LMAFY1101 - Solutions - Série 2

Statistiques descriptives

Exercice 1

1.

```
iris <- read.csv("iris.txt", sep = "")
```

Cette commande suppose que iris est dans votre répertoire de travail sur R.

2.

```
iris$Species <- factor(iris$Species, levels = c("versicolor",
  "virginica", "setosa"))
str(iris)
```

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "versicolor","virginica",..: 3 3 3 3 3 3 3 3 3 3
```

3.

```
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.30   Min.   :2.00   Min.   :1.00   Min.   :0.1   versicolor:50
 1st Qu.:5.10   1st Qu.:2.80   1st Qu.:1.60   1st Qu.:0.3   virginica :50
 Median :5.80   Median :3.00   Median :4.35   Median :1.3   setosa    :50
 Mean   :5.84   Mean   :3.06   Mean   :3.76   Mean   :1.2
 3rd Qu.:6.40   3rd Qu.:3.30   3rd Qu.:5.10   3rd Qu.:1.8
 Max.   :7.90   Max.   :4.40   Max.   :6.90   Max.   :2.5
```

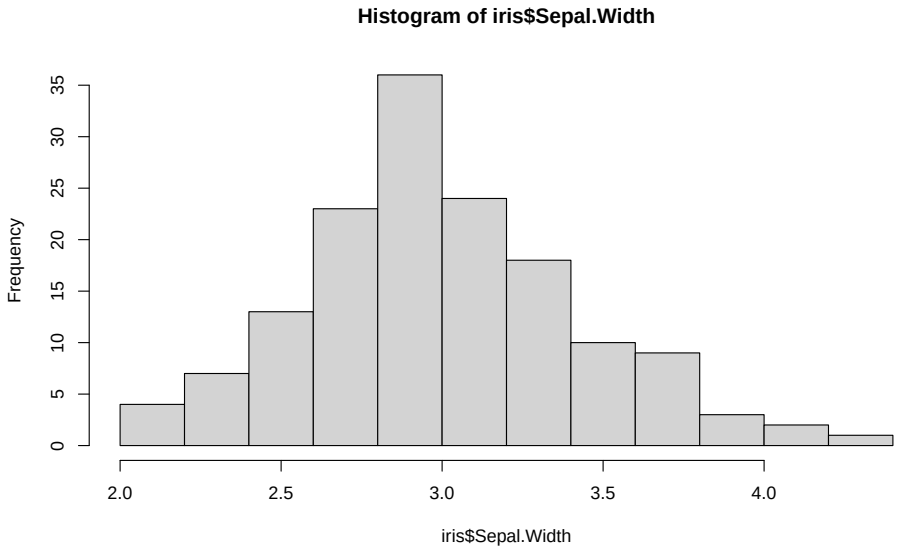
4.

```
aggregate(cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width) ~
  Species, data = iris, FUN = mean)
```

```
     Species Sepal.Length Sepal.Width Petal.Length Petal.Width
1 versicolor         5.94        2.77         4.26       1.326
2  virginica         6.59        2.97         5.55       2.026
3     setosa         5.01        3.43         1.46       0.246
```

5.

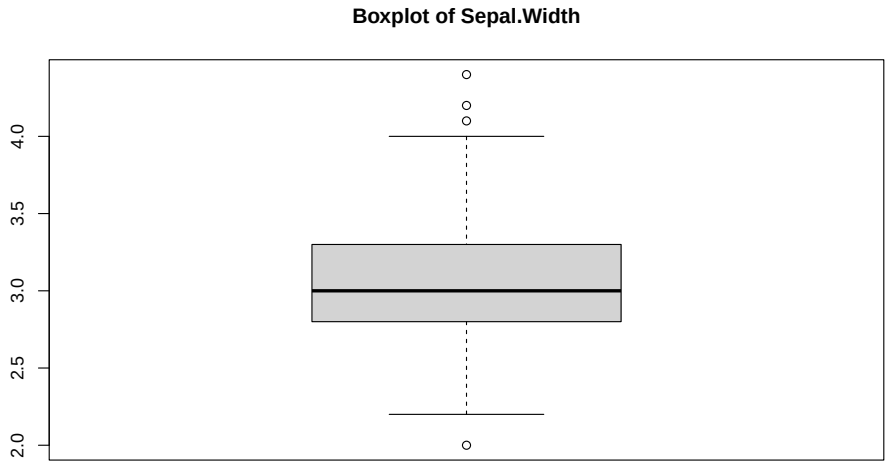
```
hist(iris$Sepal.Width)
```



6.

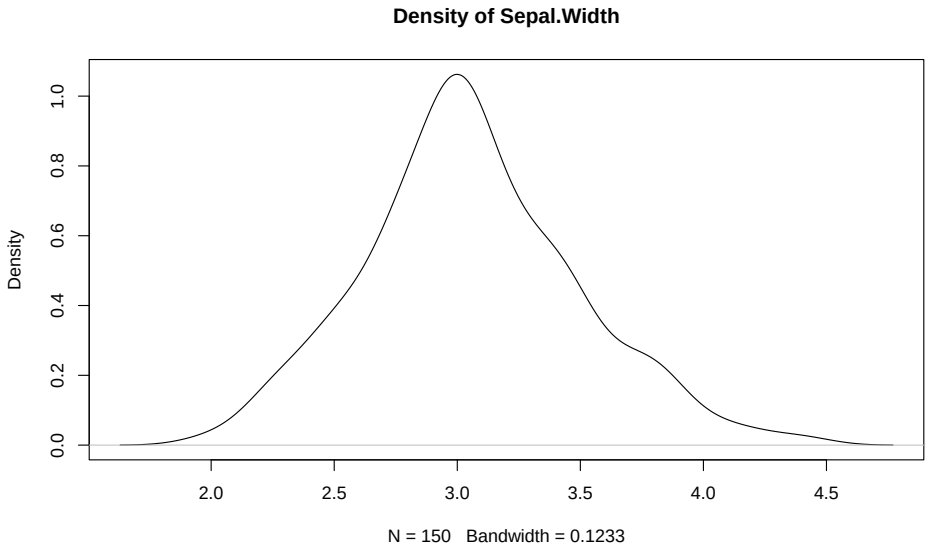
— Toute espèce
— Boxplot

```
boxplot(iris$Sepal.Width, main = "Boxplot of Sepal.Width")
```



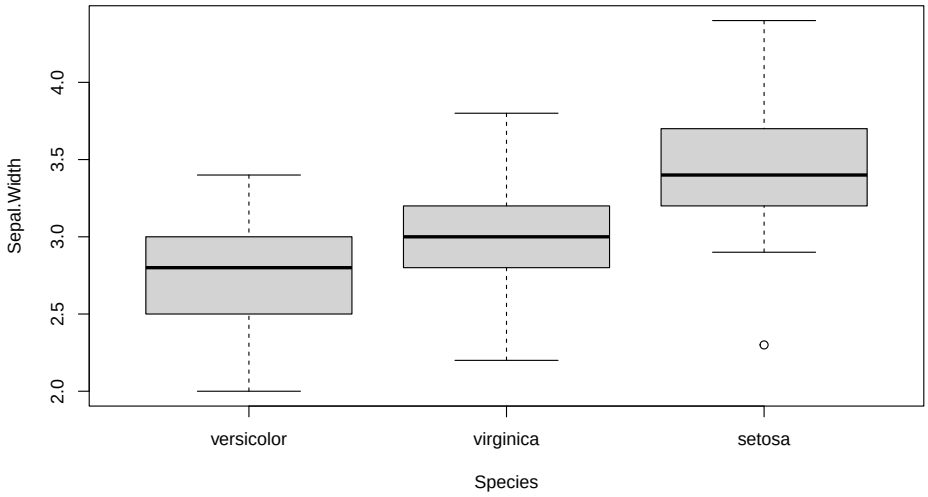
— Desnity plot

