TP2 Correction: Statistiques Descriptives

Sonia Tieo

19/10/2020

1.Chargement de données

```
data(mtcars)
```

2.Infos sur le jeu de données

Q: Décrivez le jeu de données (nombre de ligne, colonne). A quoi correspondent les observations et les variables ?

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

3.Transformation de données

Q: A quoi correspond la variable “cyl” ? Convertir les éléments de cette colonne en type “factor”. Commentez.

```
#L a variable cyl représen  te le nombre de cylindres avec trois modalités possibles (4,6 ou 8),
mtcars$cyl<-as.factor(mtcars$cyl)
```

4.Analyse univariée - Variable quantitative

Q: Pour la variable ’mpg", calculer:
1. min, max, l’intervalle des valeurs
2. la moyenne, la variance et l’écart type
3. la médiane
4. les quantiles
5. le nombre d’observations

```
mpg <- mtcars$mpg
min(mpg)## [1] 10.4
max(mpg)## [1] 33.9
range(mpg)## [1] 10.4 33.9
mean(mpg)## [1] 20.09062
```

```
var(mpg)## [1] 36.3241
sd(mpg)## [1] 6.026948
median(mpg)## [1] 19.2
quantile(mpg)##     0%    25%    50%    75%   100%## 10.400 15.425 19.200 22.800 33.900
```

Q: Tester la fonction summary sur:
- mtcars
- mpg

Commentez:

```
summary(mtcars)
```

```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt             qsec             vs               am
##  Min.   :1.513   Min.   :14.50   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :3.325   Median :17.71   Median :0.0000   Median :0.0000
##  Mean   :3.217   Mean   :17.85   Mean   :0.4375   Mean   :0.4062
##  3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5.424   Max.   :22.90   Max.   :1.0000   Max.   :1.0000
##       gear            carb
##  Min.   :3.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:2.000
##  Median :4.000   Median :2.000
##  Mean   :3.688   Mean   :2.812
##  3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :8.000
```

