TP 1

1. Manipulation de tableaux de données avec R

On va travailler sur un objet important en statistiques, les tableaux de données qui sont des objets de type `data.frame`. Pour cela, on va charger un jeu de données relatif à des mesures de qualité de l'air. On effectue la commande suivante : `data(airquality)`.

1. Décrire succintement le jeux de données à l'aide de la commande `help` ou `?`.
2. que renvoient les commandes suivantes
   (a) `head(airquality)`
   (b) `str(airquality)`
   (c) `summary(airquality)`
   (d) `colnames(airquality)`
   (e) `dim(airquality)`
3. Effectuer les commandes suivantes. Que constate-t-on ?
```
airquality$Ozone
Ozone
attach(airquality)
Ozone
detach(airquality)
Ozone
```
4. Il arrive fréquemment qu'un jeu de données contiennent des données *manquantes* (non renseignées). Une première approche que l'on va considérer est de supprimer les lignes du tableaux contenant des données manquantes. Que renvoie la commande suivantes
   (a) `sum(is.na(airquality))`
   (b) `dataTemp <- na.omit(airquality)`
   (c) `sum(is.na(dataTemp))`
5. Désormais, on travaille avec le jeu de données `dataTemp`. On va dans un premier temps donner quelques descriptifs de la variable `Ozone`.
   (a) De quel type est la variable `Ozone`
   (b) En utilisant les fonctions `mean, median, var et sd`, donner la moyenne, la médiane et la variance de la variable `Ozone`.
   (c) Commenter les résultats obtenus.
   (d) À l'aide de la fonction `quantile`, donner la valeur du premier et troisième quartile.
   (e) En déduire la taille de l'intervalle interquartile.
   (f) Donner l'indice de l'observation ayant la plus grande valeur d'Ozone.
   (g) Quel est le pourcentage d'observations ayant une valeur d'Ozone supérieur où égale à 80 ?
6. Une fois donné les descriptifs numériques relatif à la variable `Ozone`, on va proposer un descriptif visuel de cette variable. Pour cela, on considère la commande suivante `boxplot(Ozone)`. On peut constituer que deux points sont situés en dehors des moustaches. À l'aide de la commande `boxplot$out`, déterminer l'indice des observations correspondant à ces valeurs.
7. Pour obtenir le même graphique avec la fonctions `ggplot`, on effectue les commandes suivantes
```
install.packages("ggplot2")
require(ggplot2)
fig <- ggplot(data.frame(Ozone))+aes(x = "Ozone", y = Ozone) +
       geom_boxplot(color = "blue" , fill = "red")+
       labs(title = "boxplot pour la variable Ozone", x = "", y = "valeurs")
fig
```

8. Dans cette questions, on va s'intéresser à la variable `Month`.
   (a) Quel est le type de cette variable.
   (b) Comme premier descriptif numérique de cette variable, on peut donner le tableau de contingence avec la commande `table(Month)`. Quel est le pourcentage de mesure effectué les mois de juin et juillet ?
   (c) On peut donner comme descriptif visuel de cette variable un diagramme en barre (`barplot(table(Month))`) ou un camemebert (`pie(table(Month))`). Construire un diagramme en barre pour cette variable à l'aide de la fonction `ggplot`.
   (d) Construire une vecteur de même taille que `Month`, contenant le nom du mois correspondant à la valeur de `Month`.
9. Dans cette question, on va proposer descriptif simultané des variables `Ozone` et `Month`.
   (a) Donner la moyenne et l'écart-type de la variable variable Ozone par mois.
   (b) Parmis les mesure effectué au mois de juillet, quel est le pourcentage d'observations dont la valeur Ozone dépasse 50. Même question pour les observations du mois de septembre. Qu'en conclure ?
   (c) On peut considérer le descriptif visuel suivant
```
df <- data.frame(Ozone = Ozone, Month = as.factor(Month))
fig <- ggplot(df) + aes(x = Month, y = Ozone, fill = Month) + geom_boxplot()
       + labs (x= "Month", y = "Ozone", fill = "Mois")
fig
```
   (d) Que renvoie la commande suivante ?
```
tapply(X= df$Ozone, INDEX = df$Month, function(x)length(boxplot(x)$out))
```

2. Évaluation du risque de mauvaise classification

Soit (X, Y) un couple de variable aléatoire tel que X suit une loi uniforme sur $[0, 1]$ et $Y|X$ suit une loi de Bernoulli de paramètre $\eta(X)$ avec

$$\eta(X) = \frac{1}{5}\mathbb{1}_{\{X\leq 1/4\}} + \frac{2}{5}\mathbb{1}_{\{1/4<X\leq 1/2\}} + \frac{3}{5}\mathbb{1}_{\{1/2<X\leq 3/4\}} + \frac{4}{5}\mathbb{1}_{\{3/4<X\}}.$$

1. On considère le code `R` suivant :
```
x <- runif(1,0,1)
eta <- 1/5*(x <= 1/4) + 2/5*(x > 1/4   &   x <= 1/2)
      + 3/5*(x > 1/2   &   x <= 3/4) + 4/5*(x > 3/4)
y <- rbinom(1,1,eta)
pred <- ...
error <- ...
```
   (a) Que renvoie la variable `y`
   (b) Compléter le code pour que la variable `pred` renvoie la valeur de $s^*(x)$ où s^* est le classifieur Bayes.
   (c) Compléter le code pour que la variable `error` renvoie la valeur de $\mathbb{1}_{\{s^*(x)\neq y\}}$
2. On considère $(x_1, y_1), \ldots, (x_n, y_n)$, n réalisations de la variable (X, Y).
   On propose une estimation de $R(s^*)$ par $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{s^*(x_i)\neq y_i\}}$.
   Justifier.
3. Compléter la fonction ci-dessous prenant en argument un entier n et renvoyant une estimation de $R(s^*)$.

```
risk <- function(n){
x <- runif(n,0,1)
...
out <- mean()
return(out)
}
```
4. Pour différentes valeurs de n, comparer les valeurs estimées obtenues à $R(s^*)$.
5. On considère le code `R` ci-dessous. que renvoie `mean(risque)` et `sd(risque)`
```
risque <- sapply(1:100, function(int){risk(1000)})
mean(risque)
sd(risque)
```
6. Comparer l'estimation de $R(s^*)$ pour $n = 10$ et $n = 1000$.