```
density(iris$Sepal.Width) |>
  plot(main = "Density of Sepal.Width")
```



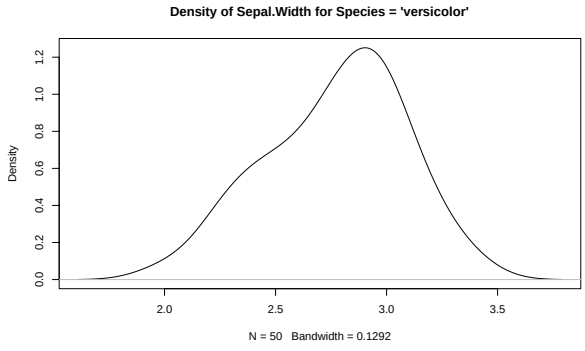
— Par espèce
  — Boxplot

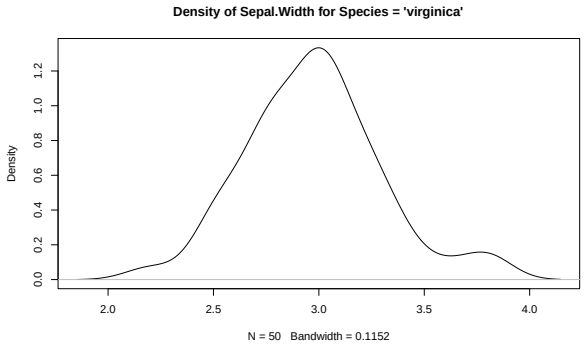
```
boxplot(Sepal.Width ~ Species, data = iris)
```

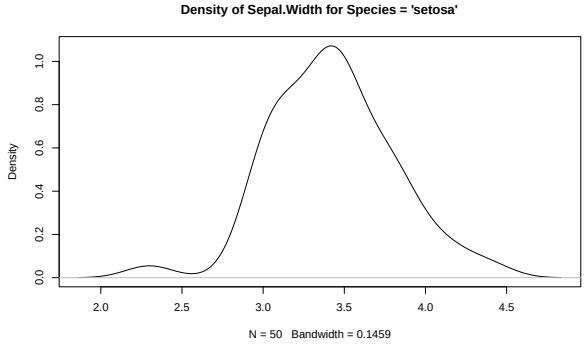


— Desnity plot

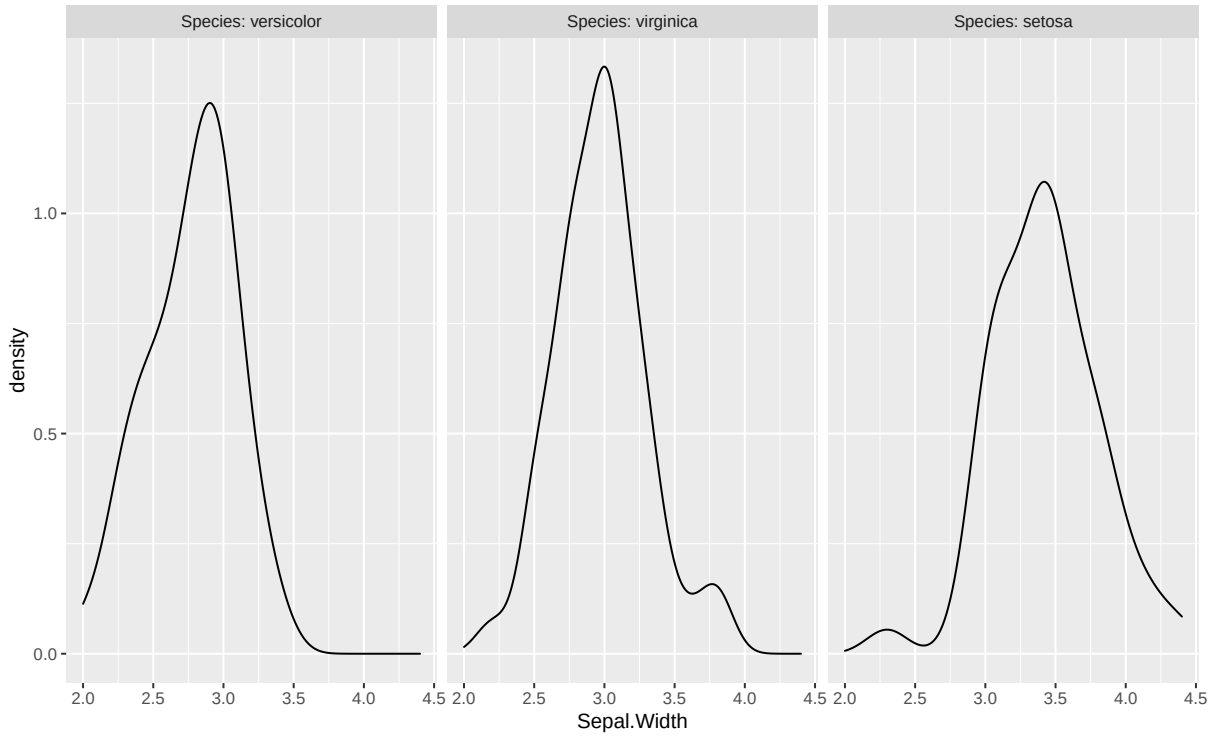
```
subset(iris, Species == "versicolor")$Sepal.Width |>
  density() |>
  plot(main = "Density of Sepal.Width for Species = 'versicolor'")
subset(iris, Species == "virginica")$Sepal.Width |>
  density() |>
  plot(main = "Density of Sepal.Width for Species = 'virginica'")
subset(iris, Species == "setosa")$Sepal.Width |>
  density() |>
  plot(main = "Density of Sepal.Width for Species = 'setosa'")
```







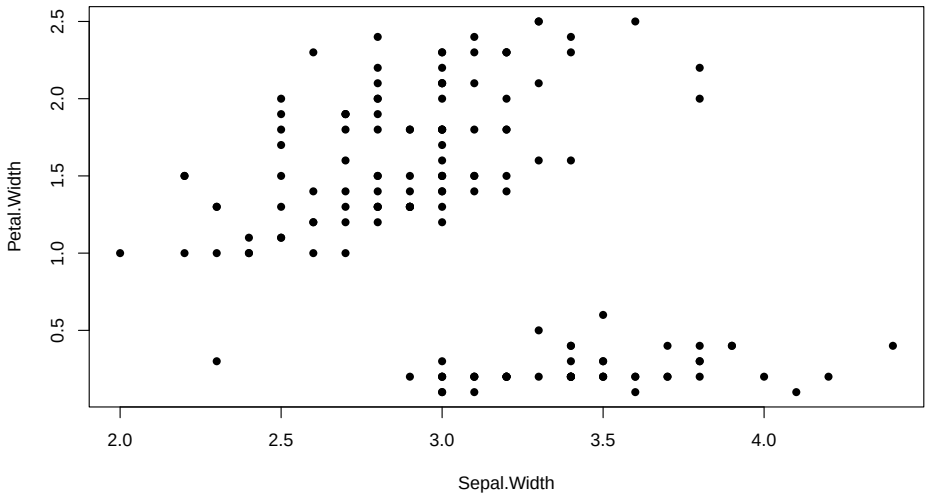
```r
# avec ggplot2 (code via esquisse::esquisser(iris))
ggplot(iris, aes(x = Sepal.Width)) + geom_density() + facet_wrap(vars(Species),
  labeller = label_both)
```



7.

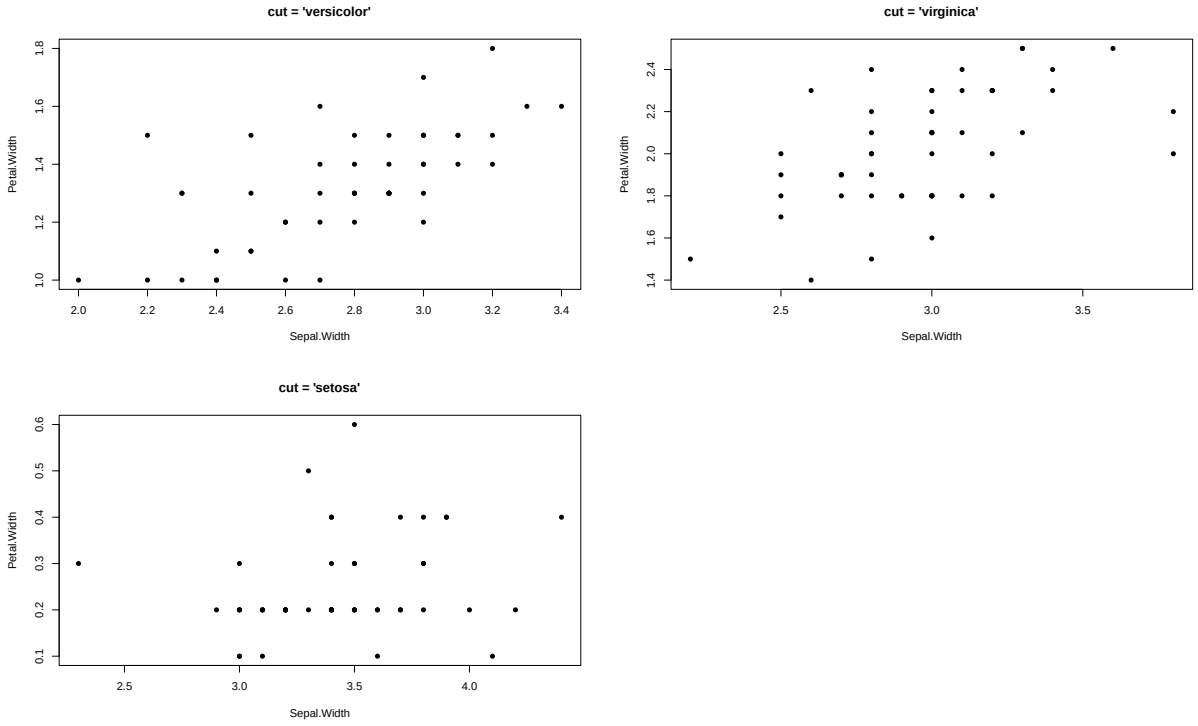
— Toute espèce