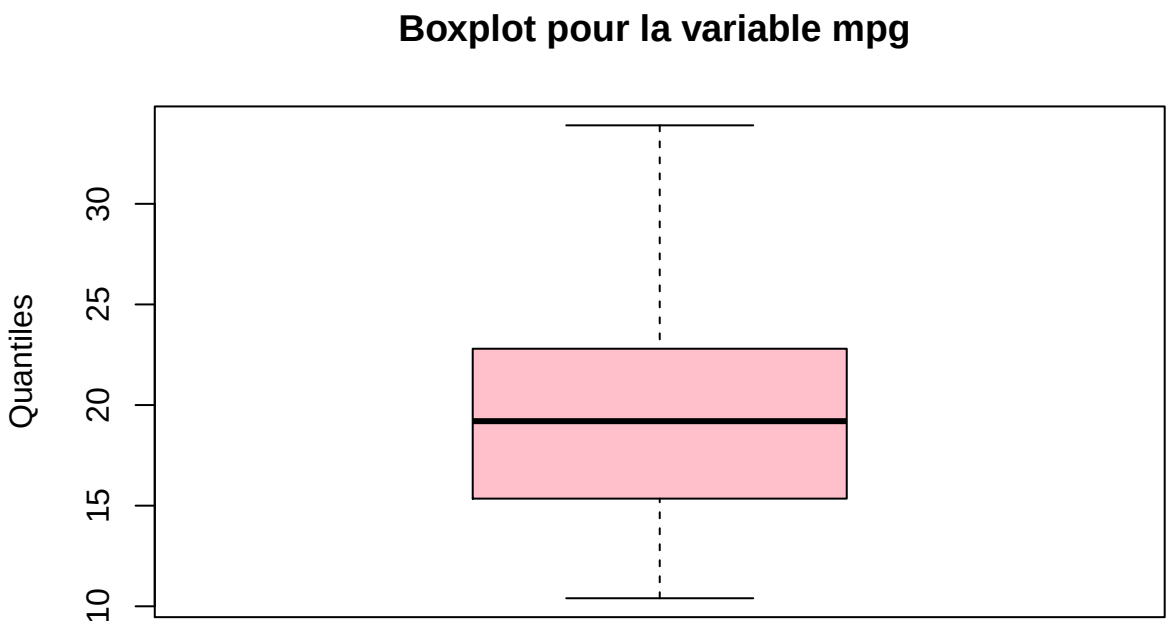
```
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

4-1.Représentation graphique

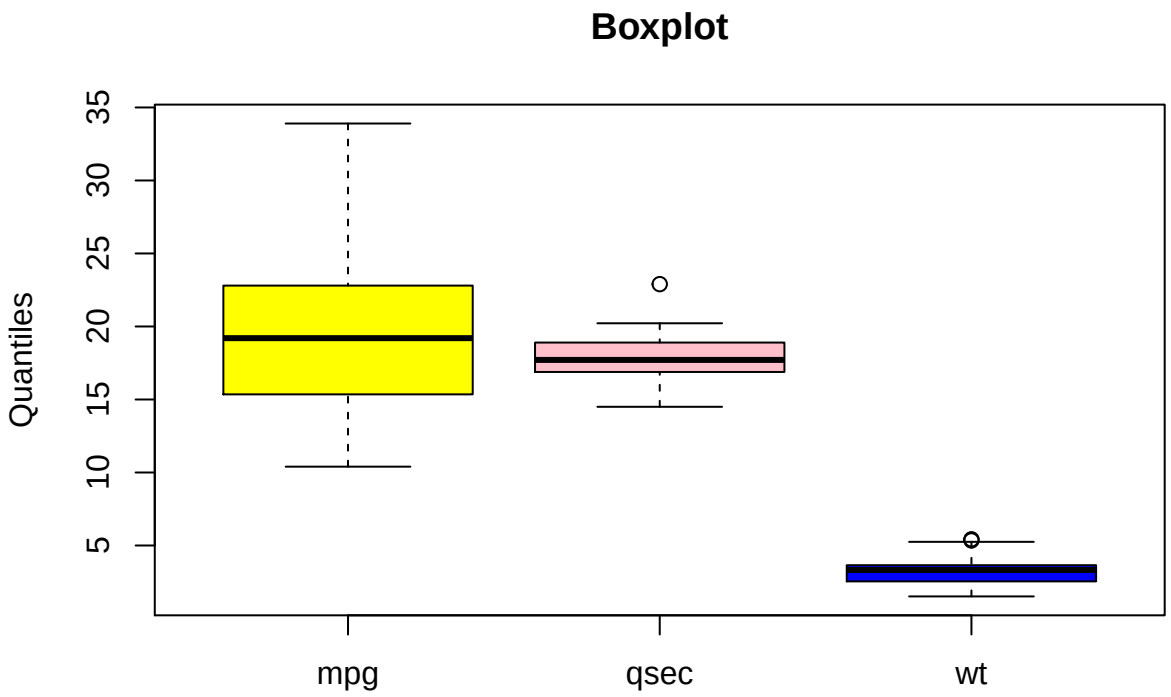
Q: Commentez:

