ANALYSE ECONOMETRIQUE DU PRIX DES DIAMANTS

Projet réalisé par Asmae DADI et Philippine RENAUDIN, étudiantes en M1 Data Science à l'UFR Science de l'Université d'Angers

Professeur :
DANIEL Christophe

Table des matières

I. Introduction

Traditionnellement, la présentation d'une bague en métal rare ornée d'une pierre précieuse à une personne, est emblématique du gage d'amour et d'engagement.

Cette pierre précieuse, reste aujourd'hui encore très majoritairement le diamant. L'industrie du diamant connaît depuis le début de 20ème siècle un essor inégalé dans le monde des pierres précieuses. Son succès quant à lui peut en grande partie être expliqué par ses nombreuses caractéristiques tels que sa dureté, sa pureté, son symbolisme, ses propriétés chimiques, optiques, électriques et thermiques, ainsi que son prix.

En effet, le diamant reste une pierre coûteuse. C'est pourquoi nous allons chercher à savoir qu'elles sont les caractéristiques du diamant qui font influencer son prix.

Les hypothèses que nous pouvons faire à ce niveau de l'étude sont les dires courants : plus le diamant a de carats plus son prix augmente. De même, plus un diamant a une clarté, une coupe et une couleur rare, plus son prix sera élevé. Pour l'explication globale du prix d'un diamant, nous pensons que les 3 variables qualitatives (coupe, couleur et clarté) et les variables représentant les dimensions et poids du diamant joueront un rôle plus important que les 2 dernières variables.

Nous travaillerons donc dans cette étude de cas, avec un jeu de données contenant plusieurs mesures de différentes caractéristiques de diamants.

II. La base de données

Notre étude aura pour support une base de données "diamonds" disponible dans la librairie ggplot2 du logiciel Rstudio. Ce jeu de données est constitué de 53940 enregistrements et de 10 variables.

Description des variables:

carat : variable quantitative représentant le poids d'un diamant en carat (1 carat=0.2 gramme)

table : variable quantitative représentant la largeur du sommet du diamant par rapport au point le plus haut.

depth : variable quantitative représentant le pourcentage de profondeur totale. Cette variable est une fonction des 3 variables suivantes (2 * z / (x + y))

x : variable quantitative représentant la longueur d'un diamant, en mm.

y: variable quantitative représentant la largeur d'un diamant, en mm.

z: variable quantitative représentant la profondeur d'un diamant, en mm.

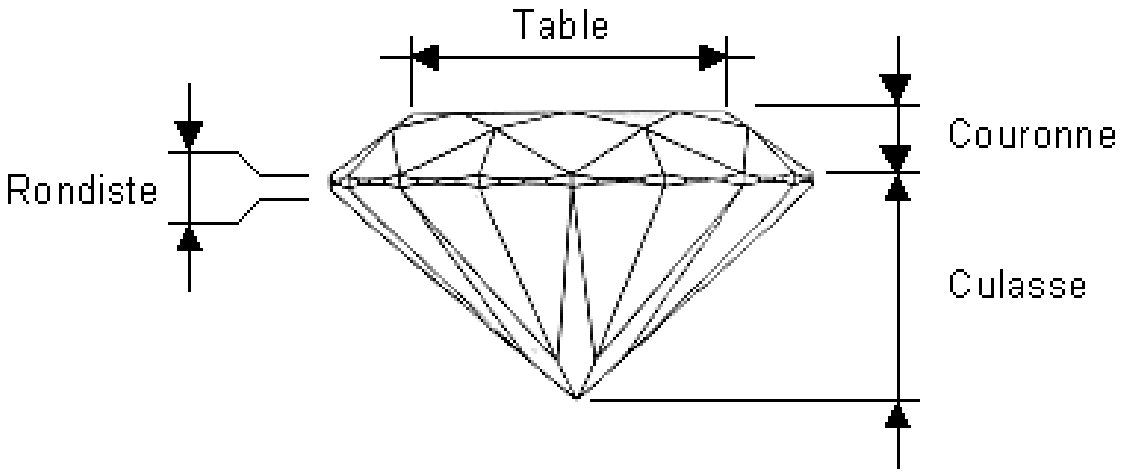
cut : variable qualitative ordonnée représentant la qualité de la coupe d'un diamant. 5 modalités : *Fair<Good<VeryGood<Premium<Ideal*

color : variable qualitative ordonnée représentant la couleur d'un diamant. 7 modalités : *D<E<F<G<H<I<J* (D représentant la meilleure couleur, J la pire. Le classement semble avoir été fait dans le sens inverse de la variable cut)

clarity : variable qualitative ordonnée mesurant la clarté d'un diamant. 8 modalités : *I1<SI2<SI1<VS2<VS1<VVS12<VVS1<IF* (I1 représentant la pire clarté, IF la meilleure)

price : variable quantitative représentant le prix d'un diamant, en dollars.

Remarque : Notre jeu de données étant très volumineux, nous allons étudier qu'un échantillon représentant 10% de ce jeu de données lors de nos différentes régressions.



III. Analyse descriptive

a. Analyse globale

```
str(diamonds)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    53940 obs. of  10 variables:
##  $ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 .
..
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5
...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4
5 ...
##  $ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num  55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
```

```
##  $ y       : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z       : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

On retrouve bien les 3 variables qualitatives **cut**, **color** et **clarity** avec leurs différentes modalités. Toutes les autres variables sont des variables quantitatives. La variable **price** semble triée par ordre croissant mais en explorant la base, on se rend compte que ce n'est pas le cas.

```
summary(diamonds)
```

```
     carat               cut        color        clarity          depth      
Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065   Min.   :43.00  
1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258   1st Qu.:61.00  
Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80  
Mean   :0.7979   Premium   :13791   G:11292   VS1    : 8171   Mean   :61.75  
3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50  
Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00  
                                   J: 2808   (Other): 2531                  
##      table           price             x                y         
##  Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000  
##  1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720  
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710  
##  Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735  
##  3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540  
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900  
##
##        z         
##  Min.   : 0.000  
##  1st Qu.: 2.910  
##  Median : 3.530  
##  Mean   : 3.539  
##  3rd Qu.: 4.040  
##  Max.   :31.800  
##
```

La variable **cut** met en évidence que dans ce jeu de données, il y a en majorité des diamants de coupe *Ideal*, puis de coupe *Premium* puis de coupe *Very Good*. Les coupes *Good* et *Fair* sont très peu représentées. Il faudra donc faire attention à respecter les proportions lors de l'échantillonnage.

En revanche pour la variable **clarity**, ce sont les modalités représentant une qualité intermédiaire qui sont les plus représentées. Il semble y avoir peu de diamants de clarté *IF* et *I1*.

La variable **color** semble être répartie de manière gaussienne, avec la couleur intermédiaire *G* la plus présente dans ce jeu de données.

Pour ce qui est des variables quantitatives, on remarque que l'étendu de la variable **carat** est assez grande. Cela veut sûrement dire qu'il y a des diamants très différents dans ce jeu de données.

Il en est de même pour la variable **price** (le diamant le moins cher étant à 326 dollars et le plus cher à 18823 dollars). La moyenne des prix est elle de 3933 dollars et la médiane de 2401 dollars. La moitié des diamants de ce jeu de données ont donc des prix inférieurs à 2400 dollars. Et le 3ème quartile indique que 75% des diamants ont des prix inférieurs à 5324 dollars. La majorité des diamants présents dans ce jeu ont donc des prix inférieurs à 6000 dollars.

Aux vues des valeurs que peuvent prendre les variables x, y et z, on peut simplement dire que les diamants de ce jeu de données ont tendances à être plus larges et profonds que longs.

Changeons maintenant le nom des colonnes pour avoir plus de lisibilité :

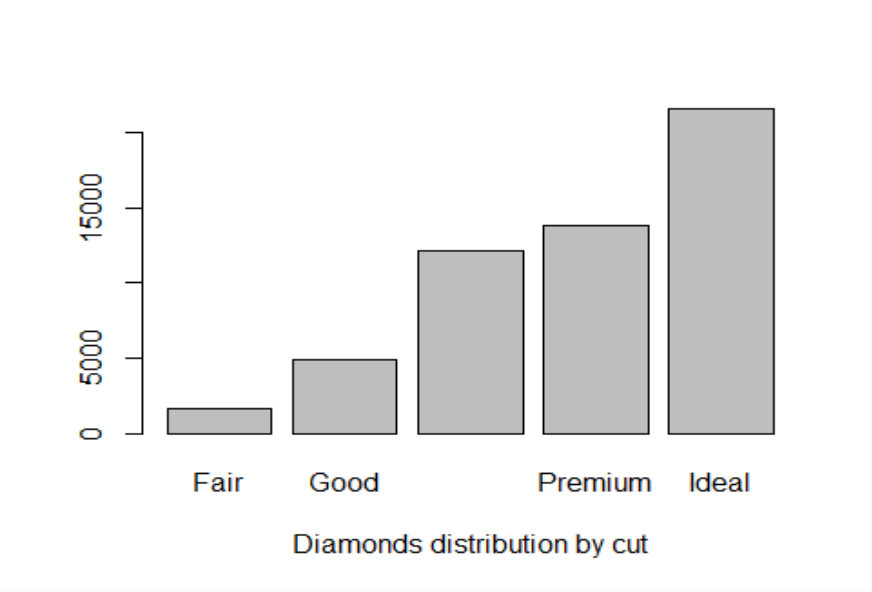
```
colnames(diamonds) <- c("carat",
                        "cut",
                        "color",
                        "clarity",
                        "fdepth",
                        "table",
                        "price",
                        "length",
                        "width",
                        "depth")
diamonds
```

```
## # A tibble: 53,940 x 10
##    carat cut       color clarity fdepth table price length width depth
##    <dbl> <ord>     <ord> <ord>    <dbl> <dbl> <int>  <dbl> <dbl> <dbl>
##  1 0.23  Ideal     E     SI2       61.5    55   326   3.95  3.98  2.43
##  2 0.21  Premium   E     SI1       59.8    61   326   3.89  3.84  2.31
##  3 0.23  Good      E     VS1       56.9    65   327   4.05  4.07  2.31
##  4 0.290 Premium   I     VS2       62.4    58   334   4.2   4.23  2.63
##  5 0.31  Good      J     SI2       63.3    58   335   4.34  4.35  2.75
##  6 0.24  Very Good J     VVS2      62.8    57   336   3.94  3.96  2.48
##  7 0.24  Very Good I     VVS1      62.3    57   336   3.95  3.98  2.47
##  8 0.26  Very Good H     SI1       61.9    55   337   4.07  4.11  2.53
##  9 0.22  Fair      E     VS2       65.1    61   337   3.87  3.78  2.49
## 10 0.23  Very Good H     VS1       59.4    61   338   4     4.05  2.39
## # ... with 53,930 more rows
```

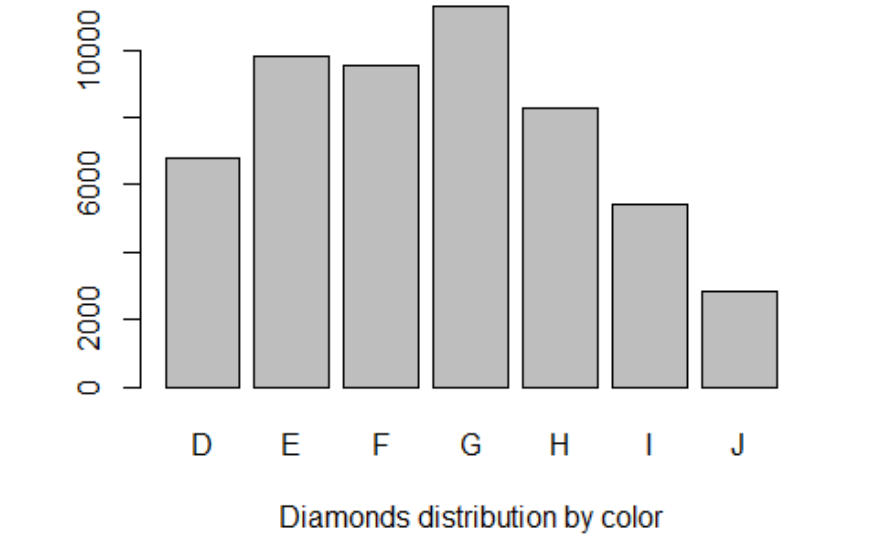
b. Analyse descriptive des variables explicatives

Représentons graphiquement les distributions des différentes variables explicatives qualitatives :

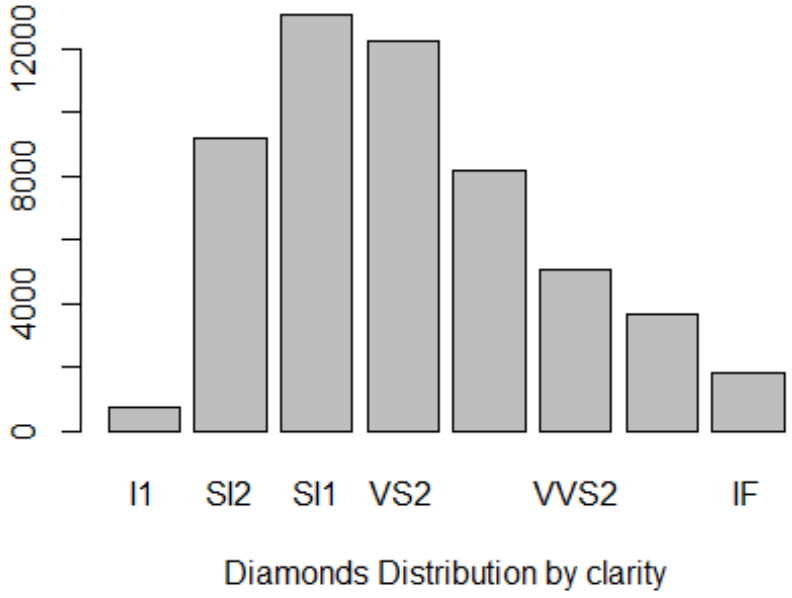
```
plot(diamonds$cut, xlab="Diamonds distribution by cut")
```