```r
plot(Petal.Width ~ Sepal.Width, data = iris, pch = 16)
```

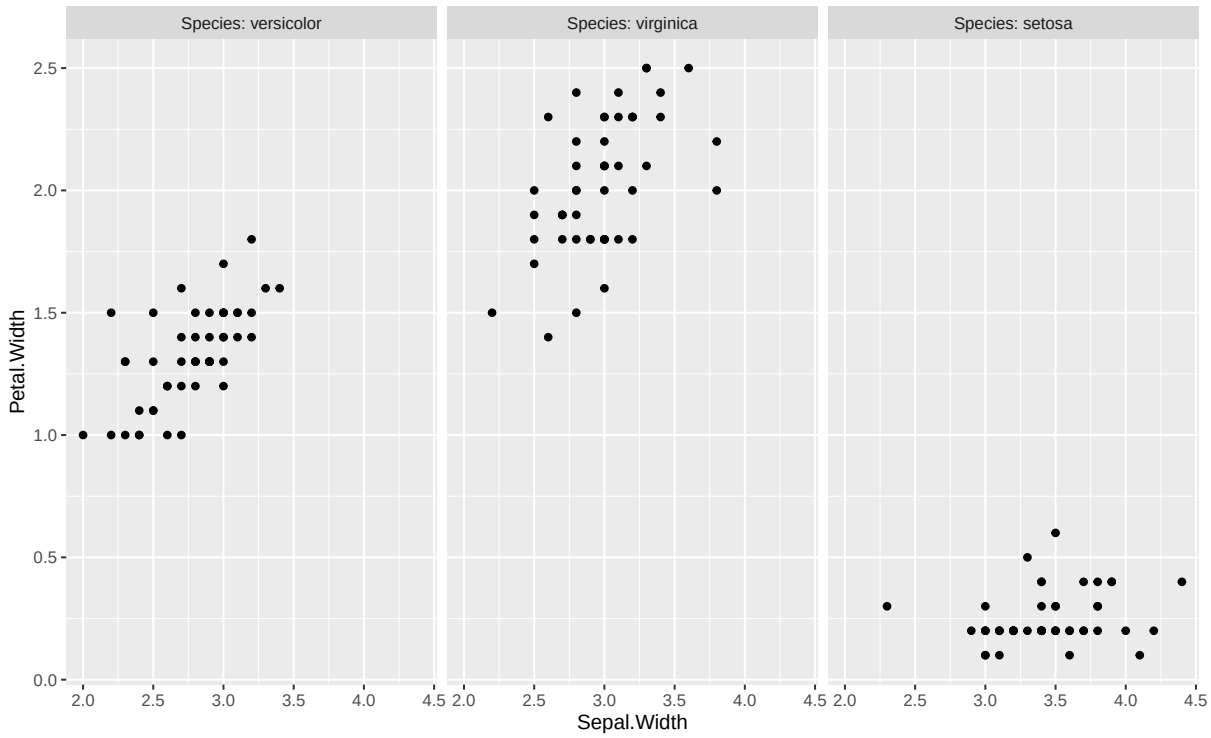


— Par espèce

```
plot(Petal.Width ~ Sepal.Width, data = subset(iris, Species ==
  "versicolor"), main = "cut = 'versicolor'", pch = 16)
plot(Petal.Width ~ Sepal.Width, data = subset(iris, Species ==
  "virginica"), main = "cut = 'virginica'", pch = 16)
plot(Petal.Width ~ Sepal.Width, data = subset(iris, Species ==
  "setosa"), main = "cut = 'setosa'", pch = 16)
```




```
# avec ggplot2 (code via esquisse::esquisser(iris))
ggplot(iris, aes(x = Sepal.Width, y = Petal.Width)) + geom_point() +
  facet_wrap(vars(Species), labeller = label_both)
```



8.