```
boxplot(mtcars$mpg, col = c("pink"),
        main = "Boxplot pour la variable mpg ",
        ylab = "Quantiles")
```



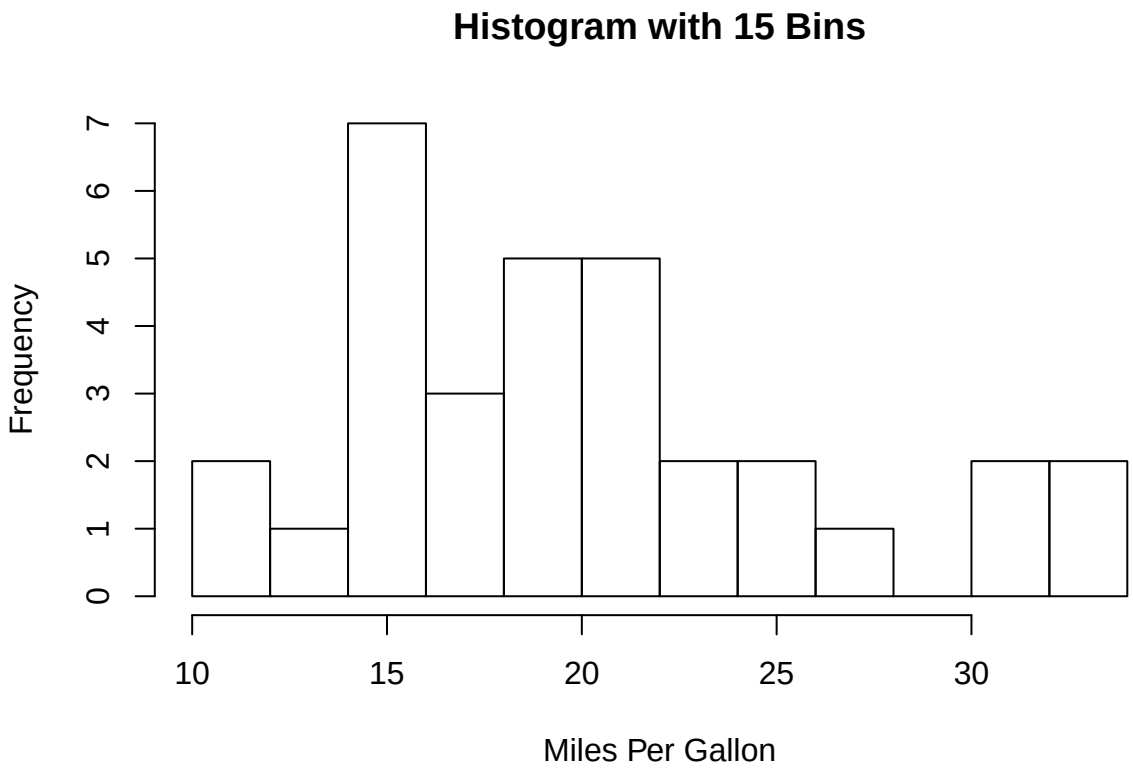
Q: Faire un boxplot pour les variables suivantes : mpg, qsec, wt
Mettre les 3 boxplots sur un même graphes, une légende, et 3 couleurs différentes(pour chaque boxplot)