```
plot(diamonds$color, xlab="Diamonds distribution by color")
```



```
plot(diamonds$clarity, xlab="Diamonds Distribution by clarity")
```

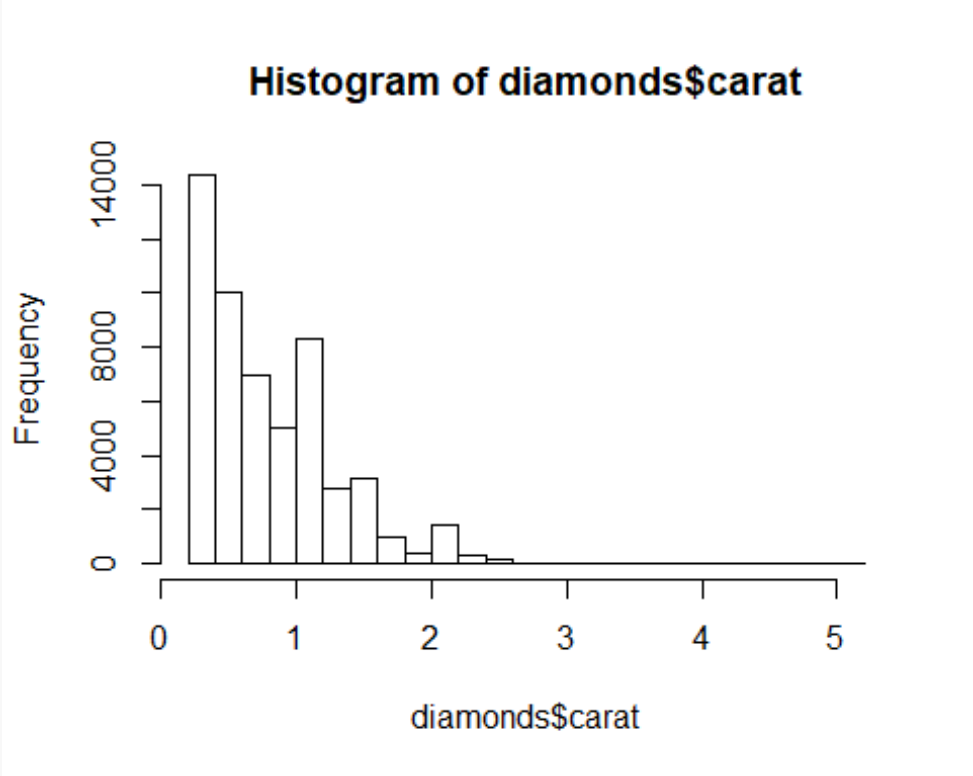


On retrouve bien graphiquement ce qui a été dit précédemment quant aux distributions des variables qualitatives.

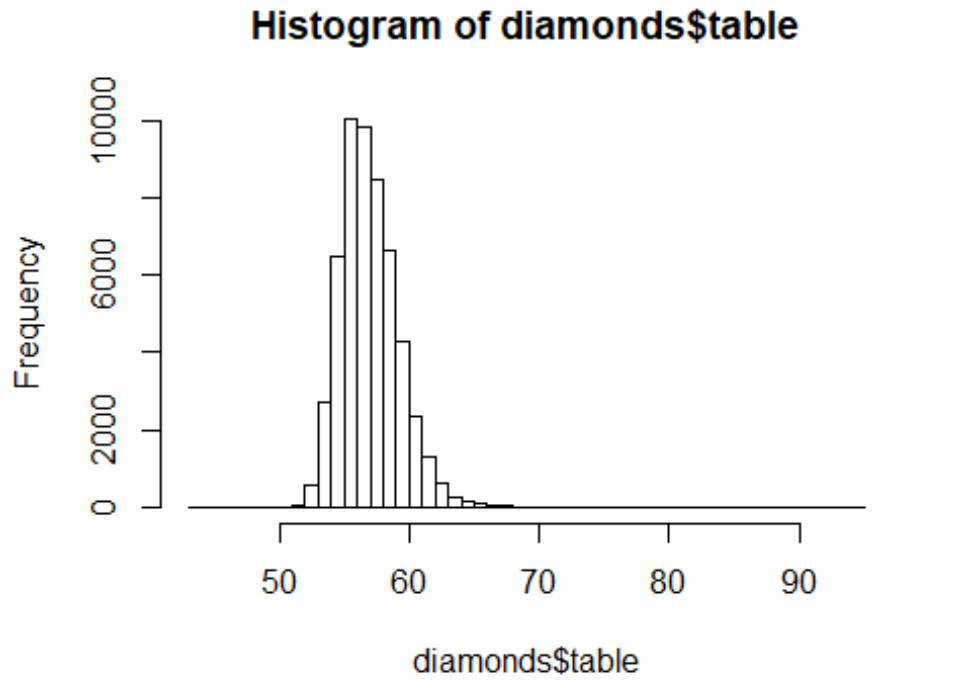
La sous-représentation de certaines modalités peut être une caractéristique du marché des diamants : peut-être ne met-on pas en vente un diamant de coupe médiocre à part s'il possède d'autres aspects plus intéressants. Pour la couleur d'un diamant, peut-être est-il rare de trouver une source de minerai donnant une "bonne" couleur. Et peut-être que peu de diamants ayant une "mauvaise" couleur sont mis en vente, à part si ceux-ci possèdent d'autres caractéristiques intéressantes. Pour la clarté, il est sûrement très rare de trouver des diamants ayant la meilleure clarté (I1) et encore une fois il est possible qu'il soit coutume de ne pas mettre en vente de diamants ayant une qualité de clarté médiocre si celui-ci ne possède pas de caractéristiques intéressantes par ailleurs.

Regardons maintenant la distribution des variables quantitatives :

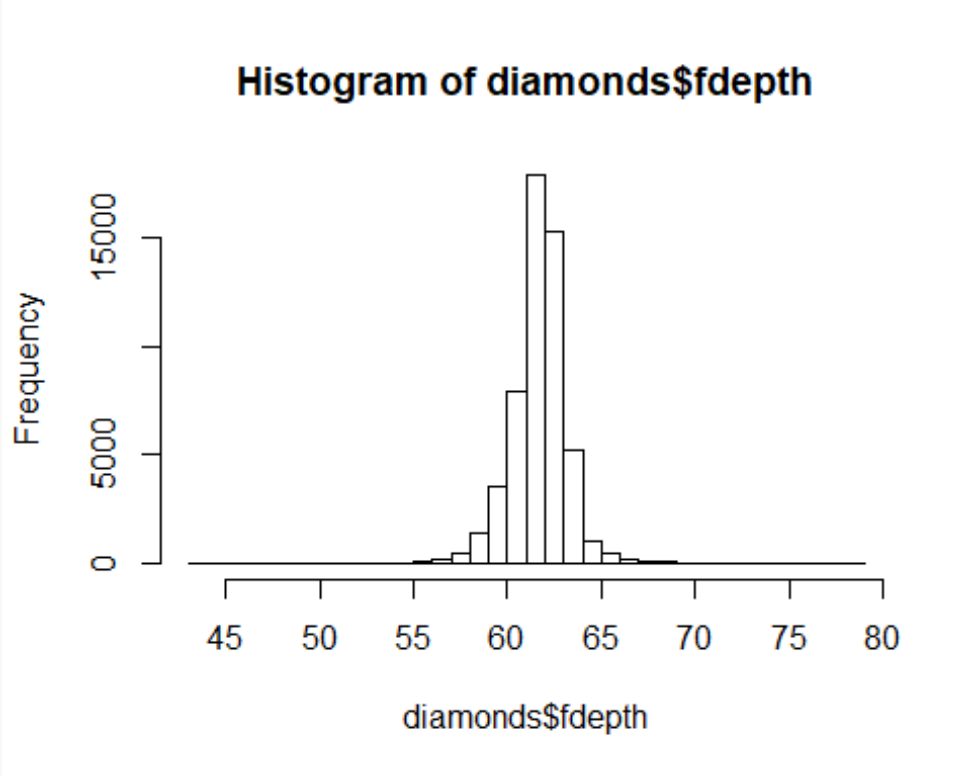
```
hist(diamonds$carat,breaks=25)
```



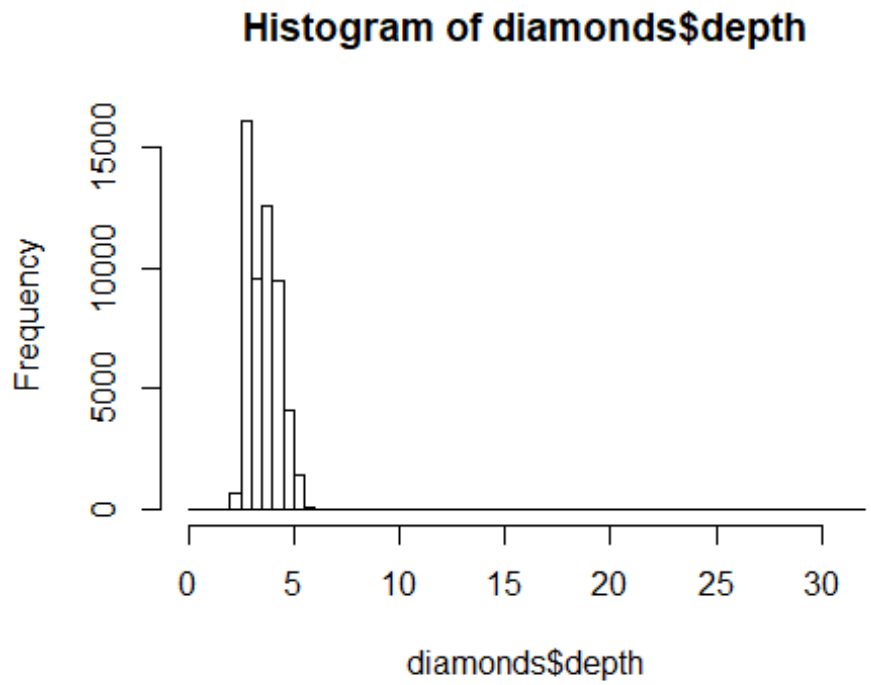

```
hist(diamonds$table,breaks=50)
```



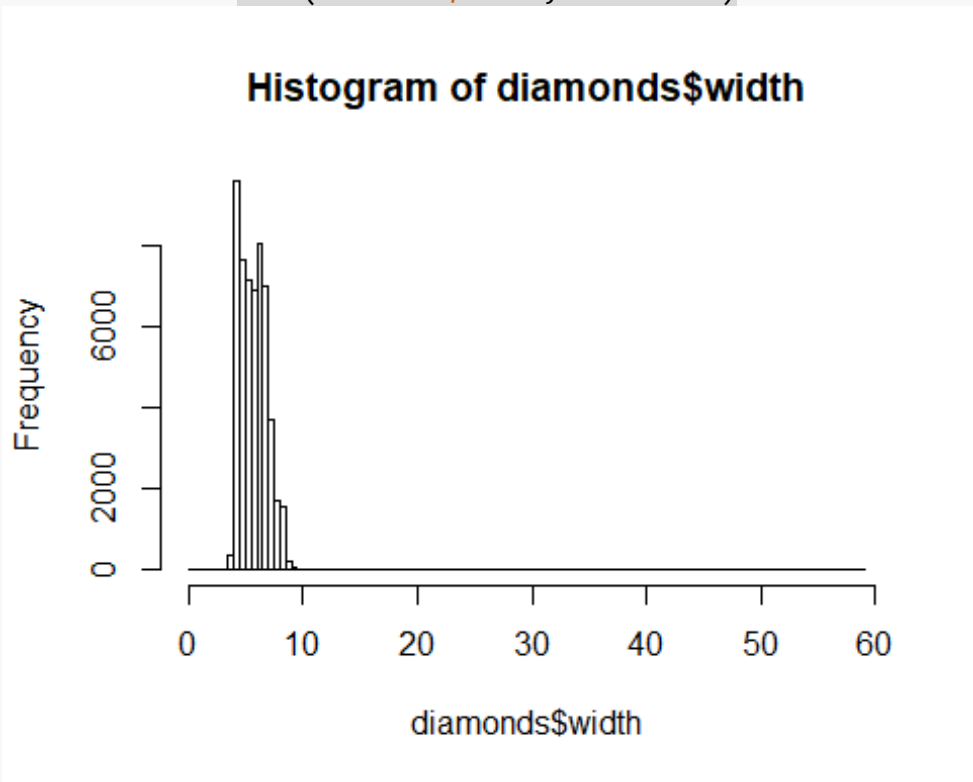
```
hist(diamonds$fdepth,breaks=50)
```



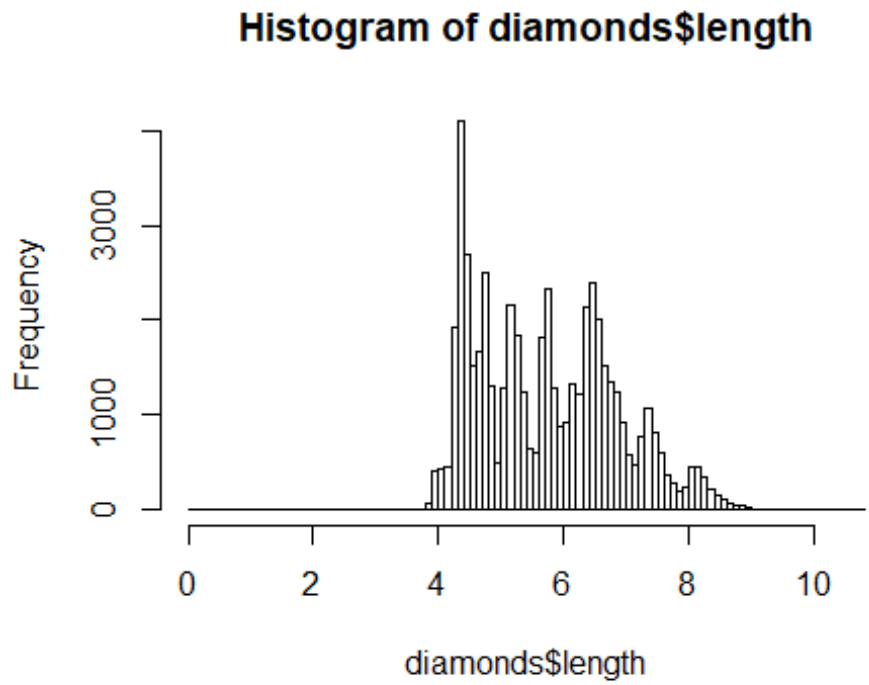

```
hist(diamonds$depth,breaks=50)
```



```
hist(diamonds$width,breaks=100)
```



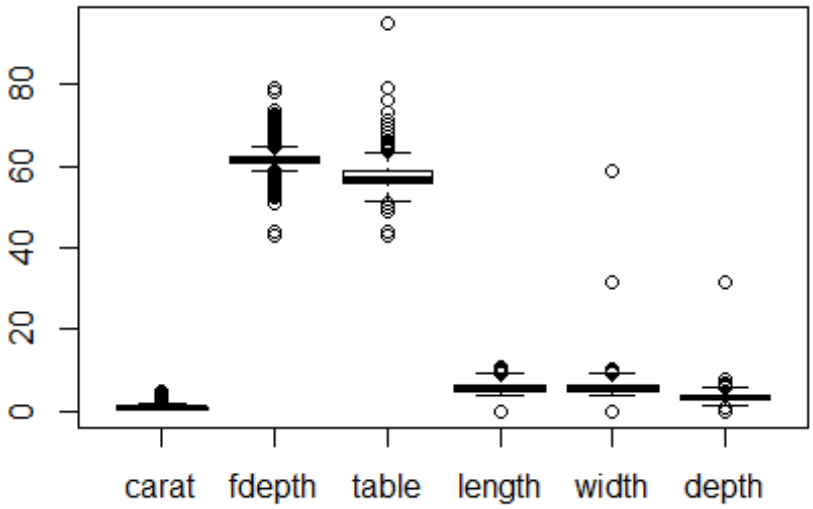

```
hist(diamonds$length,breaks=100)
```



La majorité des diamants on des poids compris entre 0.2 et 1.5 carats, ce qui correspond à des poids en gramme compris entre 0.04 et 0.3. Nous avons donc à faire dans cette base de données principalement à de petits diamants.

Les variables **table**, **fdepth**, **depth** et **width** ont des distributions gaussiennes alors que la variable **length** pas du tout. La longueur des diamants de ce jeu de données est donc très hétérogène comparé aux deux autres variables de mesure que sont **depth** et **width** .

```
boxplot(diamonds[,c(1,5,6,8,9,10)])
```

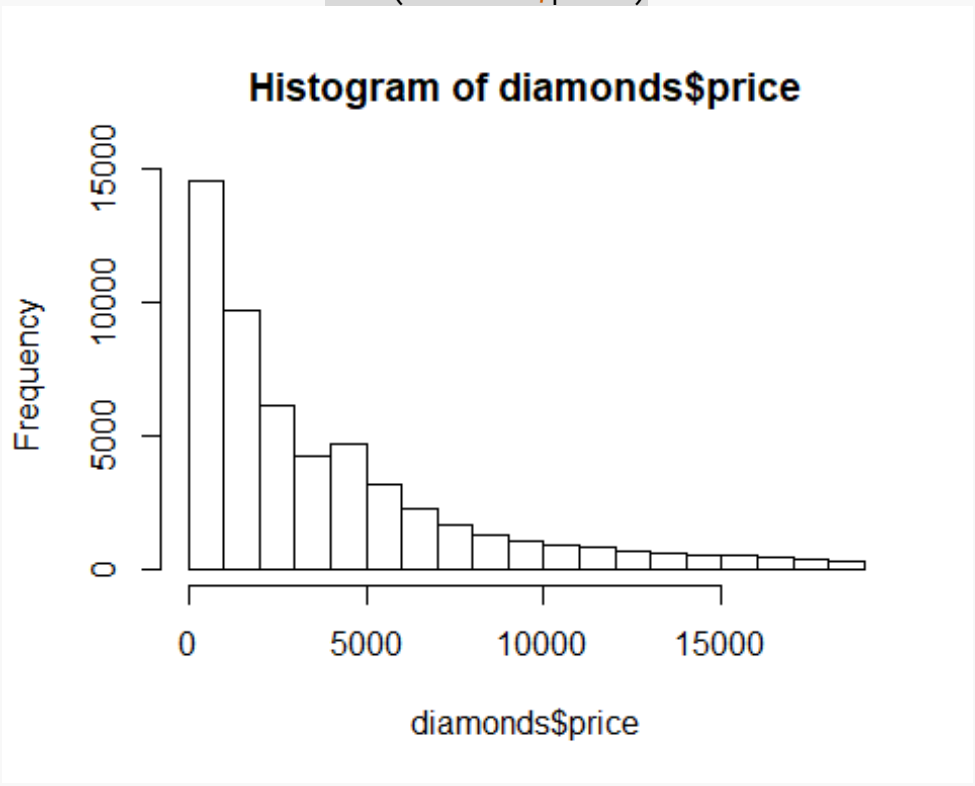


Boxplot des variables quantitatives

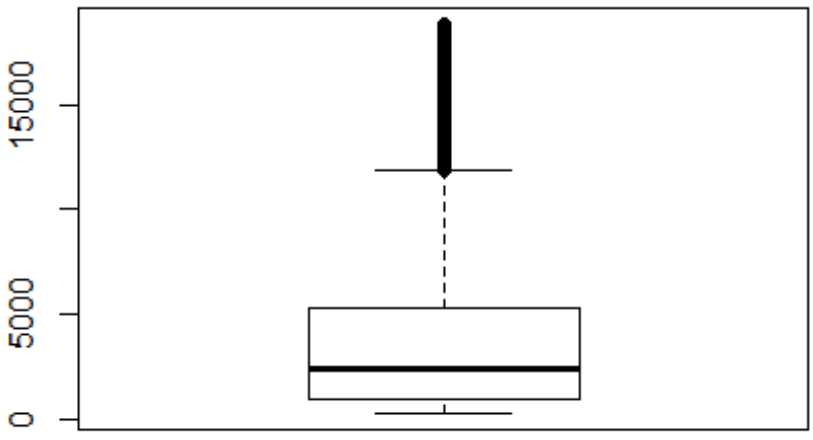
On remarque grâce à ce boxplot que les variables **fdepth** et **table** ont beaucoup de points outliers. Les variables **width** et **depth** également mais dans une moindre mesure. On veillera dans la suite à mettre ces individus outliers en supplémentaire pour les utiliser lors de l'interprétation et non lors de l'apprentissage du modèle. (Il peut s'agir d'erreurs de mesure)

c. Analyse de la variable price

```
hist(diamonds$price)
```




```
boxplot(diamonds$price)
```



Boxplot de la variable **price**

On remarque que la majorité des diamants de ce jeu de données ont des prix inférieurs à 6000 dollars (on retrouve les résultats du summary) et que ceux ayant un prix supérieur à 12000 dollars sont considérés comme outliers par le logiciel Rstudio.

Nous allons donc porter une attention particulière à ces individus pour bien comprendre quelles caractéristiques font augmenter si fortement le prix d'un diamant.