```
iris <- transform(iris, Petal.WidthC = cut(Petal.Width,
  breaks = quantile(Petal.Width, prob = c(0, 1 / 3, 2 / 3, 1)),
  labels = c("PW1", "PW2", "PW3")))
table(iris$Petal.WidthC)
```

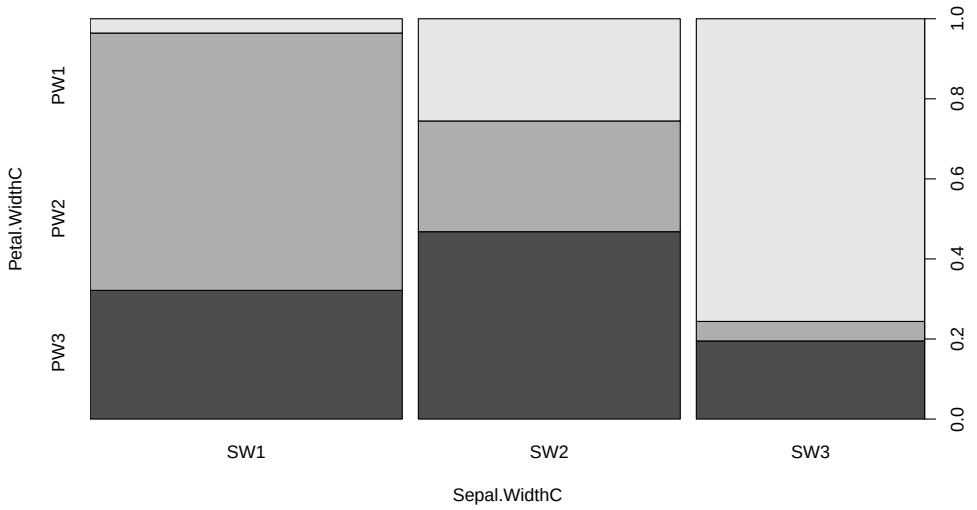
```
PW1 PW2 PW3
 45  52  48
```

9.

```
iris <- transform(iris, Sepal.WidthC = cut(Sepal.Width,
  breaks = quantile(Sepal.Width, prob = c(0, 1 / 3, 2 / 3, 1)),
  labels = c("SW1", "SW2", "SW3")))
table(iris$Sepal.WidthC)
```

```
SW1 SW2 SW3
 56  50  43
```

```
plot(Petal.WidthC ~ Sepal.WidthC, data = iris)
```



Exercice 2

```
library(ggplot2)
```

1.

— Effectifs

```
TBcut <- xtabs(~cut, data = diamonds)
TBcut
```

```
cut
     Fair      Good Very Good   Premium     Ideal
     1610      4906     12082     13791     21551
```

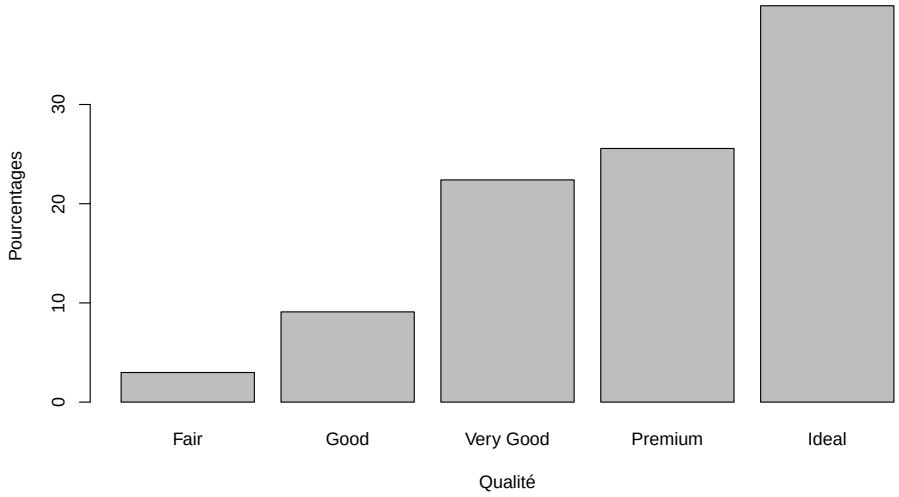
— Proportions

```
pTBcut <- proportions(TBcut) * 100
round(pTBcut, 2)
```

```
cut
     Fair      Good Very Good   Premium     Ideal
     2.98      9.10     22.40     25.57     39.95
```

— Barplot

```
barplot(pTBcut, xlab = "Qualité", ylab = "Pourcentages")
```



2.

— Effectifs

```
TBcutcolor <- xtabs(~color + cut, data = diamonds)
TBcutcolor
```

```
     cut
color Fair Good Very Good Premium Ideal
    D  163  662      1513    1603  2834
    E  224  933      2400    2337  3903
    F  312  909      2164    2331  3826
    G  314  871      2299    2924  4884
    H  303  702      1824    2360  3115
    I  175  522      1204    1428  2093
    J  119  307       678     808   896
```

— Pourcentages

