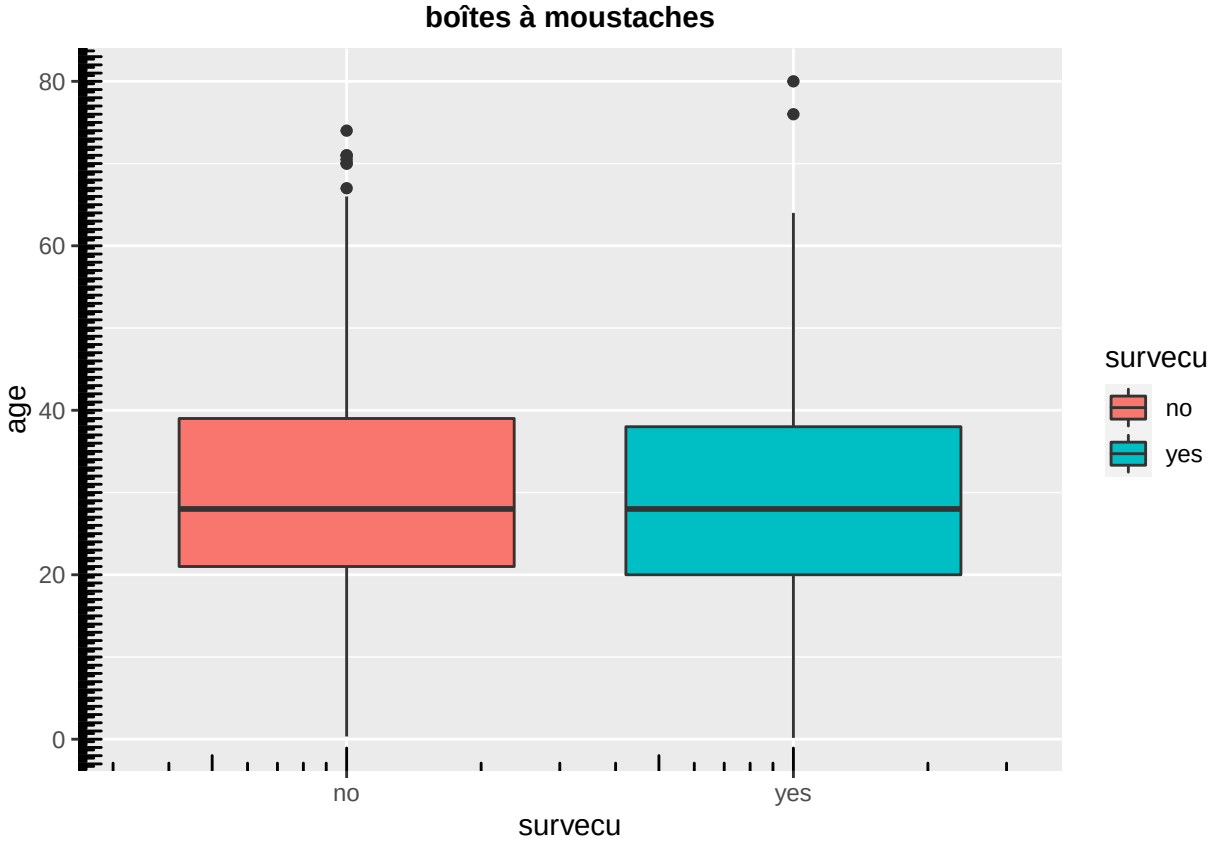
```
# Test khi-deux
chisq.test(datat$gender, datat$survived, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  datat$gender and datat$survived
## X-squared = 302.76, df = 1, p-value < 2.2e-16
```

On obtiens une p-value = 2.2e-16 qui est très largement inférieur à 0.05. On peut affirmer avec un haut niveau de confiant que le hasard à lui seul ne pourrait pas expliquer le lien entre ces deux variables. Il y a donc un lien statistique entre Le sexe des passagers (`gender`) et leur statut de survie (`survived`).

6) Étudions le lien entre l'âge des passagers et leur statut de survie

```
p<-ggplot(datat, aes(x= survecu, y=age, fill=survecu)) +
  geom_boxplot()+
  ggtitle("boîtes à moustaches")
p + theme (plot.title = element_text(size=11,face="bold",hjust = 0.5))+
  annotation_logticks()
```



Cette représentation nous montre qu'il y a eu approximativement les mêmes proportions du statut « survécu » ou pas pour les différentes tranches d'âges. On a quasiment 50 % de ceux qui ont survécus et ceux qui n'ont pas survécus entre 19 ans et 40 ans, 25 % entre 0 à 19 ans et 25 % entre 40 à 80 ans. Vérifions si il existe un lien entre ces deux variables.

Application du test de Student : Pour cela nous allons procéder à deux vérifications.

a. La distribution de la variable age est-elle proche de celle de la loi normal ?

En effet sa distribution est proche de celle de la loi normal (voir le graphe plus haut).

b. Les variances sont t-elles égales ou proches ?

```
# Ecart type
by(datat$age, datat$survived, sd, na.rm=TRUE)
```

```
## datat$survived: no
## [1] 13.92254
## ------------------------------------------------------------
## datat$survived: yes
## [1] 15.06148
```

Ici les variances sont pas égaux mais elle ne sont pas très éloignées, on peut donc utiliser le test de student mais en ne renseignant pas "var.equal = TRUE" qui est utilisé dans le cas où les deux variances sont égaux.

```
# Test de Student
t.test(datat$age~datat$survived)
```

```
##
##  Welch Two Sample t-test
##
## data:  datat$age by datat$survived
## t = 1.7707, df = 868.26, p-value = 0.07696
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.176415  3.430696
## sample estimates:
##  mean in group no mean in group yes
##          30.54537          28.91823
```

Le calcul dévoile une p-value = 0.07696. Ce résultat est supérieur à 0.05. Il n'y a donc pas de différence significative d'âge entre les passagers qui ont survécus et ceux qui n'ont pas survécus. les variables `age` et `survived` ne sont donc pas statistiquement liées.