```
diamonds_filtered = filter(diamonds, price > 6000)
summary(diamonds_filtered)
```

```
##      carat               cut        color       clarity         fdepth      
##  Min.   :0.630   Fair      : 331   D: 899   VS2    :3078   Min.   :50.80  
##  1st Qu.:1.150   Good      : 943   E:1299   SI1    :2326   1st Qu.:61.00  
##  Median :1.500   Very Good:2593   F:1878   SI2    :2204   Median :61.90  
##  Mean   :1.477   Premium   :3694   G:2856   VS1    :2050   Mean   :61.71  
##  3rd Qu.:1.690   Ideal     :3990   H:2062   VVS2   :1045   3rd Qu.:62.50  
##  Max.   :5.010                    I:1633   VVS1   : 439   Max.   :70.60  
##                                   J: 924   (Other): 409                  
##      table           price           length           width       
##  Min.   :50.00   Min.   : 6001   Min.   : 0.000   Min.   : 0.000  
##  1st Qu.:56.00   1st Qu.: 7338   1st Qu.: 6.730   1st Qu.: 6.730  
##  Median :58.00   Median : 9467   Median : 7.250   Median : 7.250  
##  Mean   :57.79   Mean   :10369   Mean   : 7.245   Mean   : 7.243  
##  3rd Qu.:59.00   3rd Qu.:12822   3rd Qu.: 7.610   3rd Qu.: 7.600  
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900  
##                                                                   
##      depth      
##  Min.   :0.000  
##  1st Qu.:4.140  
##  Median :4.480  
##  Mean   :4.467  
##  3rd Qu.:4.690  
##  Max.   :8.060  
```

Les diamants ayant des prix supérieurs à 6000 dollars ont des poids supérieurs ou égaux à 0.63 carat, ce qui est à peu près équivalent à la médiane des poids du jeu de données complet. On en conclut donc que dès l'instant ou le poids d'un diamant dépasse un certain seuil (approximativement 0.6 d'après ces données), son prix a tendance à augmenter assez fortement.

Pour la variable **cut**, qui est réputée comme bon indicateur du prix d'un diamant, le changement notable est la forte augmentation du pourcentage de diamants de coupe *Premium* dans le jeu de données filtré comparé au jeu de données complet, ainsi que la diminution assez forte du pourcentage de diamant de coupe *Ideal* dans le jeu filtré comparé au jeu complet. On en retire donc que ce n'est pas forcément la coupe *Ideal* qui fera augmenter le prix d'un diamant mais en revanche la coupe *Premium* semble être beaucoup plus récurrente chez les diamants valant plus de 6000 dollars.

Pour la variable **color** qui est aussi réputée pour influencer le prix d'un diamant, on remarque une forte diminution en pourcentage des modalités *D* et *E* qui sont les 2

“meilleures” couleurs et une très forte augmentation en pourcentage des modalités *G,H,I* et *J*, qui sont les “moins bonnes” couleurs (*J* étant la pire). On se rend donc compte que ce n’est pas forcément la meilleure couleur d’un diamant qui fera systématiquement augmenter son prix. Ce qui peut nous paraitre très étrange car il est coutume de penser que plus un diamant a une belle couleur plus il sera cher. D’après ce jeu de données, ceci ne semble pas être vrai. Attention néanmoins, il se peut que beaucoup des diamants ayant une bonne couleur aient d’autres caractéristiques tirant leur prix vers le bas…

Pour la variable **clarity** qui encore une fois est considérée comme un très bon indicateur pour le prix d’un diamant, on observe une augmentation en pourcentage des modalités *SI2*, *VS1* et *VS2* qui correspondent respectivement à des diamants ayant de petites inclusions, facilement identifiables à la loupe et à des diamants ayant de très petites inclusions, difficilement identifiables à la loupe. On note une diminution en pourcentage des modalités *VVS1* et *IF*, correspondant aux deux meilleures clartés possibles. Il se peut que cela soit dû à la sous-représentation de ces modalités dans le jeu de données complet.

Les deux variables **fdepth** et **table** sont elles très peu impactées par ce filtrage, elles ne semblent pas jouer de rôle significatif dans l’augmentation du prix d’un diamant.

Pour ce qui est des variables représentant les dimensions globales d’un diamant, on remarque juste que les diamants de plus de 6000 dollars ont tendances à être un peu plus longs, larges et profonds que la moyenne de tous les diamants du jeu complet, ce qui peut s’expliquer par leur poids plus élevé en moyenne. En revanche, la profondeur maximale pour les diamants filtrés est de 8.06 alors qu’elle est de 31,8 pour le jeu complet. Il se peut donc que l’individu ayant cette mesure soit un outlier. Il semble que la valeur de la largeur maximale soit aussi une erreur.

```
suspicious1 = filter(diamonds, depth==31.800)
suspicious2 = filter(diamonds, width==58.9)
suspicious1
```

```
##   carat cut       color clarity fdepth table price length width depth
##   0.51 Very Good E     VS1       61.8  54.7  1970   5.12  5.15  31.8
```

```
suspicious2
```

```
##   carat cut     color clarity fdepth table price length width depth
##       2 Premium H     SI2       58.9    57 12210   8.09  58.9  8.06
```

Aux vues des résultats, il semble que 31.8 et 58.9 soient des erreurs de recopie et que les bonnes valeurs soient plutôt 3.18 et 5.89. Nous décidons donc de modifier ces valeurs.

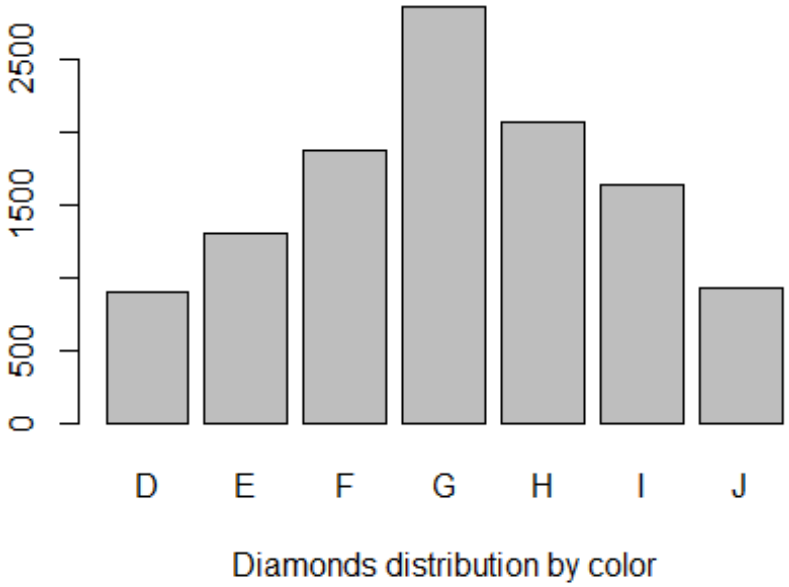
```
diamonds[48411,'depth']=3.18
diamonds[24068,'width']=5.89
diamonds[49190,'width']=3.18
summary(diamonds)
```

```
      carat               cut        color        clarity          fdepth
Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065   Min.   :43.00
1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258   1st Qu.:61.00
Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80
Mean   :0.7979   Premium   :13791   G:11292   VS1    : 8171   Mean   :61.75
3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50
Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00
                                   J: 2808   (Other): 2531
##      table           price          length           width
##  Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710
##  Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.733
##  3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :10.540
##
##      depth
##  Min.   :0.000
##  1st Qu.:2.910
##  Median :3.530
##  Mean   :3.538
##  3rd Qu.:4.040
##  Max.   :8.060
```

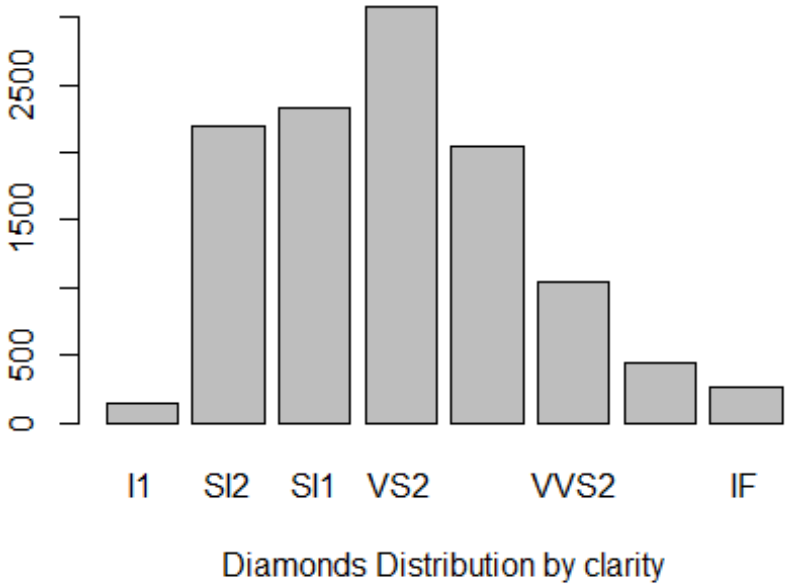
Nous allons maintenant pouvoir comparer les distributions des différentes variables, pour des prix de vente supérieurs à 6000 dollars.

```
diamonds_filtered = filter(diamonds, price > 6000)
plot(diamonds_filtered$cut, xlab="Diamonds distribution by cut")
```

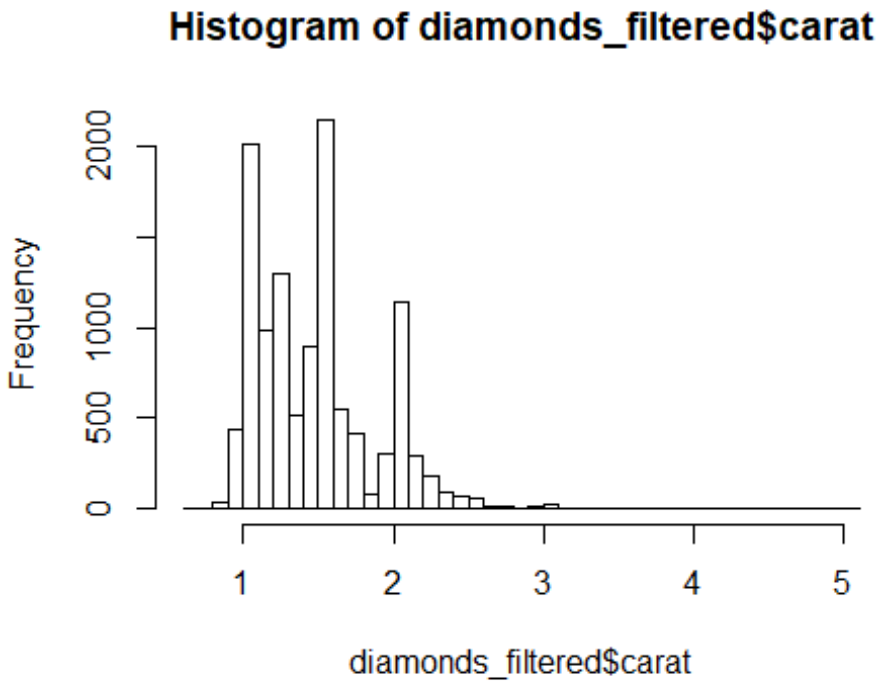
```
plot(diamonds_filtered$color, xlab="Diamonds distribution by color")
```



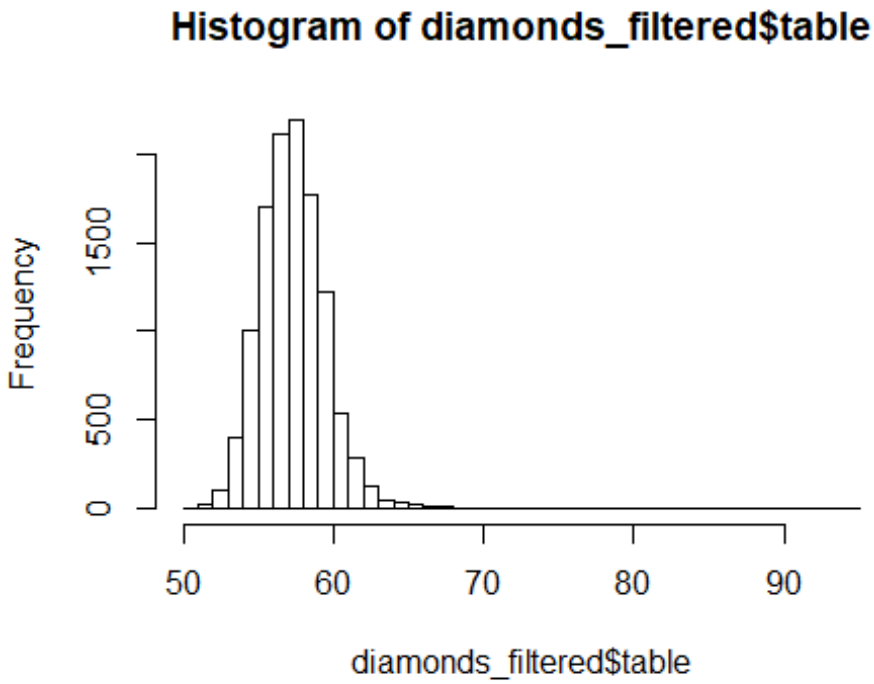

```
plot(diamonds_filtered$clarity, xlab="Diamonds Distribution by clarity")
```



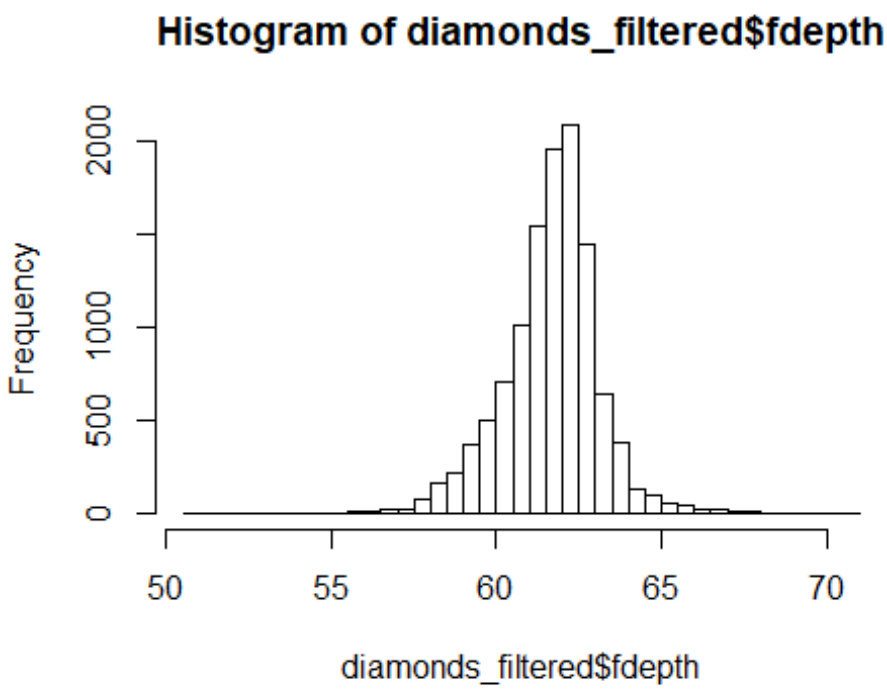
```r
hist(diamonds_filtered$carat,breaks=50)
```



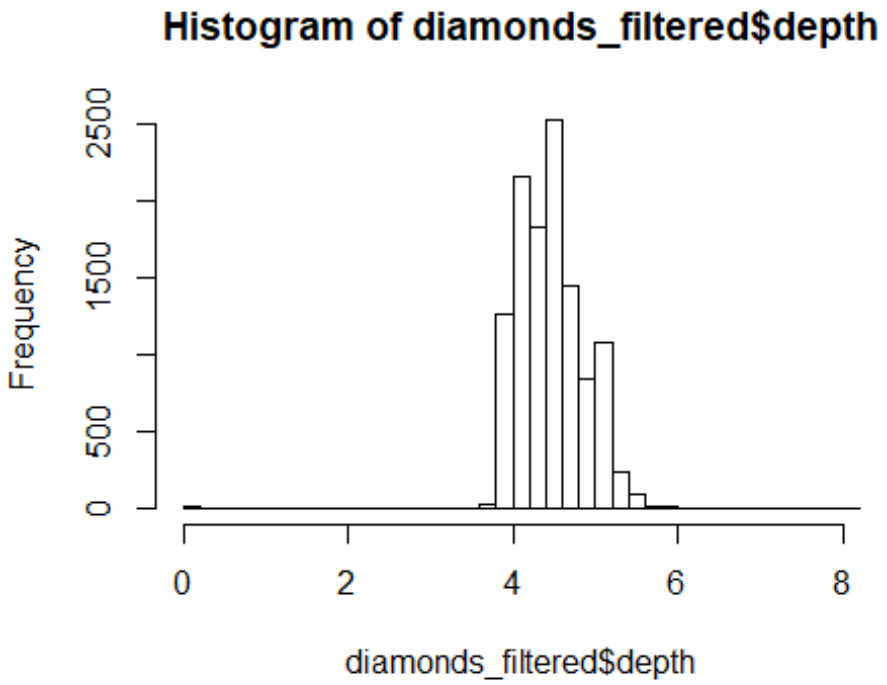

```r
hist(diamonds_filtered$table,breaks=50)
```



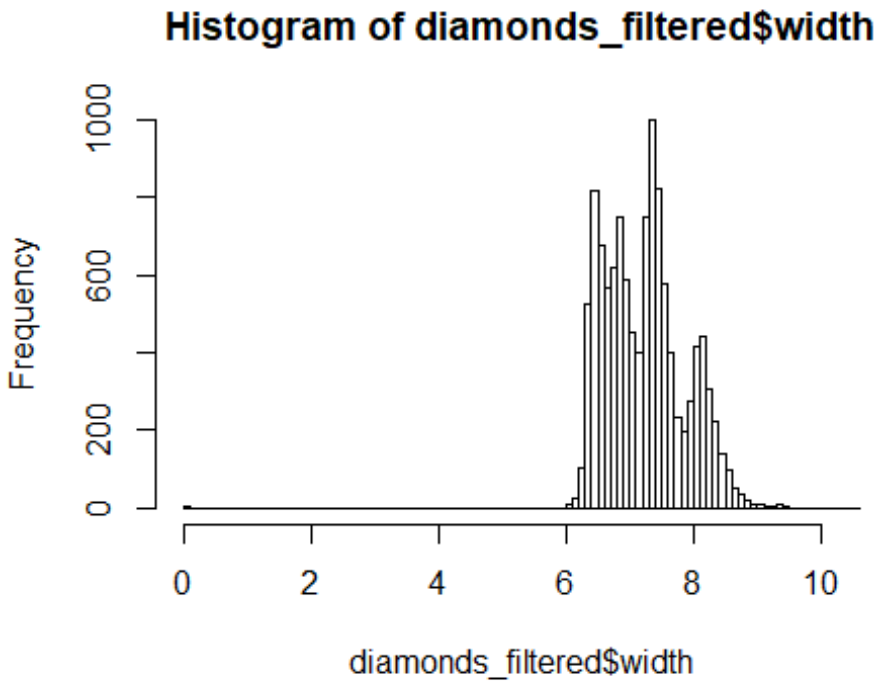
```r
hist(diamonds_filtered$fdepth,breaks=50)
```



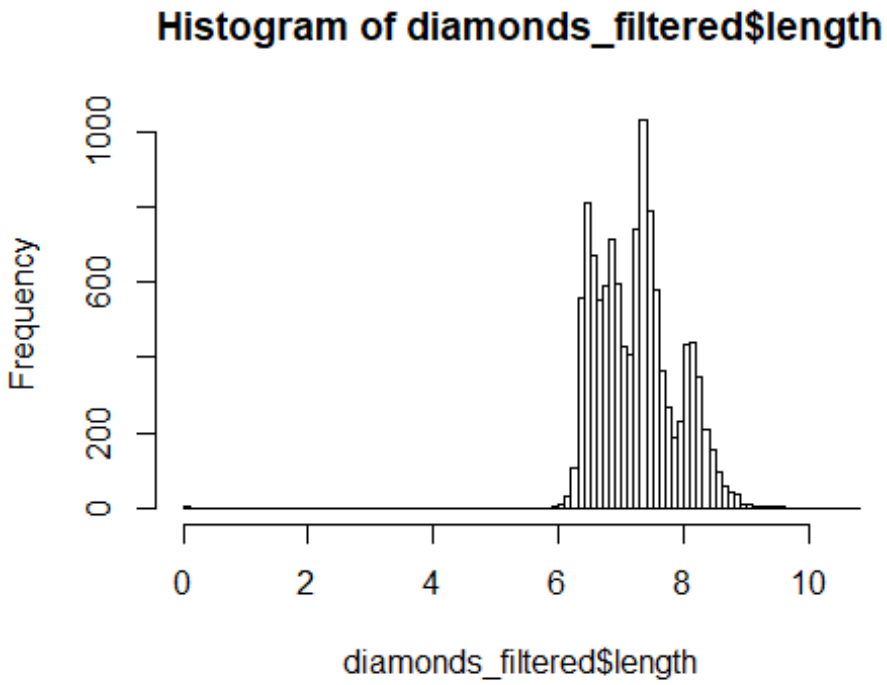