```
boxplot(mtcars[,c('mpg','qsec','wt')],
        col = c("yellow" , "pink" , "blue"),                    #Pour la couleur
        main = paste("Boxplot"),      #Pour le titre
        ylab = "Quantiles")               #Pour le titre de l'axe des ordonnées
```



Q: Faire un histogramme pour la varaible mpg avec des légendes. Regardez les options de cette fonction hist, commentez:

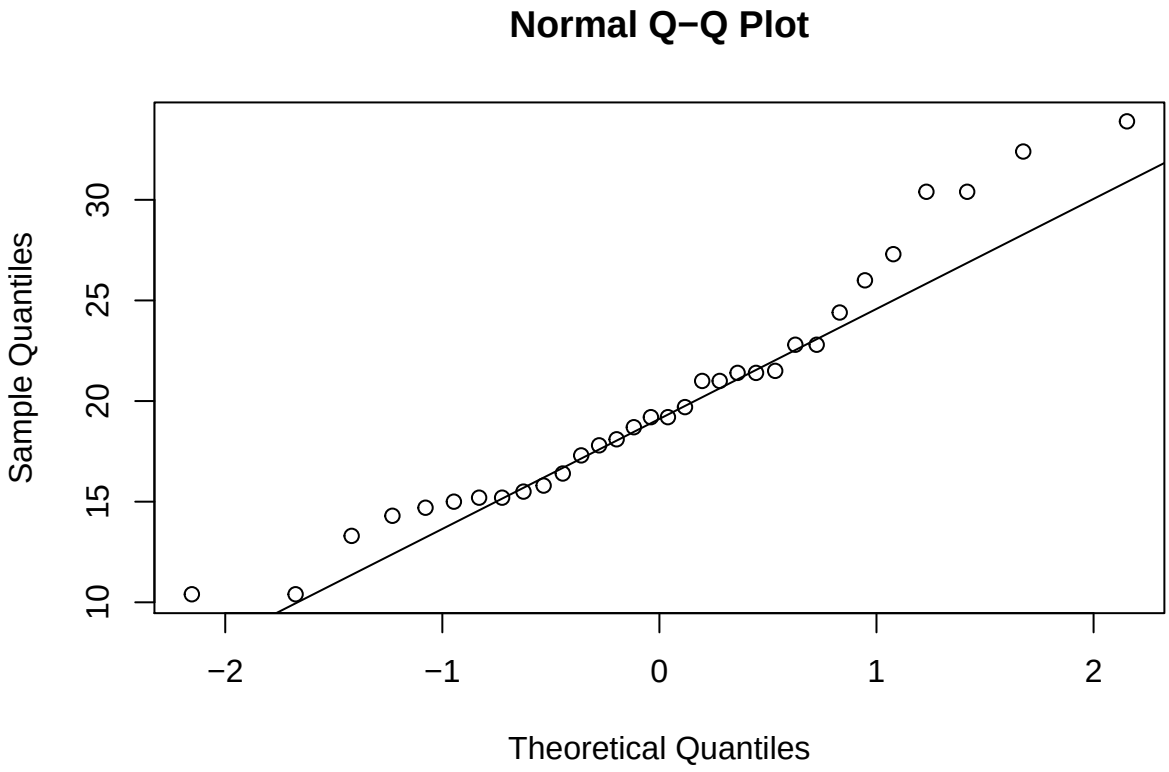
```
hist(mtcars$mpg, breaks =  15, xlab = "Miles Per Gallon",
     main = "Histogram with 15 Bins")
```



4-2. La distribution des données mpg suit-elle une loi normale ?

Q: Comparer la distribution de l'échantillon avec la distribution théorique d'une loi normale à l'aide du QQ-plot. Pour cela, executez:

```
qqnorm(mtcars$mpg)
qqline(mtcars$mpg)#la droite sample quantiles = theoretical quantiles
```




```
shapiro.test(mtcars$mpg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg
## W = 0.94756, p-value = 0.1229
```

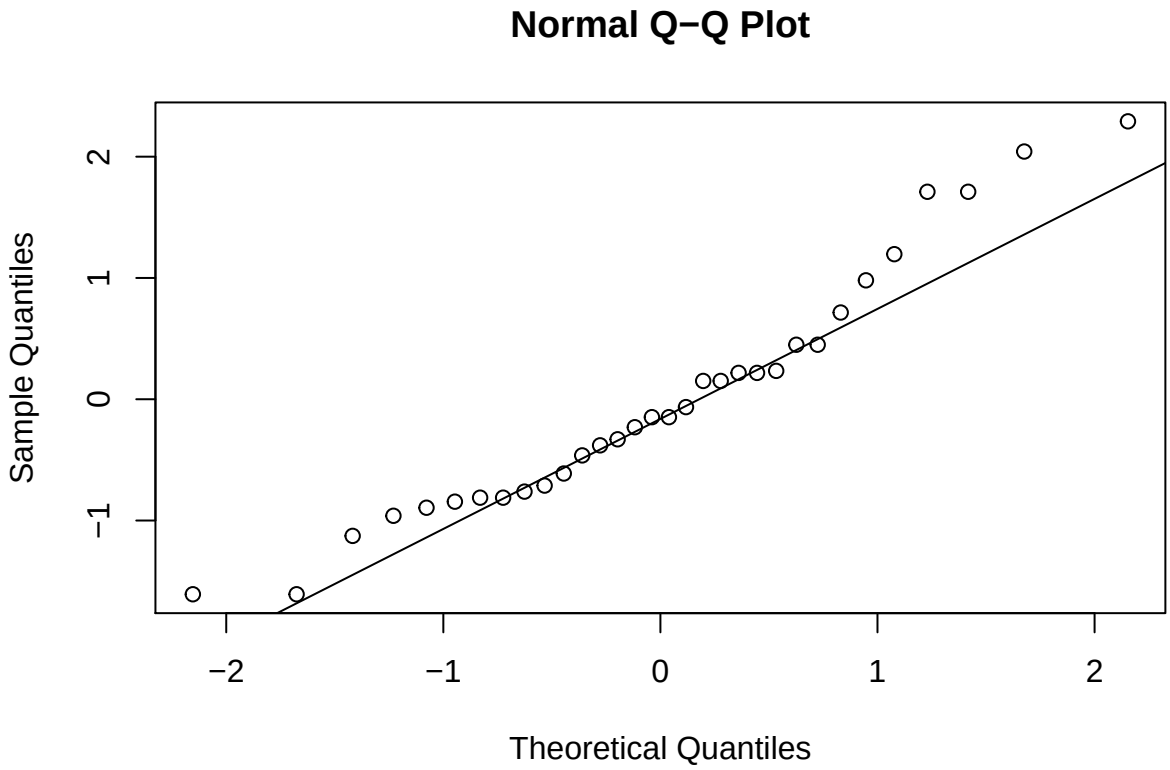
```
#Du résultat ci-dessus, la p-value > 0,05 indiquant que la distribution des données n'est pas significa
```

Q: Faites la même chose, en centrant et réduisant la variable mpg.
Indice pour réduire et centrer :

NOUV_VAR = (VARIABLE - MOYENNE(VARIABLE)) / ECARTTYPE(VARIABLE)

Conseil: stocker dans une variable "mpg.stand".
Qu'observez-vous ?