```
pTBcutcolor <- proportions(TBcutcolor, "cut") * 100
round(addmargins(pTBcutcolor, 1), 2)
```

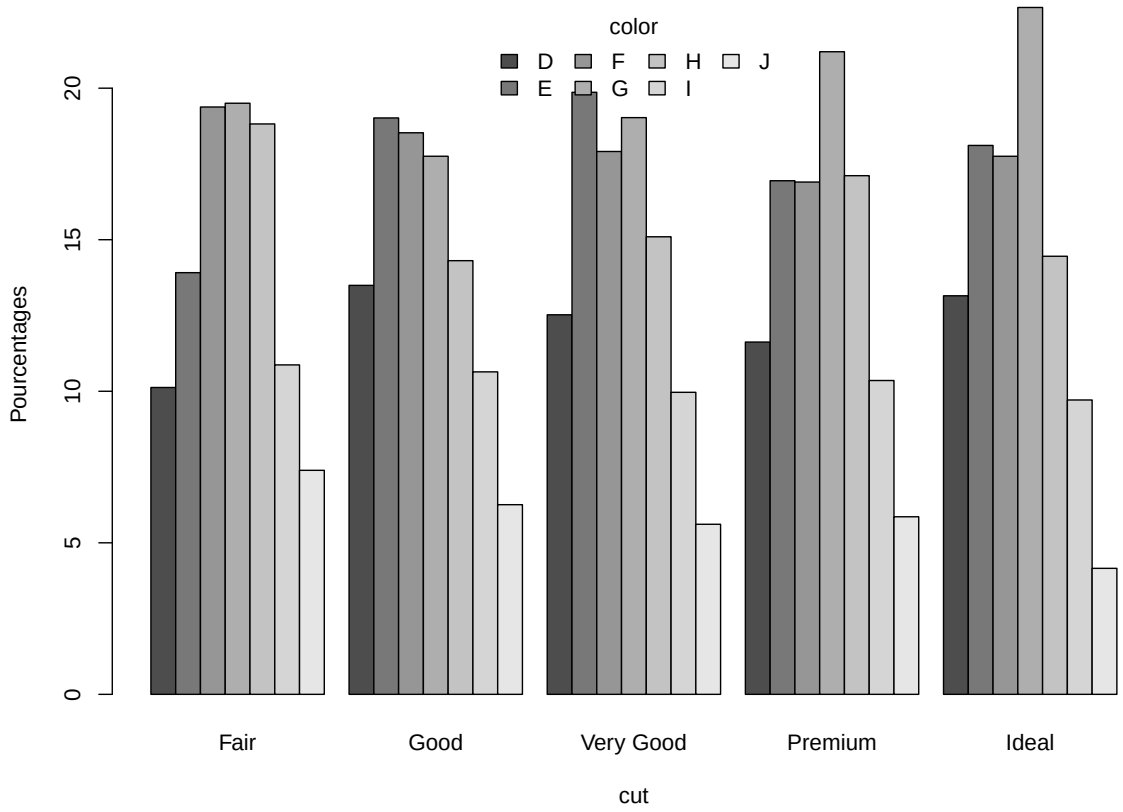
```
     cut
color   Fair   Good Very Good Premium  Ideal
  D    10.12  13.49     12.52   11.62  13.15
  E    13.91  19.02     19.86   16.95  18.11
  F    19.38  18.53     17.91   16.90  17.75
  G    19.50  17.75     19.03   21.20  22.66
  H    18.82  14.31     15.10   17.11  14.45
  I    10.87  10.64      9.97   10.35   9.71
  J     7.39   6.26      5.61    5.86   4.16
  Sum 100.00 100.00    100.00  100.00 100.00
```

→ 13.15% des diamants ayant une découpe “Ideal” ont la meilleure couleur possible (D).

3.

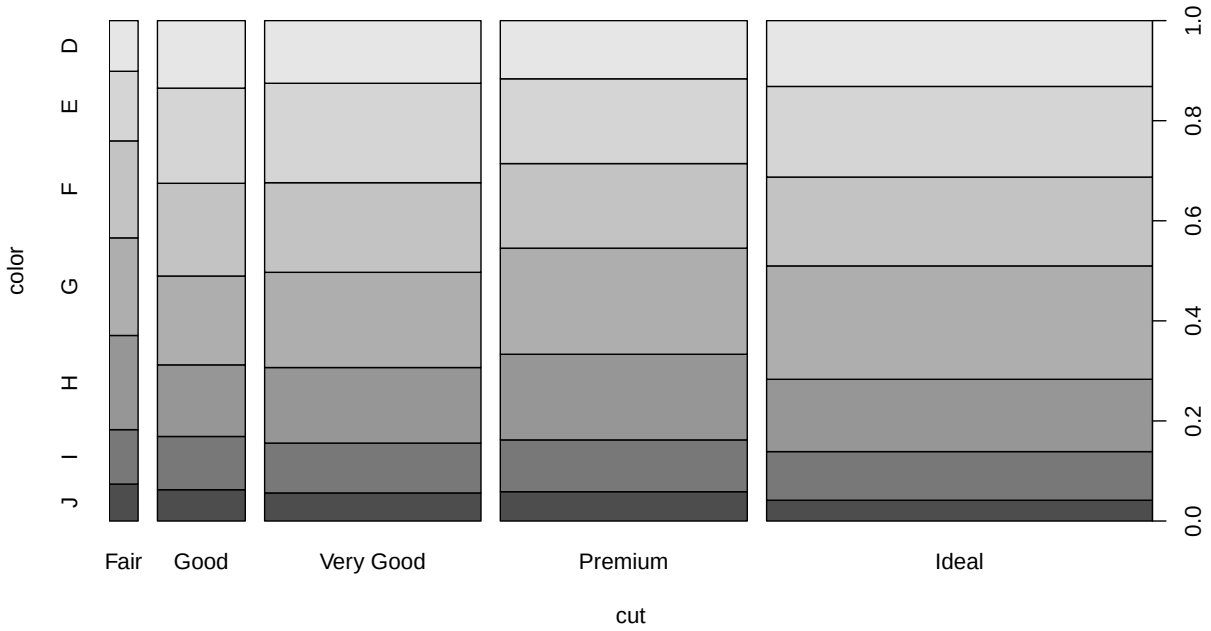
— Barplot

```
barplot(pTBcutcolor, beside = TRUE, legend = TRUE, args.legend = list(bty = "n",
  x = "top", ncol = 4, title = "color"), xlab = "cut", ylab = "Pourcentages")
```



— Spineplot

```
plot(color ~ cut, data = diamonds)
```



4.

```
summary(diamonds$price)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18823
```

```
summary(diamonds$carat)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.20    0.40    0.70    0.80    1.04    5.01
```

5.

```
aggregate(price ~ color, data = diamonds, FUN = summary)
```

```
  color price.Min. price.1st Qu. price.Median price.Mean price.3rd Qu.
1     D        357           911         1838       3170          4214
2     E        326           882         1739       3077          4003
3     F        342           982         2344       3725          4868
4     G        354           931         2242       3999          6048
5     H        337           984         3460       4487          5980
6     I        334          1120         3730       5092          7202
7     J        335          1860         4234       5324          7695
```

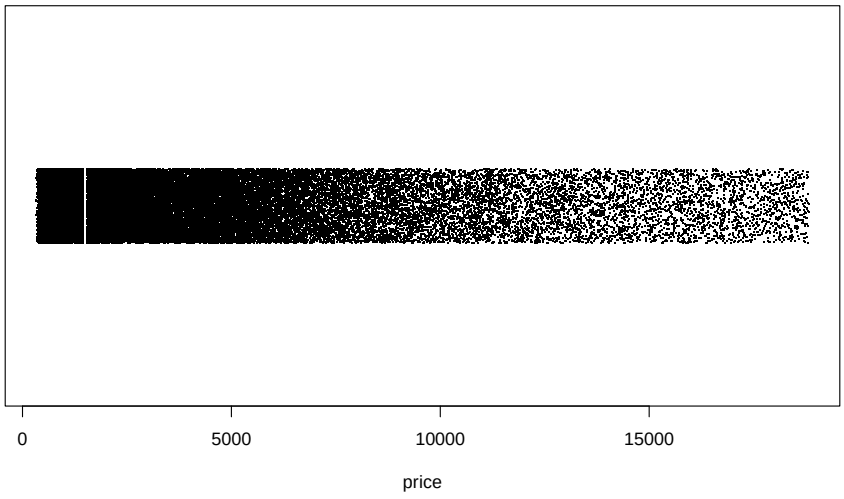
```
  price.Max.
1      18693
2      18731
3      18791
4      18818
5      18803
6      18823
7      18710
```