```r
hist(diamonds_filtered$depth,breaks=50)
```



```
hist(diamonds_filtered$width,breaks=100)
```




```
hist(diamonds_filtered$length,breaks=100)
```



On retrouve la même répartition des données pour la variable **cut**, on a toujours en majorité des diamants ayant une coupe de qualité supérieure ou égale à *Very Good*.

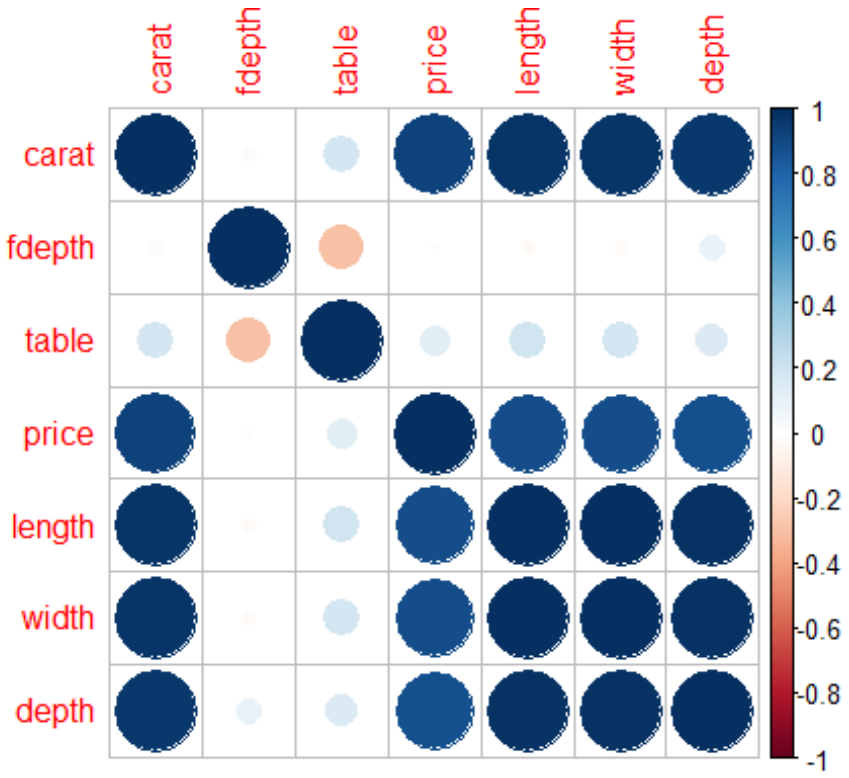
On retrouve également une répartition similaire pour la variable **clarity**. En revanche, la répartition de la variable **color** est encore plus gaussienne sur l'échantillon filtré que sur l'échantillon complet, avec la modalité *G* la plus représentée.

En ce qui concerne les variables qualitatives, les histogrammes sont légèrement décalés vers la droite par rapport à ceux du jeu complet, ce qui indique une moyenne plus élevée, comme il a déjà été mentionné auparavant.

d. Etude des corrélations

Les résultats que nous avons observés jusque-là sont des résultats marginaux, il faut prendre en compte, si elle existe, la corrélation entre les variables explicatives, qui peut fausser les interprétations des résultats.

```
corrplot::corrplot(cor(diamonds[,-c(2,3,4)]))
```



Graphique des corrélations

On remarque grâce à ce graphique des corrélations que les variables **length**, **width**, **depth** et **carat** sont très corrélées positivement et qu'elles sont également toutes très corrélées positivement à la variable à expliquer **price**. Le prix d'un diamant semble donc à première vue être déterminé par ses dimensions et son poids (en carat), en excluant les variables qualitatives de l'étude.

De plus, on décèle une légère corrélation négative entre les variables **table** et **fdepth**.

Etude des corrélations entre variables qualitatives:

```
tab_cont1=table(diamonds$cut, diamonds$color)
chisq.test(tab_cont1)

##
##  Pearson's Chi-squared test
##
## data:  tab_cont1
## X-squared = 310.32, df = 24, p-value < 2.2e-16
```

Au vu de la très faible p-value du test du khi-deux d'indépendance entre la variable **cut** et la variable **color**, on en déduit que ce deux variables ne sont pas indépendantes (on rejette Ho)

```
tab_cont2=table(diamonds$cut, diamonds$clarity)
chisq.test(tab_cont2)

##
##  Pearson's Chi-squared test
##
## data:  tab_cont2
## X-squared = 4391.4, df = 28, p-value < 2.2e-16
```

On obtient le même résultat ici, les variables **cut** et **clarity** ne sont pas indépendantes.

```
tab_cont3=table(diamonds$color, diamonds$clarity)
chisq.test(tab_cont3)

##
##  Pearson's Chi-squared test
##
## data:  tab_cont3
## X-squared = 2047.1, df = 42, p-value < 2.2e-16
```

Les variables **color** et **clarity** ne sont pas non plus indépendantes.

Remarque : Nous venons de montrer que les 3 variables qualitatives sont non-indépendantes et qu'il y a beaucoup de corrélations entre les variables quantitatives explicatives. Cela nous posera sûrement des problèmes de modélisation et d'interprétation des résultats lors de nos différentes régressions futures. Il nous faudra donc certainement trouver un modèle pouvant s'affranchir de ces problèmes de multicolinéarité.

e. Conclusion

Nous avons vu grâce à cette analyse descriptive des données que les différentes modalités des variables qualitatives ne sont pas toute également réparties. La coupe prédominante des diamants de ce jeu de données est la coupe *Ideal* suivie de la coupe *Premium* et enfin de la coupe *Very Good*. Les diamants de très bonne coupe sont donc très fortement représentés.
Pour ce qui est de la clarté, ce sont les modalités intermédiaires qui sont les plus représentées. Les clartés et les moins bonnes et les meilleures sont peu représentées.
En ce qui concerne la couleur des diamants, ce sont encore une fois les modalités intermédiaires les plus représentées suivies des meilleures puis des moins bonnes. A ce stade, on pourrait faire l'hypothèse (au vu de ces dispersions) que les variables **color** et **clarity** sont celles qui vont influencer le plus le prix d'un diamant car peu de diamants détiennent les meilleures caractéristiques pour ces variables. Donc lorsqu'un diamant possèdera ces caractéristiques rares, on peut s'attendre à une augmentation de son prix.

C'est pourquoi nous avons essayé de trouver les caractéristiques les plus communes aux diamants coûtant plus de 6000 dollars. Ce qu'il en ressort ne confirme pas forcément ce qui a été dit précédemment. Les diamants les plus cher ont principalement une coupe *Premium* mais pas forcément des couleurs rares (les diamants semblent d'ailleurs posséder les "moins bonnes" couleurs) et il en est de même pour la clarté : les modalités intermédiaires sont plus présentent que les meilleurs ou les moins bonnes. Attention toutefois ici, car cela peut-être dû à la forte sous représentations des modalités *I1* et *IF* pour cette variable.

La variable **cut** semble donc jouer un rôle plus important dans l'augmentation du prix d'un diamant que les 2 autres variables **clarity** et **color**. Il nous reste à voir maintenant ce que nous obtenons lorsque que nous étudions le jeu de données à l'aide de toutes les variables à la fois.

IV. Analyse Factorielle multiple

Ici il s'agit ici de considérer le jeu de données dans sa globalité et non en séparant données catégorielles et quantitatives. L'Analyse Factorielle Multiple (AFM) est une méthode d'analyse de données portant sur des variables (quantitatives et/ou qualitatives) structurées en groupes. Ces groupes sont dits homogènes, dans le sens où ils ne contiennent que des variables quantitatives, ou que des variables qualitatives (ils ne peuvent pas contenir les deux à la fois).

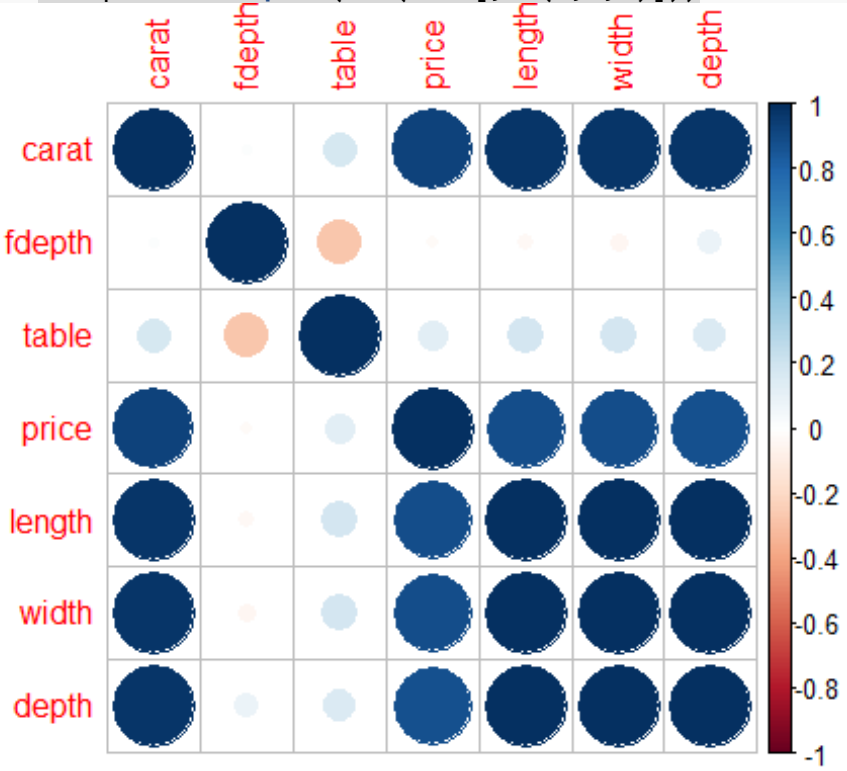
La première étape est de créer un échantillon du jeu de données complet pour rendre les temps de calculs plus rapides et les graphiques plus lisibles.

a. Echantillonnage

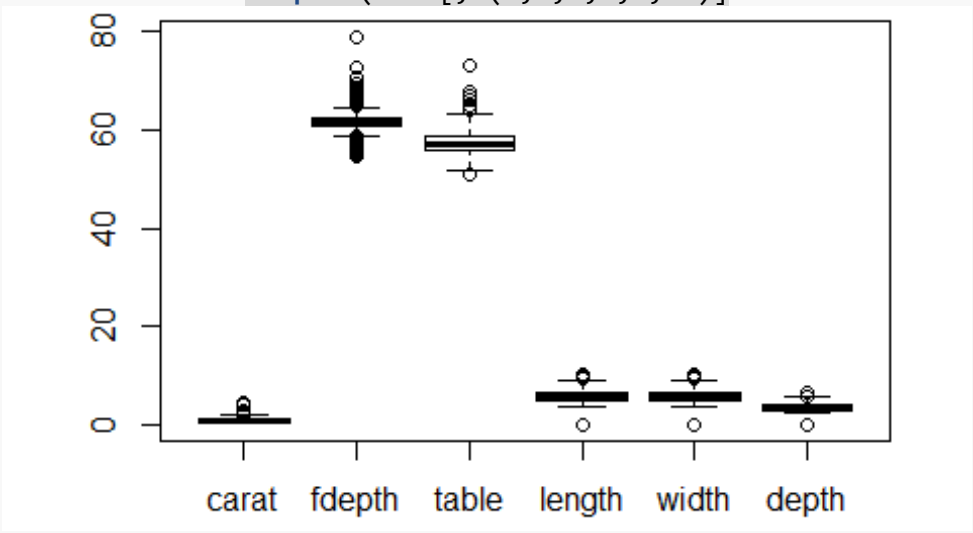
```
n=dim.data.frame(diamonds)[1]
Ind=seq.int(1,n,10)
data=diamonds[Ind,]           #On a donc un échantillon de 5394 observations,
en prenant un individus tous les 10 individus
```

```
## # A tibble: 5,394 x 10
##    carat cut       color clarity fdepth table price length width depth
##    <dbl> <ord>     <ord> <ord>    <dbl> <dbl> <int>  <dbl> <dbl> <dbl>
##  1  0.23 Ideal     E     SI2       61.5    55   326   3.95  3.98  2.43
##  2  0.3  Good      J     SI1       64      55   339   4.25  4.28  2.73
##  3  0.3  Good      I     SI2       63.3    56   351   4.26  4.3   2.71
##  4  0.23 Very Good F     VS1       60      57   402   4     4.03  2.41
##  5  0.33 Ideal     I     SI2       61.2    56   403   4.49  4.5   2.75
##  6  0.24 Very Good F     SI1       60.9    61   404   4.02  4.03  2.45
##  7  0.35 Ideal     I     VS1       60.9    57   552   4.54  4.59  2.78
##  8  0.24 Very Good D     VVS1      61.5    60   553   3.97  4     2.45
##  9  0.26 Very Good E     VVS1      63.4    59   554   4     4.04  2.55
## 10  0.7  Ideal     E     SI1       62.5    57  2757   5.7   5.72  3.57
## # ... with 5,384 more rows
```

```
corrplot::corrplot(cor(data[,-c(2,3,4)]))
```



```
boxplot(data[,c(1,5,6,8,9,10)]
```



Boxplot des variables quantitatives

On retrouve les mêmes corrélations entre variables quantitatives que sur le jeu de données complet à l'exception peut-être d'une corrélation plus faible entre les variables **depth**, **length**, **width** et **price**. Il semble également y avoir beaucoup moins de points aberrants.

b. Analyse factorielle multiple

Pour effectuer une analyse factorielle multiple, les caractéristiques des diamants de notre jeu de données doivent d'abord être regroupées selon ce qu'elles représentent.

```
#create groups for characterizing each aspect of diamond

group_weight <- c("carat")

group_size <- c("length",
                "width",
                "depth"
)
group_physical_asp <- c("cut",
                        "color",
                        "clarity"
)
group_size_frequency <- c("fdepth",
                          "table"
)
```

D'un côté nous avons des mesures des diamants à regrouper ensemble (même unité), d'un autre côté leur poids. Nous avons aussi regroupés les caractéristiques qualitatives (aspects visuels) ainsi que les fréquences et pourcentages.

Ceci étant fait, nous pouvons commencer l'application de la méthode d'AFM.