```
mpg <- mtcars$mpg
mpg.stand <- (mpg-mean(mpg))/sd(mpg)#mpg centré et réduit
qqnorm(mpg.stand)
qqline(mpg.stand)
```




```
shapiro.test(mpg.stand)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mpg.stand
## W = 0.94756, p-value = 0.1229
```

5.Analyse univariée - Variable qualitative

Q: Quelle est la différence entre les deux tableaux?

```
table.cyl <- table(mtcars$cyl)
tableprop.cyl <-  prop.table(table(mtcars$cyl))
table.cyl
```

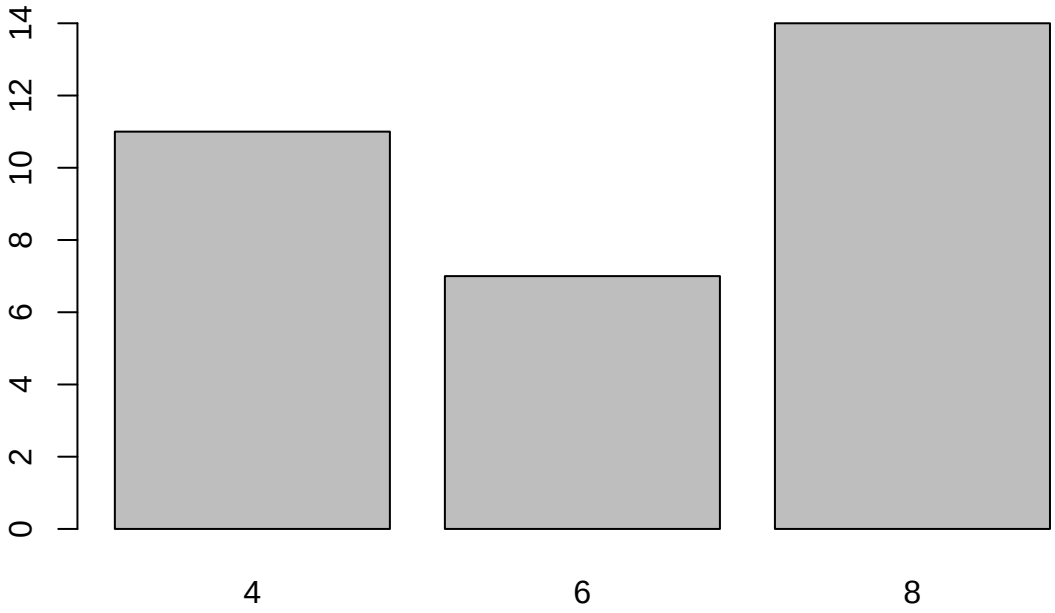
```
##
##  4  6  8
## 11  7 14
```