```
aggregate(price ~ cut, data = diamonds, FUN = summary)
```

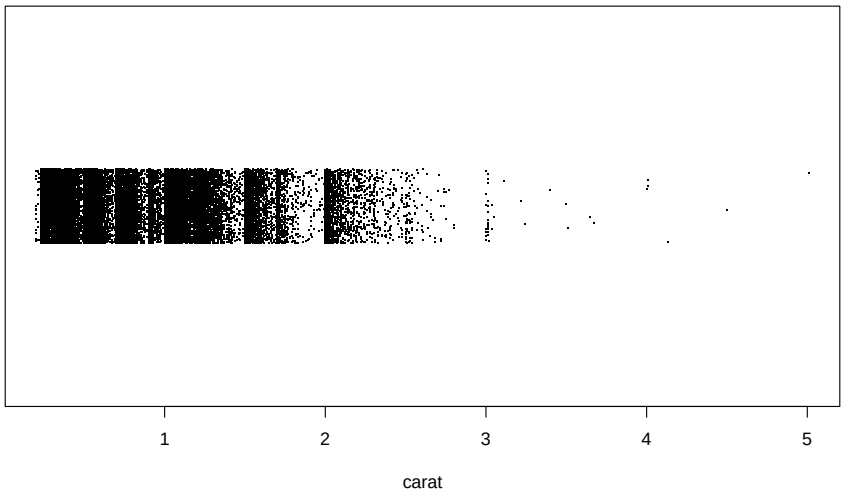
```
        cut price.Min. price.1st Qu. price.Median price.Mean price.3rd Qu.
1      Fair         337          2050         3282       4359          5206
2      Good         327          1145         3050       3929          5028
3 Very Good         336           912         2648       3982          5373
4   Premium         326          1046         3185       4584          6296
5     Ideal         326           878         1810       3458          4678
  price.Max.
1      18574
2      18788
3      18818
4      18823
5      18806
```

6.

```
stripchart(diamonds$price, xlab = "price", pch = ".", method = "jitter")
```

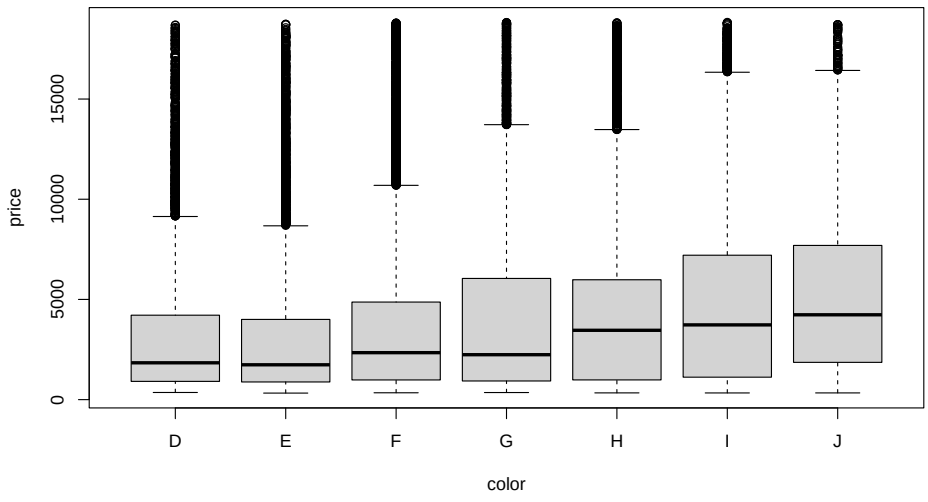


```
stripchart(diamonds$carat, xlab = "carat", pch = ".", method = "jitter")
```

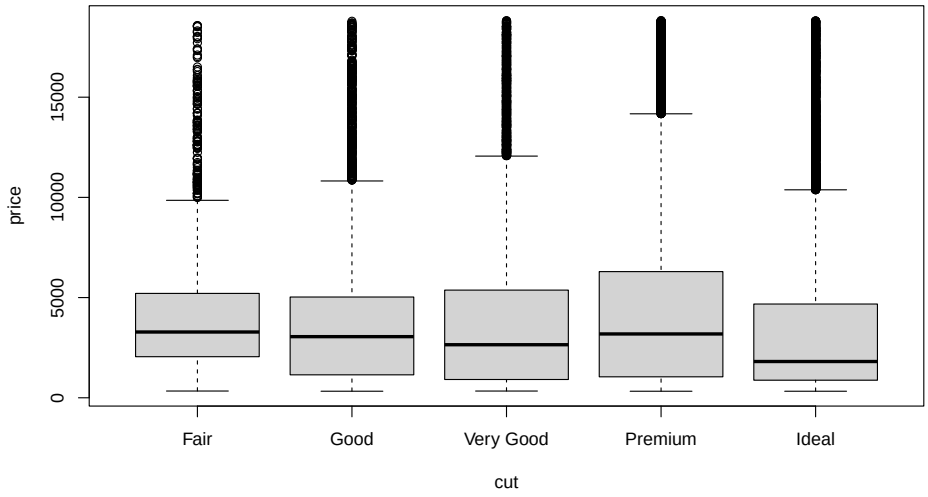


7.

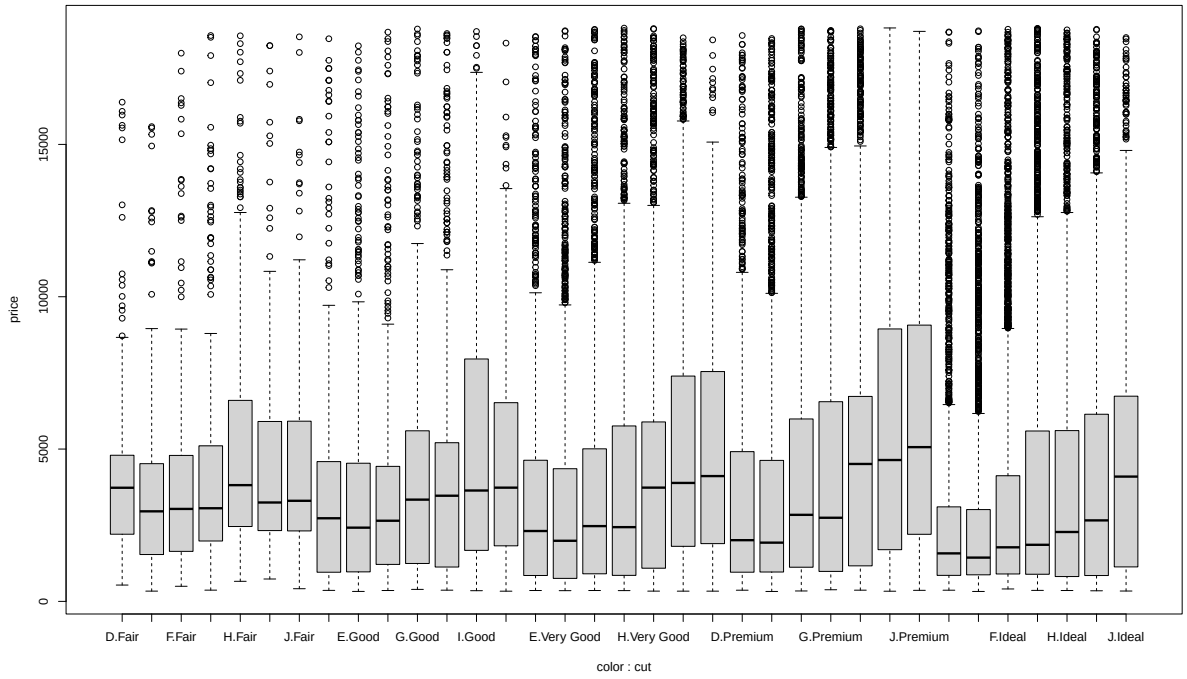
```
boxplot(price ~ color, data = diamonds)
```



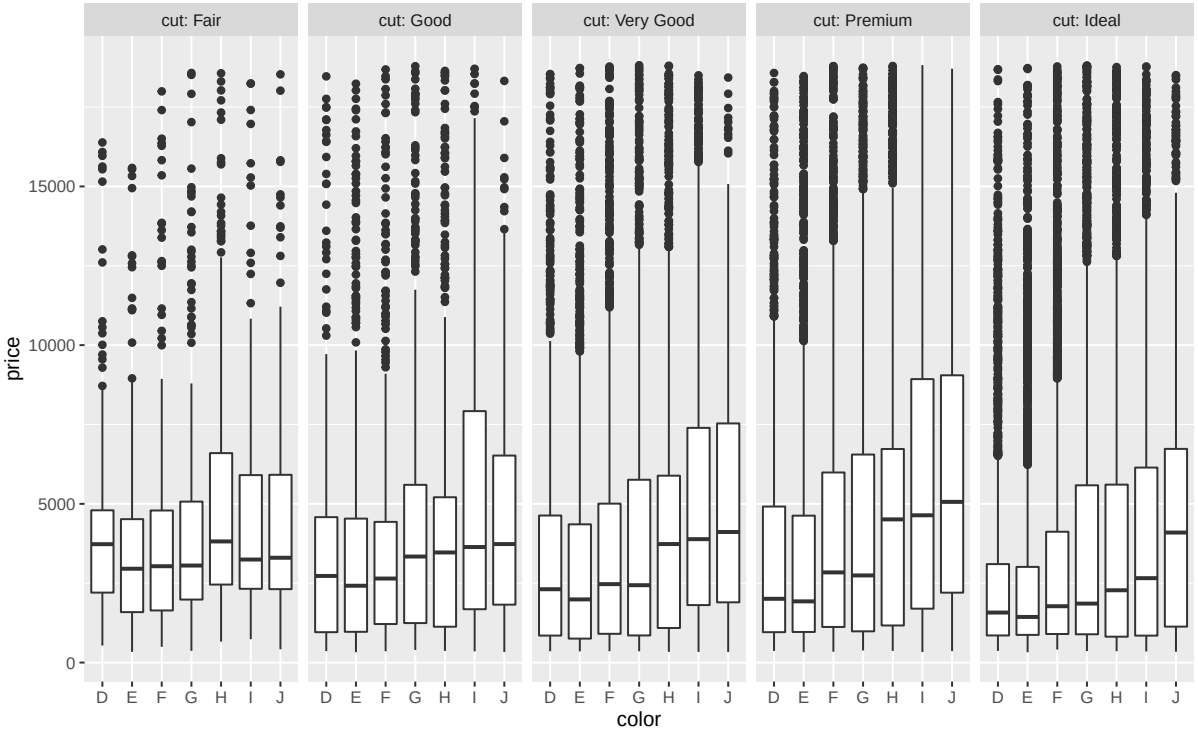