```r
#change order of column by groups

data <- data[,c(group_weight,
                group_size,
                group_physical_asp,
                group_size_frequency)]

#MFa analysis
data_mfa <- MFA(data,
                        group      = c(1,3,3,2),
                        type       = c("c","s","n","s"),
                        name.group = c("weight",
                                       "size",
                                       "physical_aspects",
                                       "size_frequencies"
                        ),
                        graph=FALSE
)
summary(data_mfa)
## Call:
## MFA(base = data, group = c(1, 3, 3, 2), type = c("c", "s", "n",
##       "s"), name.group = c("weight", "size", "physical_aspects",
##       "size_frequencies"), graph = FALSE)
##
##
Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
Variance               2.347   1.251   1.085   0.888   0.843   0.835   0.817
% of var.             13.857   7.386   6.409   5.245   4.977   4.931   4.824
Cumulative % of var.  13.857  21.243  27.652  32.897  37.874  42.805  47.628
                       Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14
Variance               0.811   0.811   0.799   0.786   0.759   0.752   0.743
% of var.              4.789   4.787   4.720   4.641   4.483   4.441   4.389
Cumulative % of var.  52.417  57.204  61.924  66.565  71.048  75.489  79.878
                      Dim.15  Dim.16  Dim.17  Dim.18  Dim.19  Dim.20  Dim.21
Variance               0.722   0.690   0.679   0.558   0.477   0.255   0.025
% of var.              4.265   4.073   4.010   3.294   2.820   1.508   0.150
Cumulative % of var.  84.142  88.215  92.225  95.520  98.339  99.848  99.998
                      Dim.22  Dim.23  Dim.24  Dim.25  Dim.26
Variance               0.000   0.000   0.000   0.000   0.000
% of var.              0.002   0.001   0.000   0.000   0.000
Cumulative % of var.  99.999 100.000 100.000 100.000 100.000



Groups
                    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
weight          |  0.894 38.119  0.800 |  0.021  1.666  0.000 |  0.028  2.597
size            |  0.894 38.109  0.800 |  0.019  1.551  0.000 |  0.030  2.738
physical_aspects|  0.457 19.455  0.020 |  0.461 36.860  0.020 |  0.741 68.239
```

```
size_frequencies|  0.101  4.316  0.008 |  0.750 59.924  0.426 |  0.287 26.425
                   cos2
weight            0.001 |
size              0.001 |
physical_aspects  0.051 |
size_frequencies  0.062 |
##
Individuals (the 10 first)
                   Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
1                | -1.918  0.029  0.273 |  0.345  0.002  0.009 | -0.434  0.003
2                | -0.936  0.007  0.029 |  1.173  0.020  0.046 |  1.467  0.037
3                | -0.697  0.004  0.021 |  0.686  0.007  0.020 |  1.017  0.018
4                | -1.845  0.027  0.218 | -1.134  0.019  0.082 |  0.208  0.001
5                | -0.926  0.007  0.058 |  0.435  0.003  0.013 | -0.900  0.014
6                | -1.343  0.014  0.122 | -1.583  0.037  0.169 |  1.502  0.039
7                | -1.248  0.012  0.104 |  0.075  0.000  0.000 | -0.921  0.014
8                | -2.189  0.038  0.194 | -1.234  0.023  0.062 |  1.087  0.020
9                | -2.185  0.038  0.211 | -0.461  0.003  0.009 |  1.236  0.026
10               | -0.487  0.002  0.032 |  0.625  0.006  0.052 |  0.087  0.000
##                    cos2
## 1                 0.014 |
## 2                 0.072 |
## 3                 0.044 |
## 4                 0.003 |
## 5                 0.055 |
## 6                 0.152 |
## 7                 0.057 |
## 8                 0.048 |
## 9                 0.068 |
## 10                0.001 |
##
## Continuous variables
                   Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
carat            |  0.449 38.119  0.894 |  0.069  1.666  0.021 | -0.080  2.597
length           |  0.947 12.781  0.896 |  0.102  0.280  0.010 | -0.184  1.048
width            |  0.943 12.680  0.889 |  0.105  0.294  0.011 | -0.192  1.131
depth            |  0.942 12.648  0.887 |  0.191  0.977  0.037 | -0.135  0.559
fdepth           | -0.008  0.002  0.000 |  0.708 31.382  0.502 |  0.426 13.074
table            |  0.360  4.315  0.129 | -0.676 28.541  0.456 |  0.430 13.351
##                    cos2
## carat             0.028 |
## length            0.034 |
## width             0.037 |
## depth             0.018 |
## fdepth            0.181 |
## table             0.185 |
##
## Categories (the 10 first)
                     Dim.1     ctr    cos2  v.test     Dim.2     ctr    cos2
Fair               |   1.133   0.554   0.043   9.590 |   2.508   9.553   0.212
```

```
Good              |   0.427   0.214   0.020   6.120 |  -0.051   0.011   0.000
Very Good         |   0.050   0.008   0.001   1.312 |  -0.272   0.851   0.027
Premium           |   0.676   1.734   0.172  19.490 |  -1.018  13.826   0.389
Ideal             |  -0.662   2.462   0.258 -25.552 |   0.664   8.709   0.259
D                 |  -0.618   0.669   0.065 -11.082 |  -0.095   0.055   0.002
E                 |  -0.543   0.773   0.080 -12.342 |  -0.278   0.712   0.021
F                 |  -0.195   0.097   0.010  -4.356 |  -0.101   0.092   0.003
G                 |  -0.201   0.121   0.013  -4.970 |   0.127   0.170   0.005
H                 |   0.532   0.600   0.060  10.646 |   0.120   0.107   0.003
##                   v.test     Dim.3     ctr    cos2  v.test
## Fair              29.071 |   3.527  25.101   0.419  43.895 |
## Good              -1.009 |   1.194   7.816   0.155  25.177 |
## Very Good         -9.732 |   0.381   2.223   0.054  14.652 |
## Premium          -40.187 |   0.021   0.008   0.000   0.898 |
## Ideal             35.085 |  -0.757  15.047   0.337 -42.958 |
## D                 -2.326 |   0.421   1.452   0.030  11.101 |
## E                 -8.650 |   0.272   0.905   0.020   9.084 |
## F                 -3.096 |   0.195   0.453   0.010   6.408 |
## G                  4.300 |  -0.261   0.952   0.022  -9.467 |
## H                  3.290 |  -0.210   0.439   0.009  -6.195 |
```

Cette sortie nous donne beaucoup d'informations diverses sur les contributions des différentes variables aux axes, ainsi que la contribution de chaque groupe de variables créés. Nous les exploiterons plus en détails dans la suite.

Nous pouvons néanmoins d'ores et déjà remarquer que si l'on ne retient que 2 composantes principales, nous n'expliquons que 21% de la variance. Pour avoir un pourcentage de 50% il faudrait retenir au minimum 8 composantes.

Pour des raisons pratiques nous allons nous limiter dans cette étude à deux composantes principales.

Regardons de plus près comment ces deux axes sont constitués, du point de vue des groupes, puis des variables et enfin des individus.

i. Information des groups

```r
groups_mfa_results <- get_mfa_var(data_mfa, "group")

groups_mfa_results$cos2
```

```
##                        Dim.1        Dim.2        Dim.3        Dim.4
## weight           0.800082870 0.0004339829 0.0007946055 3.792526e-09
## size             0.799639512 0.0003762359 0.0008831805 4.950656e-07
## physical_aspects 0.019565566 0.0199537872 0.0514878185 7.134782e-02
## size_frequencies 0.007782834 0.4261636302 0.0623903912 1.863646e-04
##                        Dim.5
## weight           5.436709e-06
## size             9.875899e-06
```

```
## physical_aspects 6.574343e-02
## size_frequencies 1.746916e-07
```

Les groupes les mieux représentés dans le plan F1/F2 sont les groupes “weight”, “size” et “size_frequencies” qui regroupent toutes les variables quantitatives.

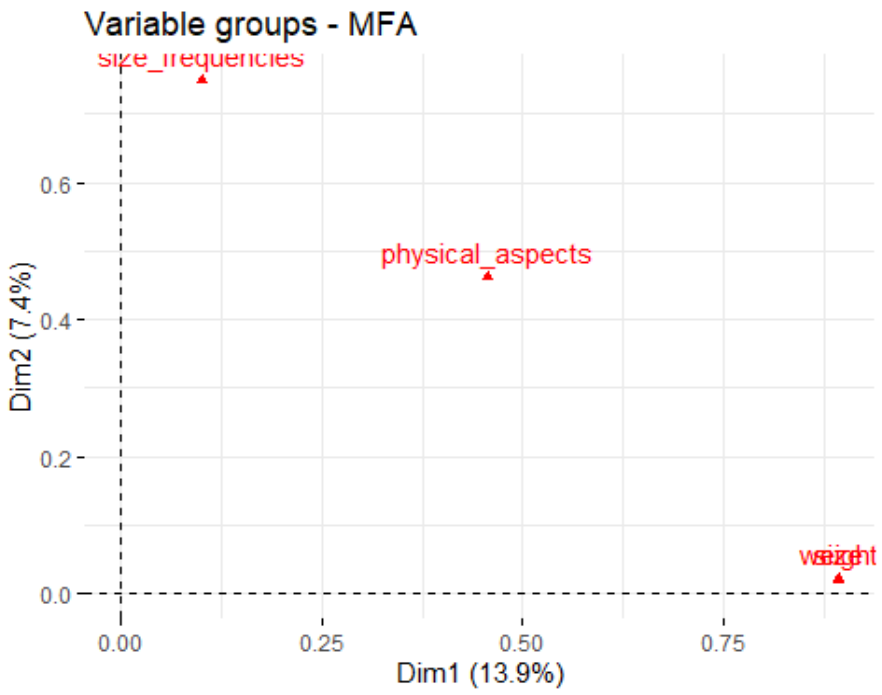
```
groups_mfa_results$contrib
```

```
##                     Dim.1     Dim.2     Dim.3        Dim.4       Dim.5
## weight           38.11925  1.665567  2.597427  0.006933386  0.27666060
## size             38.10893  1.550810  2.738388  0.079216426  0.37288107
## physical_aspects 19.45533 36.859881 68.239408 98.149243942 99.29352058
## size_frequencies  4.31649 59.923742 26.424777  1.764606246  0.05693775
```

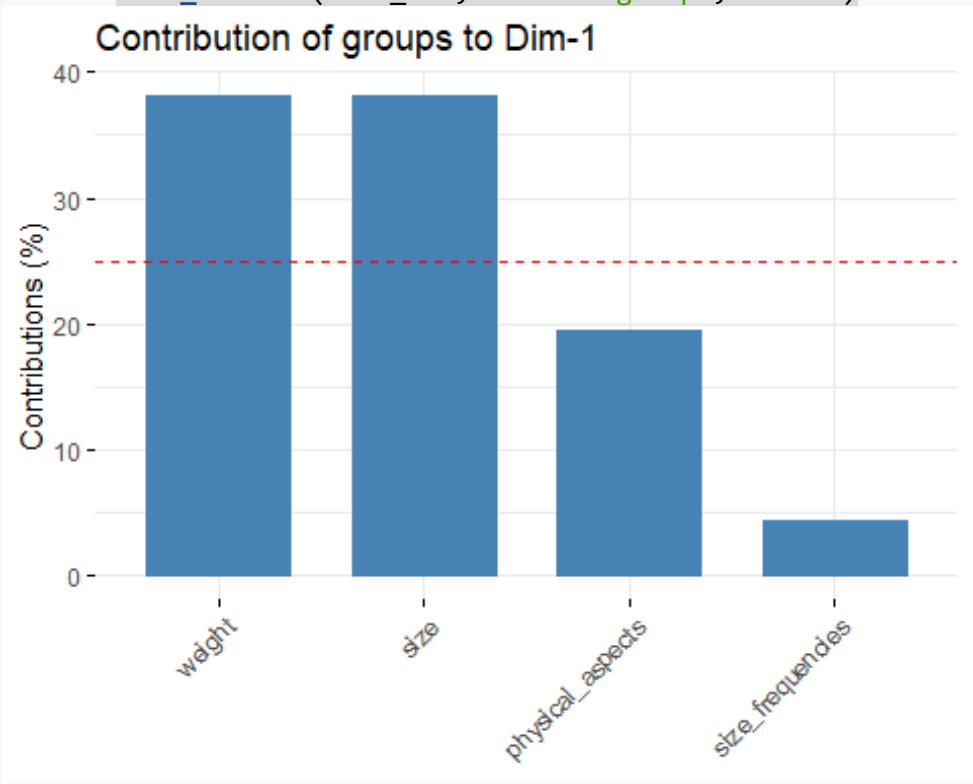
Le tableau des contributions montre que l’axe 1 est majoritairement défini par les groupes “weight” et “size” (contributions quasi égales) suivi par le groupe “physical_aspects”. Le deuxième est quant à lui principalement défini par les groupes “size_frequencies” et “physical_aspects”.

Essayons de visualiser graphiquement ces informations :

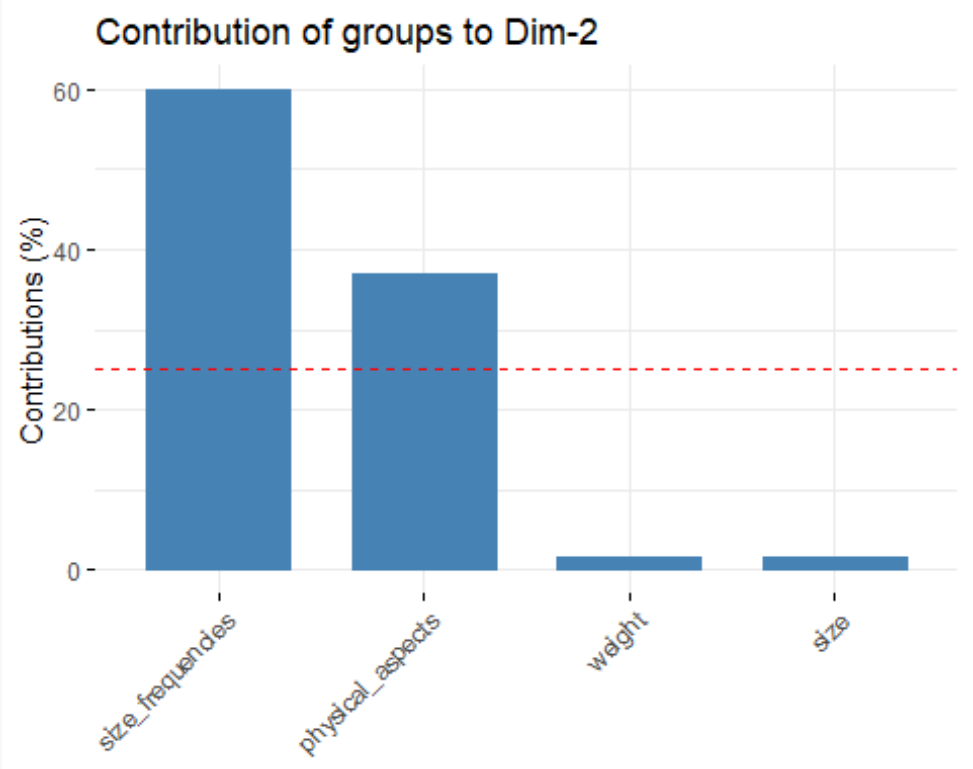
```
fviz_mfa_var(data_mfa, "group")
```



```
fviz_contrib(data_mfa, choice="group", axes=1)
```




```
fviz_contrib(data_mfa, choice="group", axes=2)
```

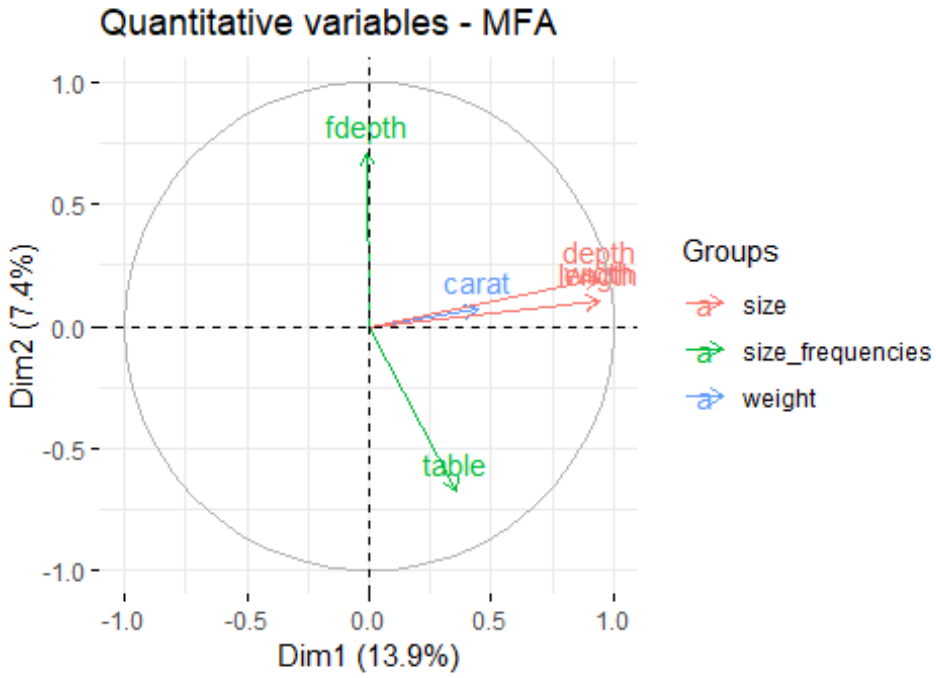


On retrouve bien dans ce graphique tout ce qui a été mentionné précédemment. Les groupes “weight” et “size” sont quasiment confondus sur le plan F1/F2. Cela signifie que les 2 groupes sont très fortement positivement corrélés. (Ce qui semble logique car on a déjà mis en lumière une forte corrélation positive entre les variables définissant ces 2 groupes). Le groupe “physical_aspects” est quant à lui plus ou moins corrélés aux trois autres groupes. En revanche, le groupe “size_frequencies” semble totalement décorrélé des groupes “weigth” et “size”.

Il en ressort que les variables qualitatives sont légèrement corrélées au différentes variables quantitatives, ce qui ne va pas nous arranger car les variables quantitatives sont déjà très corrélées entre elles.

ii. Informations des variables quantitatives