```
tableprop.cyl
```

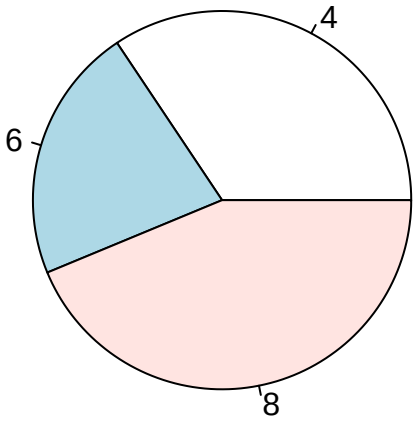
```
##
##       4       6       8
## 0.34375 0.21875 0.43750
```

Q: Executer et commenter le code suivant:
Q: Ajoutez-y des légendes au barplot et pie plot.
Q: Faire pareil avec tableprop.cyl

```r
barplot(table.cyl)
```




```r
pie(table.cyl)
```



6. Analyse bivariée - 2 Variables quantitatives

On choisit les variables : mpg (miles per gallon) et wt (poids des véhicules)

Q: Donner la covariance (cov) et la correlation de pearson et de spearman (cor) des 2 variables quantitatives:

```r
wt <- mtcars$wt
mpg <- mtcars$mpg
cov(mpg, wt)
```

```
## [1] -5.116685
```

```r
cor(mpg, wt)
```

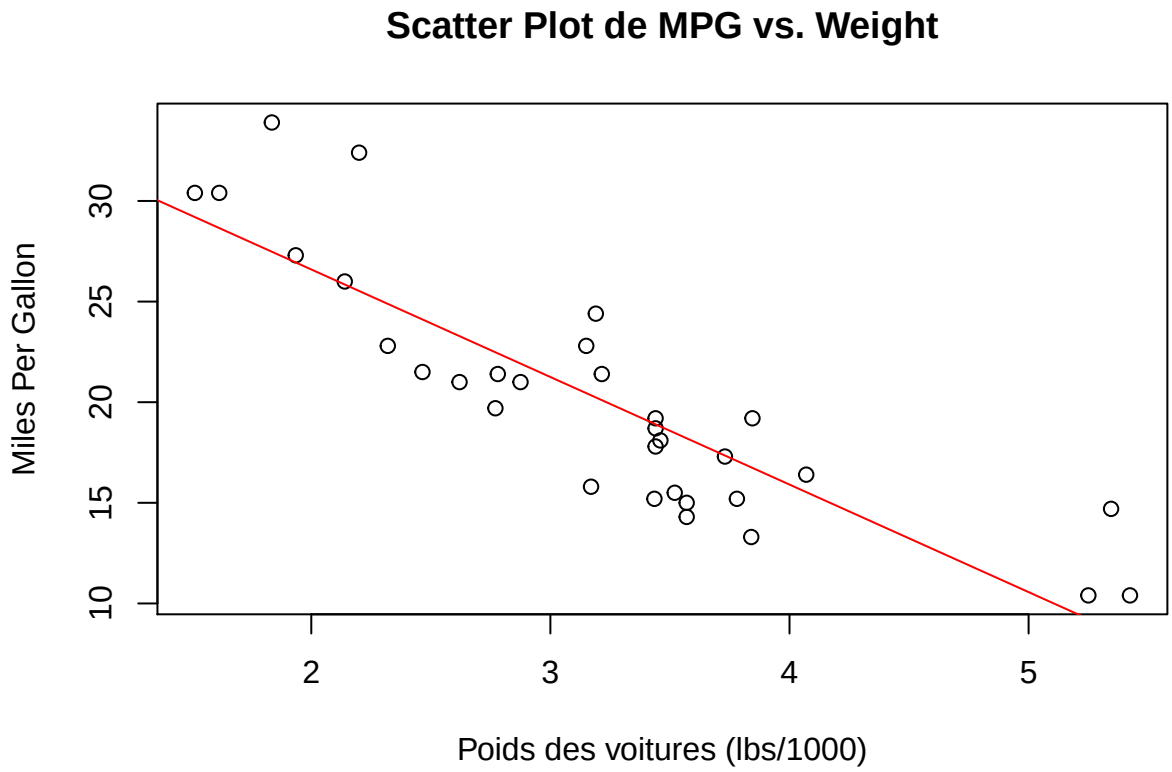
```
## [1] -0.8676594
```