```
boxplot(price ~ cut, data = diamonds)
```




```
boxplot(price ~ color + cut, data = diamonds)
```

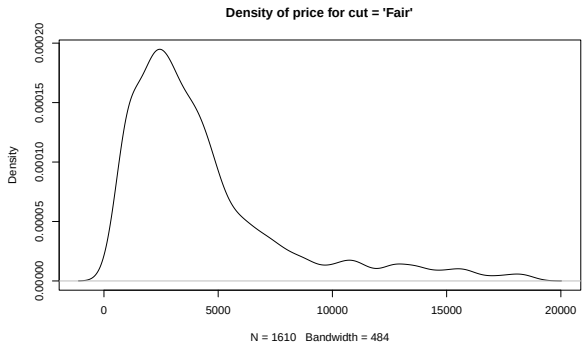


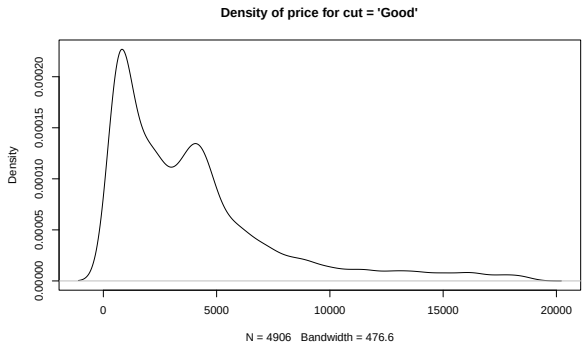