```
fviz_mfa_var(data_mfa)
```



Ce graphique des variables quantitatives nous donne le cercle des corrélations des variables, celles-ci étant colorées en fonction de leur groupe d’appartenance.

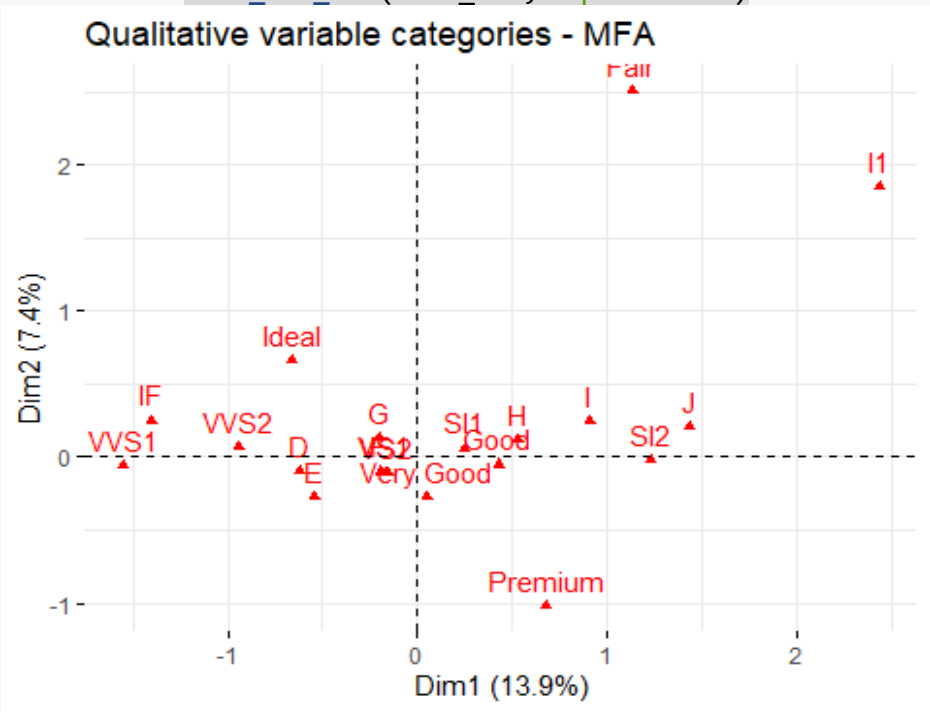
On voit de suite que les 3 variables du groupe “size” sont toutes positivement très corrélées entre elles et très corrélées à l’axe 1.

La variable carat semble elle être moyennement bien représenté sur le plan F1/F2, nous ne pouvons donc pas dire grand-chose sur elle ici.

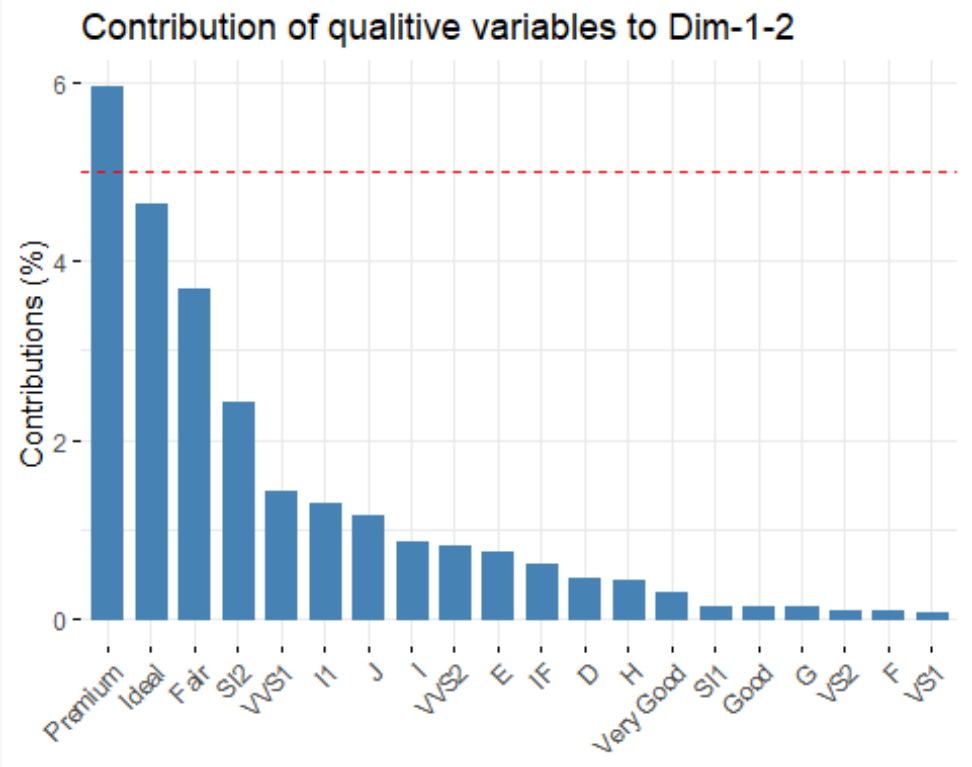
Les 2 variables du groupe “size_frequencies” sont très corrélées à l’axe 2.

iii. Information des variables qualitatives

```
fviz_mfa_var(data_mfa, 'quali.var')
```




```
fviz_contrib(data_mfa, choice="quali.var", axes=1:2)
```



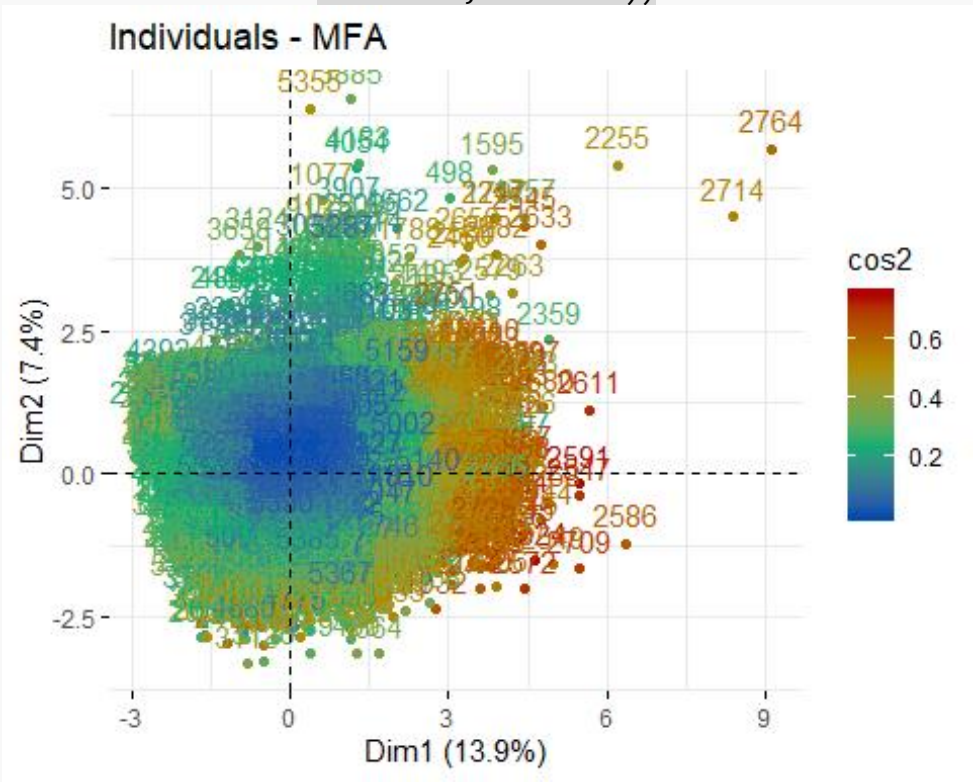
Beaucoup de modalités sont proches du centre du graphique, ce qui veut dire qu'elles n'ont pas d'influence particulière. Les modalités se détachant le plus sont pour la variable **cut** : *Fair, Ideal* et *Premium*, pour la variable **color** : *D,E,I,J* et pour la variable **clarity** : *IF, VVS1, VVS2, SI2* et *I1*.

Les modalités *VVS1* et *IF* semble assez fortement associées, comme les modalités *D* et *E*. Les modalités *I,J* et *SI2* sont également assez fortement associées.

En termes de contribution, les différentes modalités des variables qualitatives contribuent très peu au plan F1/F2. Hormis la variable **cut** qui a tendance à contribuer le plus à travers ses modalités *Premium, Ideal* et *Fair*.

iv. Information des individus

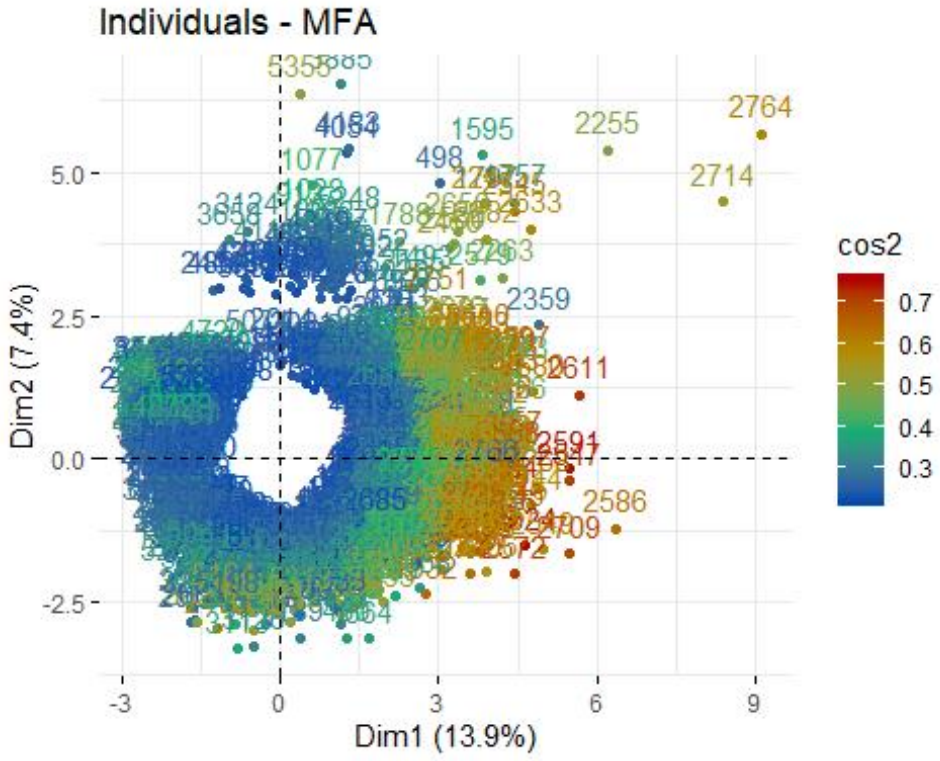
```
fviz_mfa_ind(data_mfa, col.ind = "cos2", gradient.cols=c("#024cb2","#02b26f",
                                "#b29002","#b20202"))
```



Les individus se trouvant aux alentours du centre du graphique semblent mal représentés sur le plan F1/F2. En revanche ceux étant plus éloignés sont mieux représentés. Ils le sont encore plus quand ils sont du côté positif du premier axe.

Essayons de visualiser uniquement les individus ayant un cos2 supérieur au pourcentage de variance expliquée par les 2 axes:

```
fviz_mfa_ind(data_mfa, col.ind = "cos2",
gradient.cols=c("#024cb2","#02b26f","#b29002","#b20202"), select.ind =
list(cos2=0.22))
```



On voit bien que les individus se trouvant au centre du graphique ont disparus. La répartition des individus est assez circulaire dans ce plan. On retrouve légèrement la séparation en 4 groupes observée par l'ACP, bien que les limites soient ici plus floues, sûrement à cause des variables qualitatives.

c. Conclusion

On retrouve ici les résultats de l'analyse exploratoire des données concernant les variables quantitatives (corrélation, définition des axes, etc…). On apprend en revanche que les variables qualitatives sont corrélées aux variables quantitatives bien que peu de modalités aient une réelle apparente influence sur la définition des axes et donc dans l'explication du prix d'un diamant.

Le faible pourcentage de variance expliquée par 2 axes et le nombre de modalités totales des variables qualitatives posent des problèmes de clarté des résultats et de justesse des interprétations. C'est pourquoi nous décidons à partir de ce moment de ne plus prendre en compte les variables qualitatives dans notre modèle. Nous n'utilisons donc plus que les variables quantitatives pour expliquer la variable **price** dans la suite de notre étude.

V. Analyse en Composantes Principales

Procédons maintenant à une analyse en composantes principales. L'Analyse en Composantes Principales (ACP) est une méthode d'analyse factorielle applicable uniquement sur des variables quantitatives. Elle consiste à résumer l'information contenue dans le jeu de données afin de faciliter l'interprétation des corrélations existantes entre les variables.

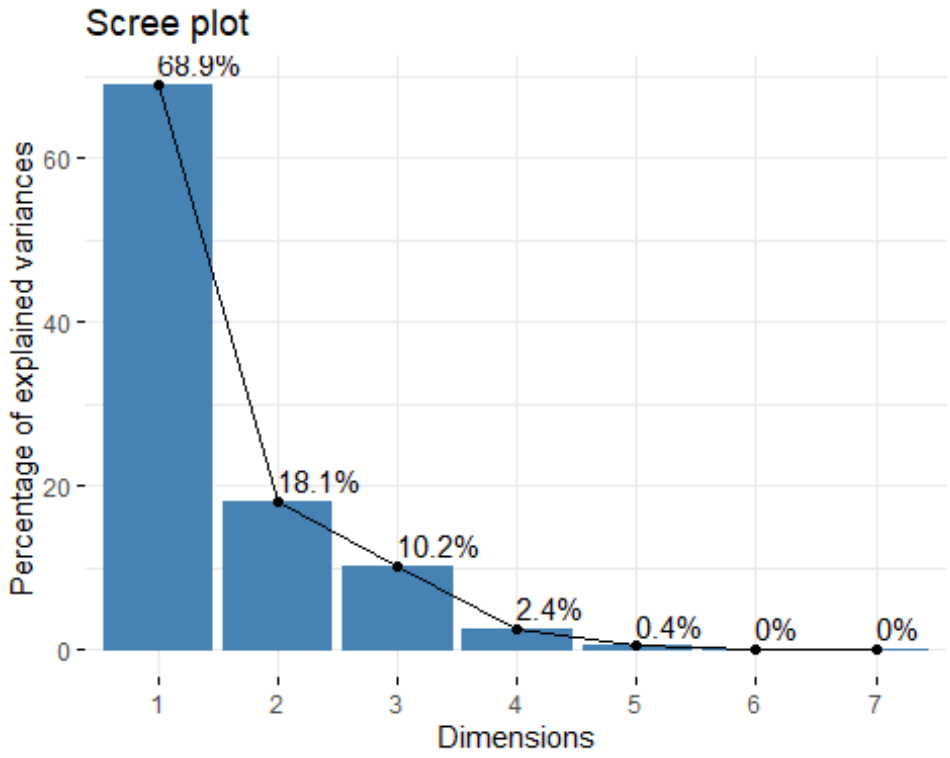
La commande permettant de réaliser l'ACP est la suivante:

```
data_acp=PCA(data,quali.sup=c(2,3,4),graph = FALSE)
```

a. Choix des axes

Regardons en fonction du pourcentage d'inertie projetée sur chaque axe combien de dimensions il faudrait garder dans la suite de l'étude.

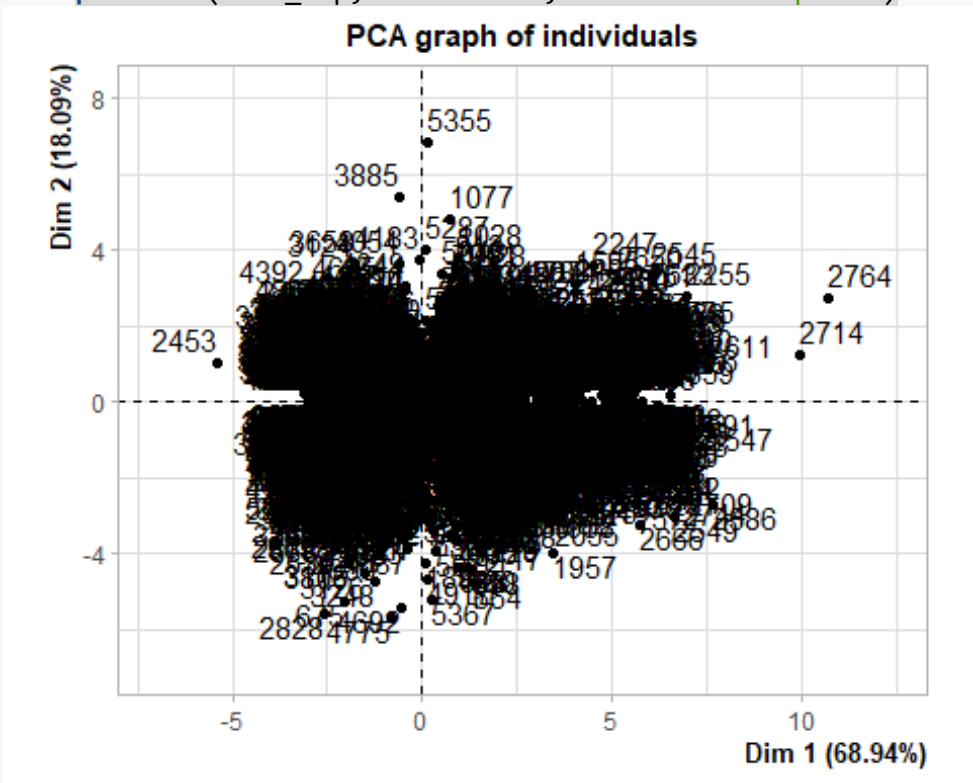
```
fviz_eig(data_acp, addlabels = TRUE)
```



On voit très clairement qu'une seule dimension explique déjà presque 70% de l'information contenue dans le jeu de données. On prend donc très peu de risque en ne retenant qu'un axe et surtout on diminue grandement le nombre de variables explicatives. Nous préfèreront retenir 2 axes (87% de variance expliquée), ce qui facilite la représentation graphique.

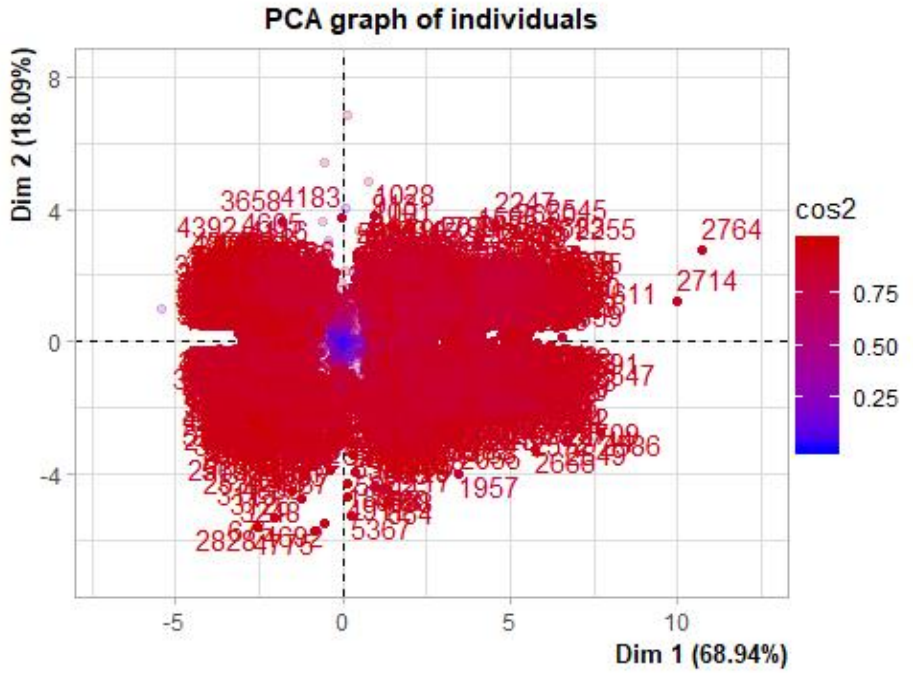
b. Etude des individus