```r
cor(mpg, wt, method='spearman')
```

```
## [1] -0.886422
```

Q: Représentez le nuage de plot: mpg vs weight(wt).
Indice(étape):
1. plot(x=wt, y=mpg) avec x,y = coordonnées des points *OU* plot(wt ~ mpg) où le ~ indique "wt en fonction de mpg".
2. Trouver la doite de régression avec la méthode des moindres carrés(modèle lineaire simple): lm(wt ~ mpg), commentez la commande summary(lm(wt ~ mpg)).
3. Ajouter la doite de régression au plot avec la fonction abline.

```
plot(wt, mpg, main = "Scatter Plot de MPG vs. Weight",
     xlab = "Poids des voitures (lbs/1000)", ylab = "Miles Per Gallon")
# add trendline
abline(lm(mpg~wt, data = mtcars) , col ="red")
```



Q: Faire le test de corrélation entre les variables, en supposant HO l'hypothèse de corrélation de pearson entre les variables:
Utilisez cor.test()
Commentez

```
cor.test(wt, mpg , method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  wt and mpg
## t = -9.559, df = 30, p-value = 1.294e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9338264 -0.7440872
## sample estimates:
```

```
##        cor
## -0.8676594
```

7.Analyse bivariée - Variables qualitatives

Q: Faire la table de contingence pour les variables: cyl et am avec table() **Q:** Table compte les modalités, comment obtenir les proportions ?

```
cyl <- as.factor(mtcars$cyl)
am <- as.factor(mtcars$am)
table(cyl,am)
```

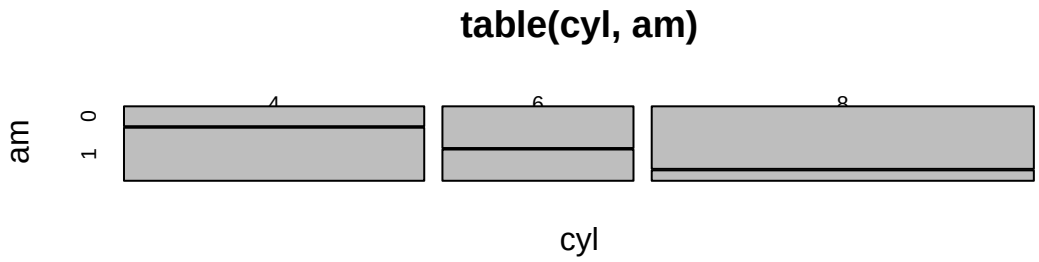
```
##    am
## cyl  0  1
##   4  3  8
##   6  4  3
##   8 12  2
```

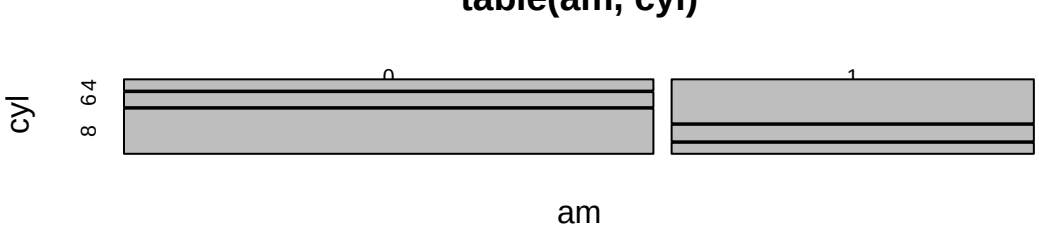
```
prop.table(table(cyl,am))
```