```
# avec ggplot2 (code via esquisse::esquisser(diamonds))
ggplot(diamonds, aes(x = color, y = price)) + geom_boxplot() +
  facet_wrap(facets = vars(cut), ncol = 5, labeller = label_both)
```

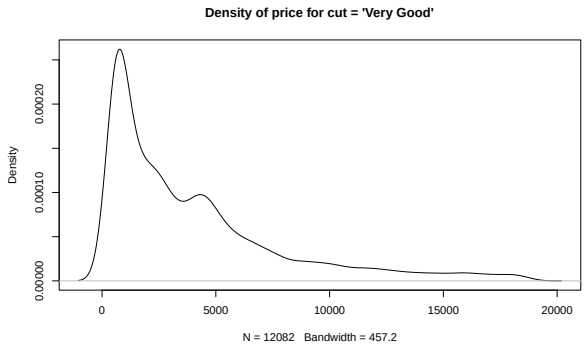


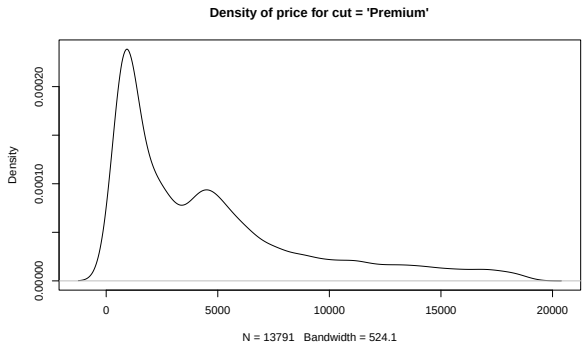
8.

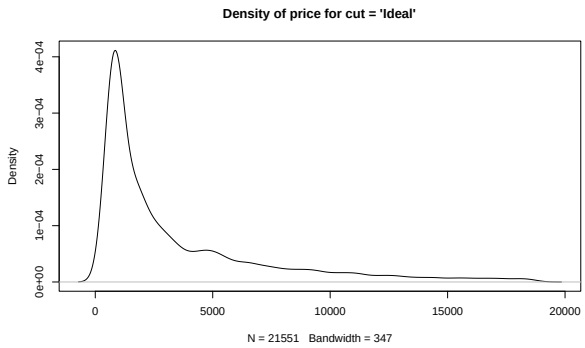
```
subset(diamonds, subset = cut == "Fair")$price |>
  density() |>
  plot(main = "Density of price for cut = 'Fair'")
subset(diamonds, subset = cut == "Good")$price |>
  density() |>
  plot(main = "Density of price for cut = 'Good'")
subset(diamonds, subset = cut == "Very Good")$price |>
  density() |>
  plot(main = "Density of price for cut = 'Very Good'")
subset(diamonds, subset = cut == "Premium")$price |>
  density() |>
  plot(main = "Density of price for cut = 'Premium'")
subset(diamonds, subset = cut == "Ideal")$price |>
  density() |>
  plot(main = "Density of price for cut = 'Ideal'")
```



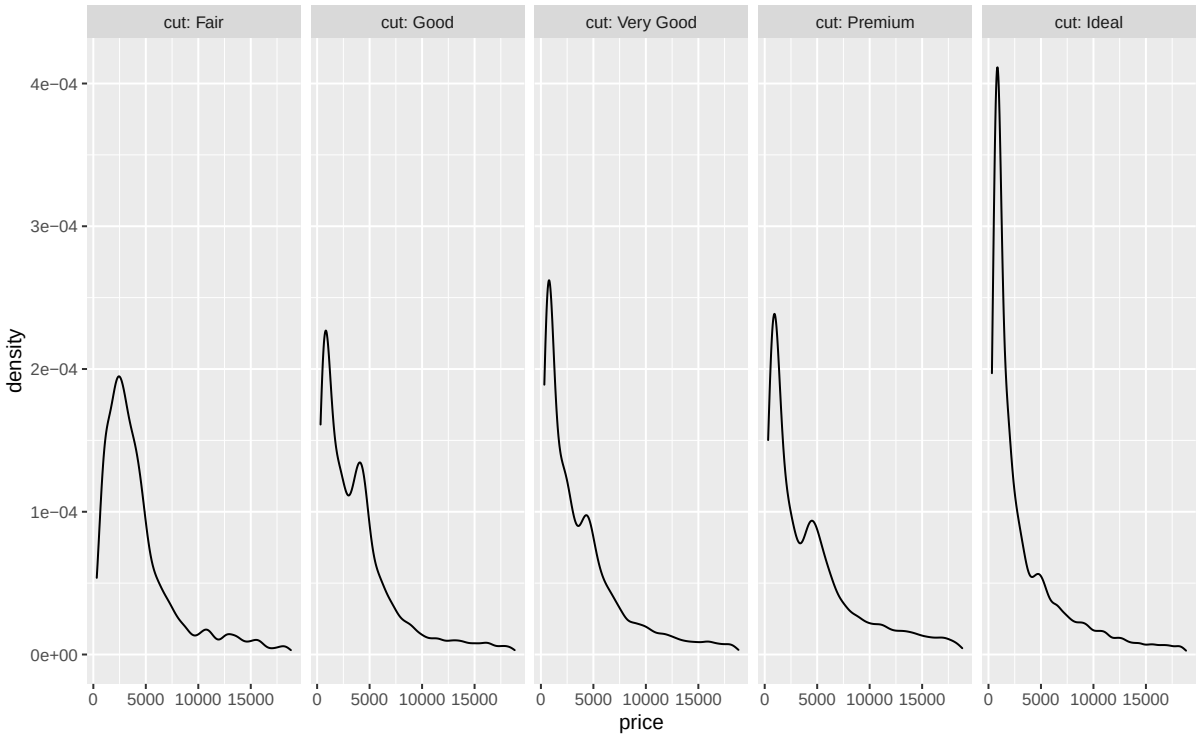






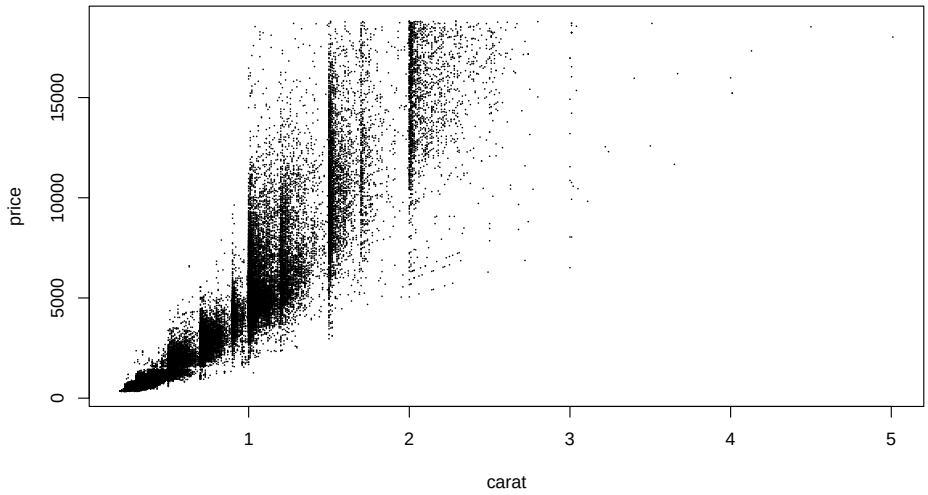



```
# avec ggplot2 (code via esquisse::esquisser(diamonds))
ggplot(diamonds, aes(x = price)) + geom_density() + facet_wrap(vars(cut),
  ncol = 5, labeller = label_both)
```



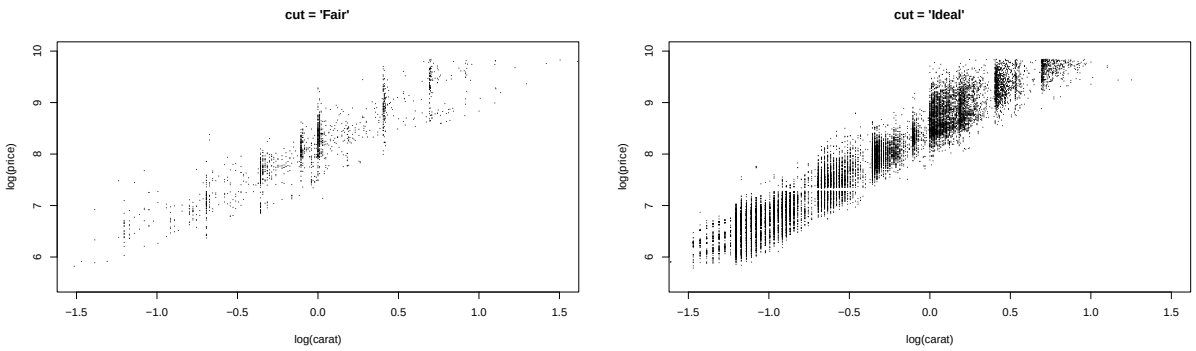
9.

```
plot(price ~ carat, data = diamonds, pch = 16, cex = 0.2)
```



10.

```
plot(log(price) ~ log(carat), data = subset(diamonds, cut ==
  "Fair"), pch = 16, cex = 0.2, ylim = c(5.5, 10), xlim = c(-1.5,
  1.5), main = "cut = 'Fair'")
plot(log(price) ~ log(carat), data = subset(diamonds, cut ==
  "Ideal"), pch = 16, cex = 0.2, ylim = c(5.5, 10), xlim = c(-1.5,
  1.5), main = "cut = 'Ideal'")
```



Dans les deux cas, on constate une relation linéaire positive. Les deux pentes semblent très similaires.