```
plot.PCA(data_acp,choix="ind", invisible = 'quali')
```



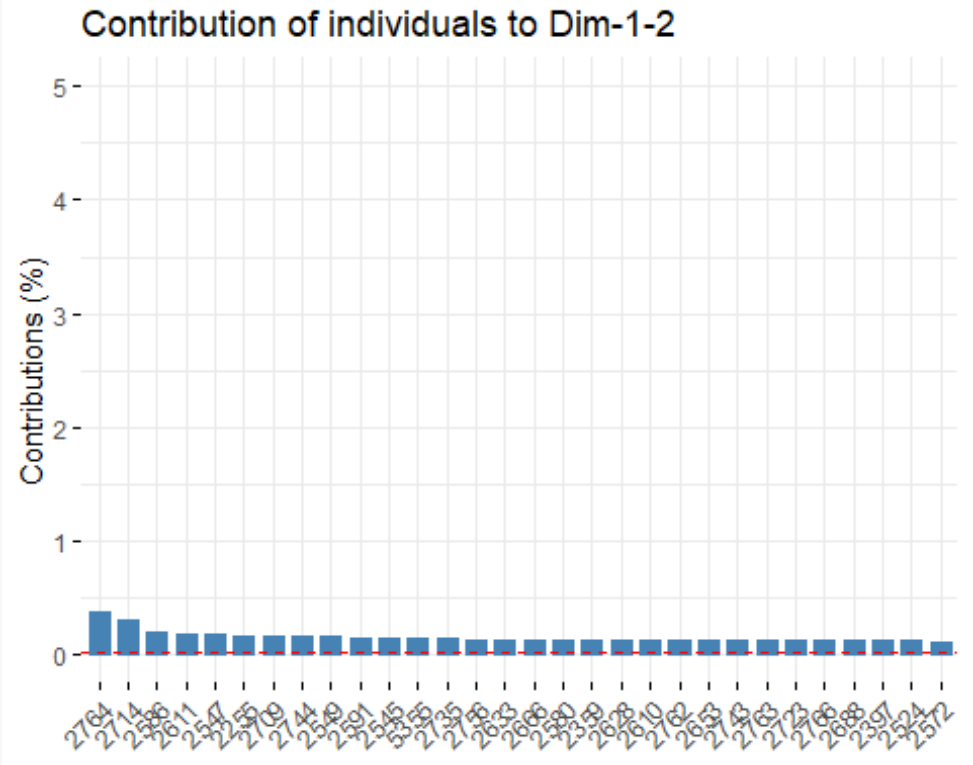
On peut voir sur le graphe des individus qu'ils sont pour la quasi-totalité tous regroupés à un même endroit du schéma. Le nuage semble également se diviser en quatre groupes, un groupe dans chaque cadran du repère. On tentera de voir à quoi cela peut correspondre avec l'étude des variables. On remarque néanmoins qu'une dizaine d'individus se distinguent des autres. Nous allons tout d'abord vérifier si ces points ne seraient pas des outliers (mal représentés dans le plan à cause de la projection sur ces axes).
Comme la part d'inertie expliquée par le plan F1/F2 est de 87%, nous allons reconstruire le nuage des individus en ne gardant que ceux dont le cos2 est supérieur à 0.8.

```
plot.PCA(data_acp,choix="ind",select="cos2 0.8",unselect=0.8,
         habillage='cos2',invisible='quali')
```



Nous voyons donc que plusieurs individus ont disparu de la représentation (notamment au centre), ce qui rend le nuage encore plus clairement divisé en 4 régions.

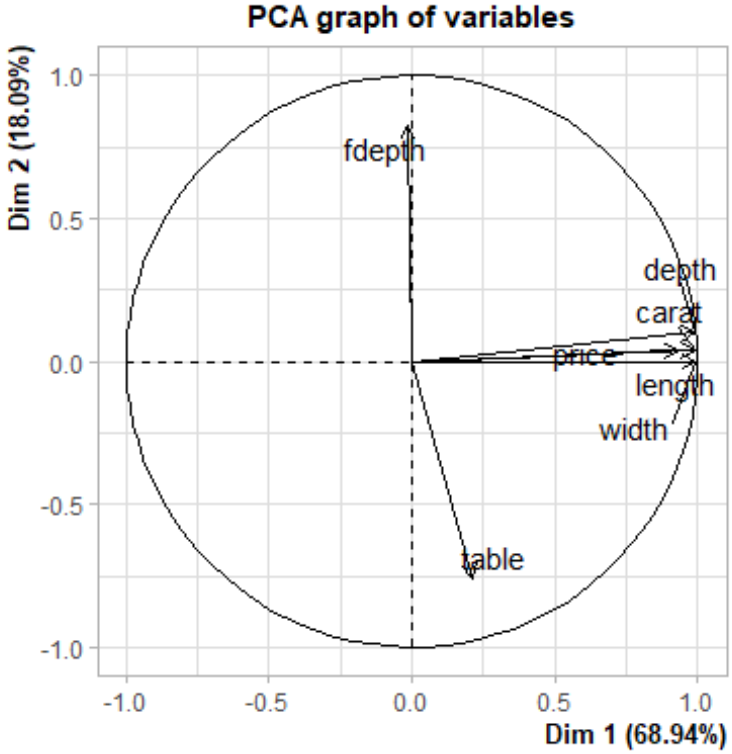
```
fviz_contrib(data_acp, choice = "ind", axes = 1:2, ylim=c(0,5), top=30)
```



Les individus 2764 et 2714 sont ceux qui contribuent le plus à la dimension 1, comme vu sur le nuage des individus précédemment. Leur contribution étant d'ailleurs très élevée par rapport aux autres individus, ils pourraient être considérés comme outliers. Nous choisissons de ne pas les considérer comme tel ici.

c. Etude des variables

```
plot.PCA(data_acp,choix="var")
```



Toutes les variables sont bien représentées ici (longueur de vecteurs proche de 1)

Le cercle des corrélations des variables montre que les variables **price**, **carat**, **depth**, **width** et **length** sont très fortement positivement corrélées. On retrouve ici la forte corrélation positive évoquée en début d'analyse. De plus, ces variables sont très proche de l'axe 1 donc on en déduit que la dimension 1 met en opposition les diamants ayant des prix élevés, une longueur, largeur et profondeur importante et un poids (en carat) élevé aux diamants ayant un prix faible, une longueur, largeur, profondeur faible et un poids (en carat) faible.

La variable **table** elle est quasiment orthogonale aux autres et contribue fortement à la dimension 2. La variables **fdepth** est quasiment confondue avec l'axe 2, elle explique donc majoritairement cet axe et elle est quasiment orthogonale aux variables définissant la dimension 1. On en déduit donc que la dimension 2 met en opposition les diamants ayant une forte de valeur pour la variable **fdepth** et une faible valeur pour la variable **table** et ceux ayant une faible valeur pour **fdepth** mais une forte pour **table**. On retrouve donc que ceux deux variables sont négativement corrélées.

d. Conclusion

Cette analyse en composante principale nous a montré que l'on explique plus de 85% de la variance en ne gardant seulement que 2 dimensions. L'étude de la projection des individus dans le plan F1/F2 met en évidence la séparation des individus en 4 groupes, chacun situé dans un cadran du repère. Cela pourrait vouloir dire que l'on peut séparer les diamants de notre jeu de données en 4 catégories différentes.
L'étude des variables nous a permis de donner un sens à chaque axe : l'axe 1 met en opposition les diamants longs, lourds, larges et profonds ayant un prix relativement élevé à ceux ayant les caractéristiques inverses et l'axe 2 met en opposition les diamants ayant une forte valeur de **fdepth** et une faible valeur de **table** à ceux ayant les caractéristiques inverses. On peut alors grâce à cela donner un sens aux 4 groupes mentionnés précédemment.

Le premier groupe (cadran supérieur gauche) rassemblerait les diamants ayant de petites dimensions et un prix relativement faible, ayant une forte valeur de **fdepth** et une faible valeur de **table**.
Le deuxième groupe (cadran supérieur droit) rassemblerait les diamants ayant de grandes dimensions et un prix relativement élevé, ayant une forte valeur de **fdepth** et une faible valeur de **table**.
Le troisième groupe (cadran inférieur gauche) rassemblerait les diamants ayant de petites dimensions et un prix relativement faible, ayant une faible valeur de **fdepth** et une forte valeur de **table**.

Le quatrième groupe (cadran inférieur droit) rassemblerait les diamants ayant de grandes dimensions et un prix relativement élevé, ayant une faible valeur de **fdepth** et une forte valeur de **table**.

Cependant, en ne gardant qu'un seul axe, on pourrait conclure que pour expliquer le prix d'un diamant, il suffit de connaitre sa longueur, largeur, profondeur et poids au vu de la forte corrélation positive de ces variables avec la variable **price**. Et donc que plus un diamant est lourd, long, large et profond, plus son prix sera élevé. Attention néanmoins à la forte corrélation de ces variables explicatives entre elles qui provoque peut-être de la redondance : certaines variables ne seraient peut-être pas nécessaires pour estimer le prix d'un diamant.

VI. Régression linéaire multiple

Nous proposons ici de traiter non plus analytiquement notre base de données mais économétriquement à l'aide de plusieurs régressions multiples. Celles-ci nous permettrons de trouver un modèle linéaire pouvant expliquer au mieux les valeurs de notre variable cible Y et ayant un pouvoir prédictif satisfaisant.

Modèle mathématiques : $Y = X\beta + \varepsilon$

Avec : $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_n \end{pmatrix}$ et $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

$\rightarrow\ y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$

- y_i : valeur de la variable cible pour l'individu i
- β : vecteur des coefficients du modèle (β_0 représente la constante)
- X : matrice des valeurs observées des variables Xi
- ε_i : terme d'erreur (résidu) pour l'individu i

Dans une régression linéaire multiple, la variable à expliquer Y est une variable quantitative et toutes les variables explicatives le sont également. La régression multiple maximise la variance de Y.

De plus, pour que les résultats d'une régression multiple soient exploitables, il faut que le modèle vérifie plusieurs postulats:

1. P1 : les résidus epsilon sont centrés (espérance nulle)
2. P2 : les résidus ont même variance (homoscédasticité)
3. P3 : les résidus sont indépendants entre eux
4. P4 : les résidus sont des variables aléatoires Gaussiennes (param (0,sigma2))

Vérification des Postulats

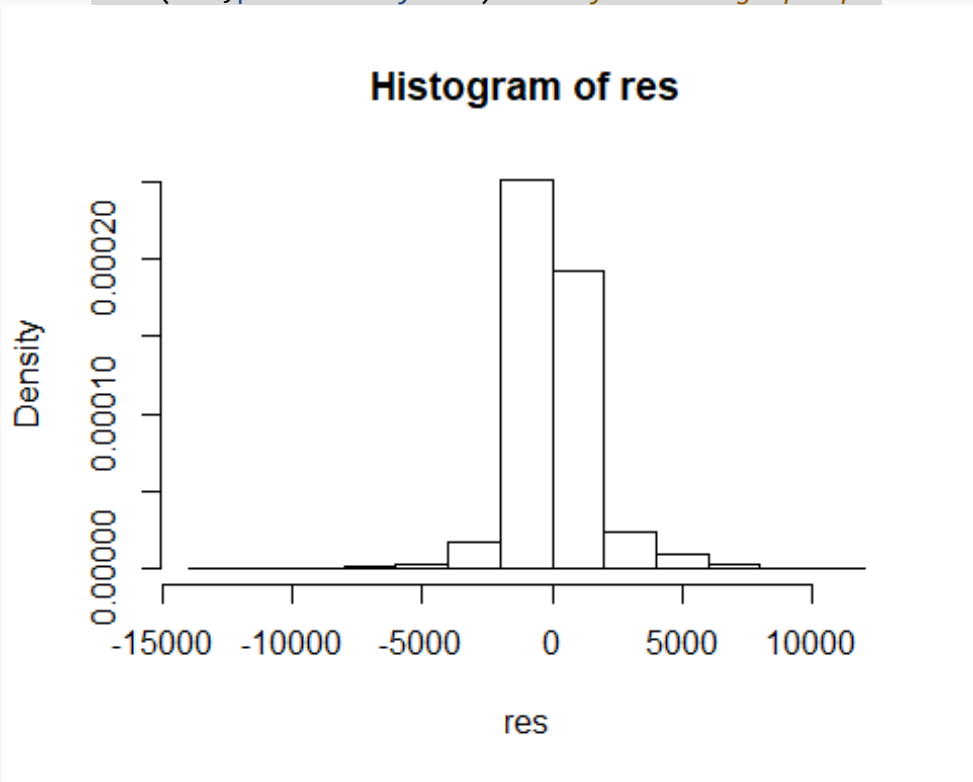
```
data=data[-c(2453,2714,2754,2764,5182,5286),] #on enlève les diamants qui son
t considérés comme outliers par la commande influenceIndexPlot(Reg_1)
Reg_1=lm(data$price~data$carat + data$fdepth + data$table + data$length + dat
a$width+ data$depth)
vif(Reg_1)
```

```
##  data$carat data$fdepth  data$table data$length  data$width  data$depth
##   24.998189   28.243201    1.148146  939.982169  912.517295 1777.776131
```

On remarque grâce au test vif que nos variables sont très corrélées entre elles (on le savait déjà grâce à l'ACP effectuée précédemment), cela va donc sûrement poser problème pour la régression multiple qui supporte très mal les soucis de multicolinéarité.

Vérifions si les résidus sont distribués de manière gaussienne:

```
res=Reg_1$residuals
hist(res,probability = T) #vérification graphique
```




```
shapiro.test(res[1:5000]) #on ne prend que les 5000 premiers résidus car le t
est sur R n'autorise pas plus de 5000 observations
##  Shapiro-Wilk normality test
##
## data:  res[1:5000]
## W = 0.84528, p-value < 2.2e-16
```

A première vue les résidus semblent suivre une loi normale centrée mais le test de Shapiro réfute violemment cette hypothèse.

Vérifions l'homoscédasticité des résidus:

```
bptest(Reg_1)
##  studentized Breusch-Pagan test
##
## data:  Reg_1
## BP = 1142.9, df = 6, p-value < 2.2e-16
```

Au vu de la valeur de la p_value, on rejette également l'hypothèse nulle d'égalité des variances des résidus.

Vérifions pour finir l'indépendance des résidus:

```
dwtest(Reg_1)
##  Durbin-Watson test
##
## data:  Reg_1
## DW = 1.4915, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

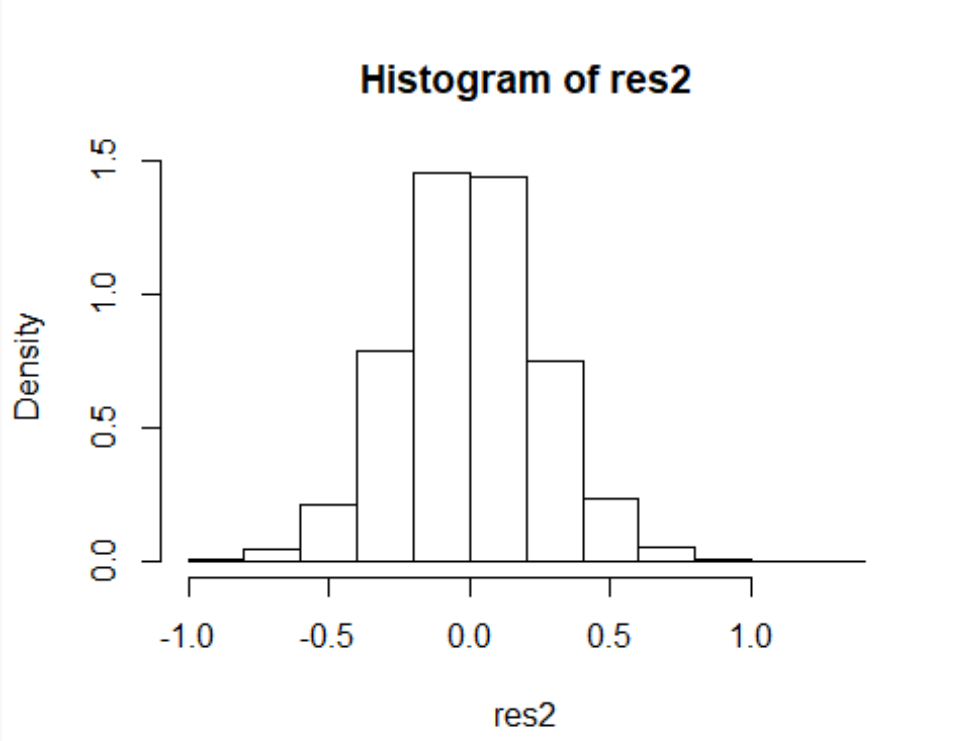
La p-value nous indique le rejet de l'hypothèse nulle d'indépendance des résidus. De manière générale l'indépendance des résidus est supposée vraie de par la façon dont les données ont été collectées. Or ici, nous n'avons pas accès à un document explicatif sur la façon dont cette base de données a été créée. Il se peut très bien que tous les diamants proviennent d'un même bijoutier auquel cas, effectivement les données sont biaisées.