```
##    am
## cyl       0       1
##   4 0.09375 0.25000
##   6 0.12500 0.09375
##   8 0.37500 0.06250
```

Q: Obtenir le mosaique plot cyl vs am et am vs cyl avec mosaicplot()

```
par(mfrow=c(2,1))# pour mettre deux graphiques côte à côte
mosaicplot(table(cyl,am))# P(am | cyl=4), P(am | cyl=6), P(am | cyl=8)
mosaicplot(table(am,cyl))# P(cyl | am=0), P(cyl | am=1)
```

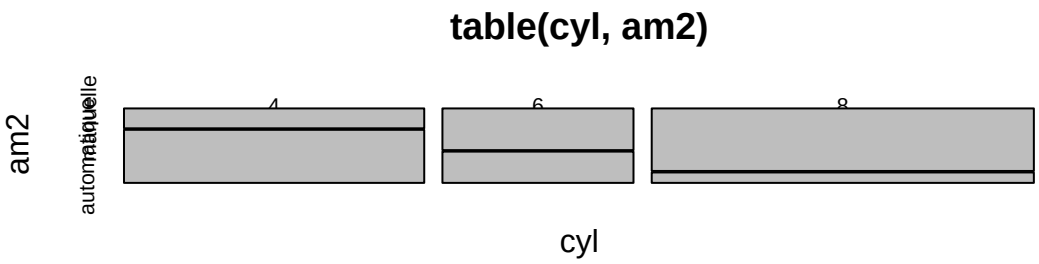


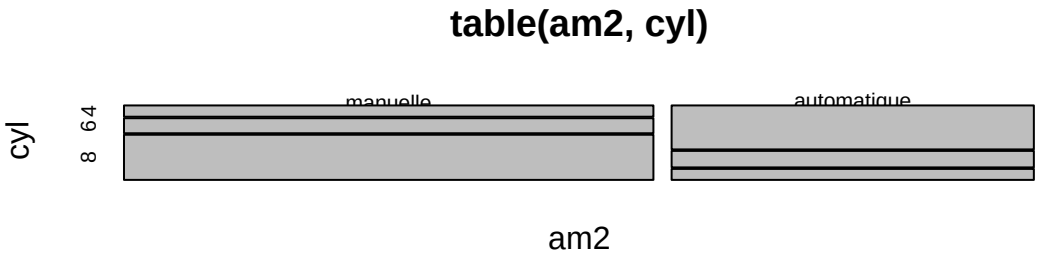


Q: Pour aller plus loin, transformer les données de la variable am:
Si 0, mettre “automatique”

Si 1, mettre “manuelle”

```
am2 <- factor(am , labels = c("manuelle" , "automatique"))
par(mfrow=c(2,1))# pour mettre deux graphiques côte à côte
mosaicplot(table(cyl,am2))# P(am | cyl=4), P(am | cyl=6), P(am | cyl=8)
mosaicplot(table(am2,cyl))# P(cyl | am=0), P(cyl | am=1)
```





8.Analyse bivariée - Variable quantitative vs variable qualitative

Q: Calculer la moyenne de “miles per gallon” en fonction des 3 catégories de volumes de cylindres en exécutant:

```
tapply(mtcars$mpg, mtcars$cyl, mean)
```

```
##        4        6        8
## 26.66364 19.74286 15.10000
```

Q: Faire de même en remplaçant la fonction mean par la fonction summary:

```
tapply(mtcars$mpg, mtcars$cyl, summary)
```

```
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.40   22.80   26.00   26.66   30.40   33.90
##
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.80   18.65   19.70   19.74   21.00   21.40
##
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.40   15.20   15.10   16.25   19.20
```

Q: Faire de même pour les variables mpg VS am:

```r
tapply(mtcars$mpg, am2, mean)
```

```
##    manuelle automatique
##    17.14737    24.39231
```

Q: Réaliser un boxplot pour les variables mpg en fonction de cyl:

```r
boxplot(mpg ~ cyl, main = "Boxplot of Miles/Gallon for Automatic/Manual",
        col = c("yellow" , "pink" ,"blue"))
```