Au vu du rejet de tous les postulats, nous allons modifier notre variable à expliquer Y pour voir si les résidus vérifient cette fois les hypothèses nécessaires pour pouvoir mener à bien notre régression linéaire multiple.

```
Reg_2=lm(log(data$price) ~ data$carat + log(data$fdepth) + log(data$table) +d
ata$length + data$width + data$depth)
res2=Reg_2$residuals
vif(Reg_2)
```

```
##       data$carat log(data$fdepth)  log(data$table)      data$length
##        25.001279        28.592095         1.157434        959.380053
##       data$width       data$depth
##       900.894836      1790.263463
```

```
hist(res2,probability = T) #vérification graphique
```



```
shapiro.test(res2[1:5000])
##  Shapiro-Wilk normality test
##
## data:  res2[1:5000]
## W = 0.99868, p-value = 0.0004184

bptest(Reg_2)
##  studentized Breusch-Pagan test
##
## data:  Reg_2
## BP = 34.975, df = 6, p-value = 4.358e-06

dwtest(Reg_2)
##  Durbin-Watson test
##
## data:  Reg_2
## DW = 1.6444, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

Malheureusement, même en passant au log du prix, les résidus ne vérifient toujours aucun des postulats (même si l'on note une nette amélioration pour la normalité et l'homoscédasticité). Même en essayant plusieurs transformations sur notre variable Y et nos variables explicatives (suppression des variables très corrélées, variables au carré, formules de boxCox), nous n'obtenons jamais vérification des postulats par les résidus. Nous sommes donc obligées à ce stade d'abandonner la régression multiple car non pertinente.

Nous allons maintenant effectuer des régressions différentes de la régression multiple en ce que celles-ci peuvent régler les problèmes de multicolinéarité des variables.

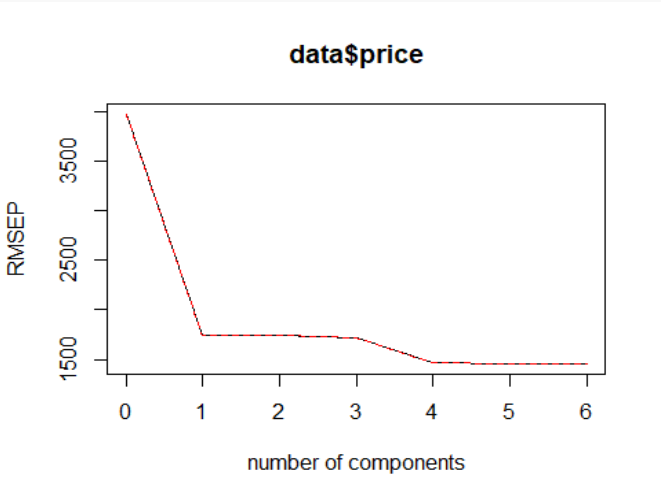
VII. Régression sur Composantes Principales

La régression sur Composantes Principales consiste à expliquer la variable Y non plus à l'aide des variables d'origine mais à l'aide de composantes principales obtenues par une méthode d'analyse factorielle (ACP par exemple). L'avantage de considérer des variables synthétiques est qu'elles sont non corrélées entre elles, on contourne ainsi grâce à cette méthode les problèmes de multicolinéarité des variables explicatives. La régression sur composantes principale maximise la variance de X (matrice des variables explicatives) et non la variance de Y. On portera donc une attention particulière au pourcentage de variance expliquée pour Y par nos différents modèles.

```
Reg_pcr=pcr(data$price~ data$carat + data$fdepth + data$table + data$length +
data$width + data$depth, scale=T, validation ='LOO') #On standardise les donn
ées car elles n'ont pas toutes la même unité
summary(Reg_pcr)
```

```
## Data:    X dimension: 5388 6
##  Y dimension: 5388 1
## Fit method: svdpc
## Number of components considered: 6
##
## VALIDATION: RMSEP
## Cross-validated using 5388 leave-one-out segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV            3974     1748     1739     1714     1468     1459     1457
## adjCV         3974     1748     1739     1714     1468     1459     1457
##
## TRAINING: % variance explained
##             1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X             66.64    87.73    99.47    99.98    99.99   100.00
## data$price    80.68    80.87    81.43    86.41    86.57    86.63
```

```
validationplot(Reg_pcr)
```

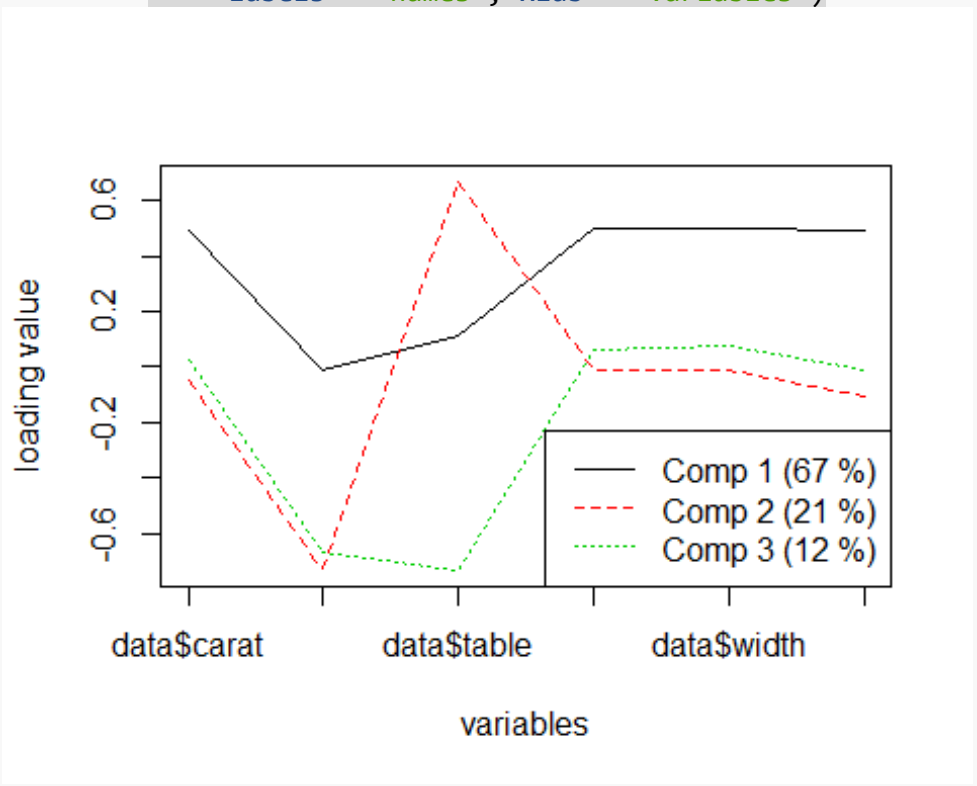


Le résumé de cette régression PCR nous montre que le nombre de composantes à retenir pour minimiser l'indicateur PRESS est de 6. Il n'y a donc aucune réduction de dimension. En regardant bien les pourcentages de variances expliquées pour X et Y, nous décidons de ne retenir que les 3 premières composantes car celles-ci expliquent 99.47% de la variance de X et plus de 80% de la variance de Y, ce qui est plutôt bon.

a. Signification des axes

Pour définir chaque axe retenu, nous allons représenter le graphe des loadings.

```
plot(Reg_pcr, "loadings", comps = 1:3, legendpos = "bottomright",
     labels = "names", xlab = "variables")
```




```
Yloadings(Reg_pcr)[,1:3]
```

```
##    Comp 1    Comp 2    Comp 3
## 1785.0902 -155.9995  353.9599
```

On retrouve donc les résultats de l'ACP pour les 2 premières dimensions : l'axe 1 met en opposition les diamants ayant des mesures (poids, longueur, largeur, profondeur) élevées à ceux en ayant des plus faibles.

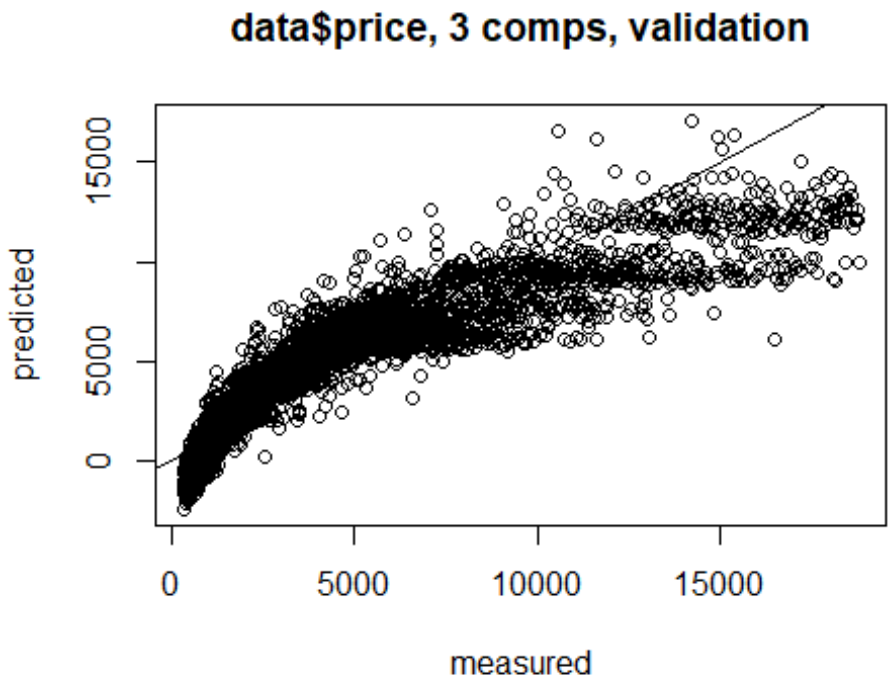
Pour l'axe 2, on trouve une forte corrélation positive avec la variable **table** et une forte corrélation négative avec la variable **fdepth**. L'axe 2 met donc en opposition les diamants ayant une forte valeur de **table** et une faible valeur de **fdpeth** à ceux ayant au contraire une faible valeur de **table** mais une valeur de **fdepth** élevée.

Enfin, l'axe 3 met lui en opposition les diamants ayant de fortes valeurs de **fdepth** et **table** à ceux ayant des faibles valeurs pour les deux variables.

La variable à expliquer Y est elle très positivement corrélée à la dimension 1, comme déjà vu lors de l'ACP. Il semblerait donc que les principaux facteurs explicatifs du prix d'un diamant soit son poids (en carat) et ses dimensions. C'est en effet la pensée commune.

b. Pouvoir prédictif du modèle

```
predplot(Reg_pcr,ncomp=3, line=T)
```



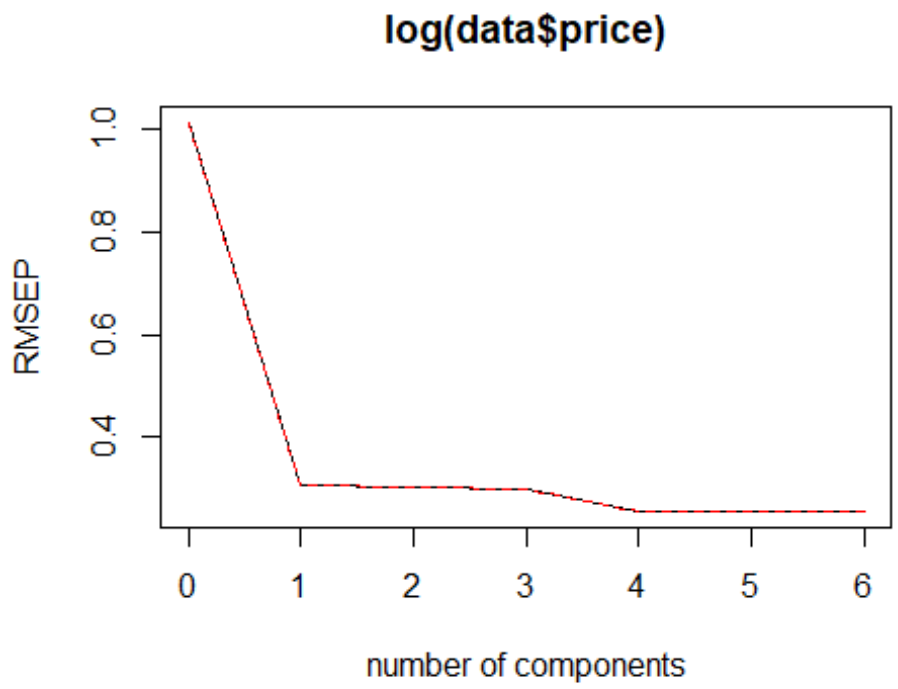
On voit grâce à ce graphique que la prédiction faite par notre modèle n'est pas très satisfaisante. La courbe obtenue est logarithmique et non linéaire. De plus, la présence de prix prédits négatifs montre que le modèle peut surement être amélioré. Il semble aussi que le modèle estime bien les prix des diamants vendus entre 1000 et 10000 dollars mais beaucoup moins bien les autres fourchettes de prix.

Nous allons essayer de voir si expliquer le log du prix au lieu du prix lui-même améliore le modèle:

```
Reg_pcr2=pcr(log(data$price)~data$carat + data$fdepth + data$table + data$len
gth + data$width + data$depth, scale=T, validation ='LOO') #On standardise le
s données car elles n'ont pas toutes la même unité
summary(Reg_pcr2)
```

```
## Data:    X dimension: 5388 6
##  Y dimension: 5388 1
## Fit method: svdpc
## Number of components considered: 6
##
## VALIDATION: RMSEP
## Cross-validated using 5388 leave-one-out segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           1.014   0.3054   0.3021   0.2971   0.2545   0.2544   0.2543
## adjCV        1.014   0.3054   0.3021   0.2971   0.2545   0.2544   0.2543
##
## TRAINING: % variance explained
##                  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X                  66.64    87.73    99.47    99.98    99.99   100.00
## log(data$price)    90.93    91.14    91.43    93.71    93.72    93.73
```

```
validationplot(Reg_pcr2)
```

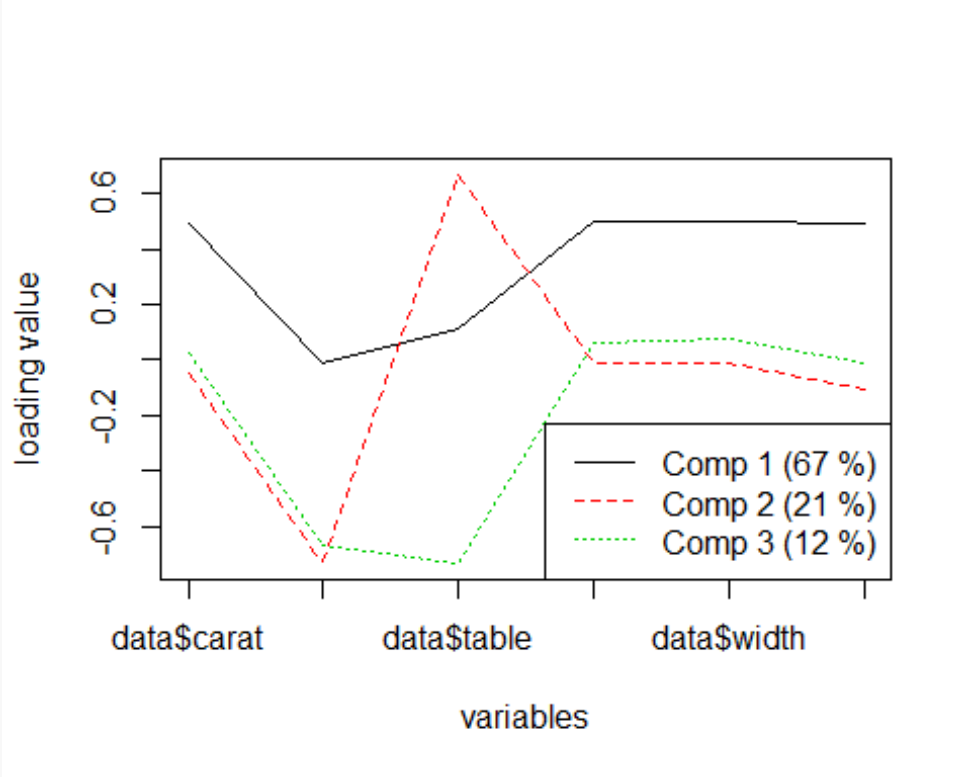


Encore une fois, il faut 6 composantes pour minimiser l'indicateur PRESS mais on se rend compte que 4 composantes donne quasiment la même valeur. On voit donc le début d'une réduction de dimension. En observant les pourcentages de variances expliquées, en gardant

également 3 composantes, on explique cette fois-ci plus de 91% de la variance de Y (log(price)) et plus de 99% de la variance de X. Ce modèle semble donc meilleur.

Vérifions si la signification des axes est toujours la même avec ce modèle:

```
plot(Reg_pcr2, "loadings", comps = 1:3, legendpos = "bottomright",
     labels = "names", xlab = "variables")
```

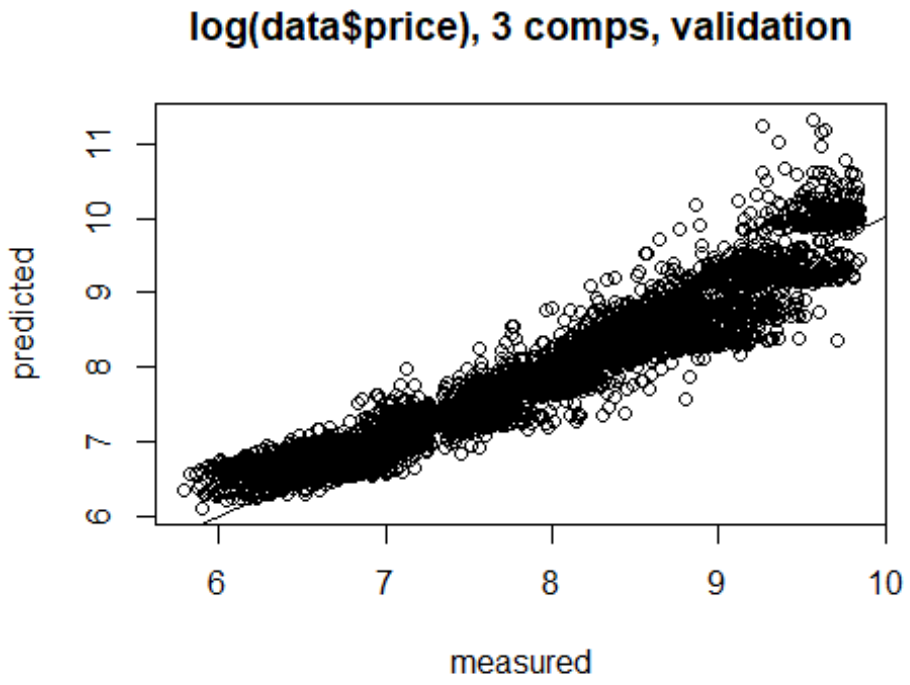



```
Yloadings(Reg_pcr2)[,1:3]
```

```
##      Comp 1      Comp 2      Comp 3
##  0.48353728 -0.04064062  0.06547932
```

La définition des axes est inchangée et la variable Y (log(price)) est toujours plus corrélée à l'axe 1, donc on conserve les interprétations faites plus haut.

```
predplot(Reg_pcr2,ncomp=3, line=T)
```



On obtient cette fois une courbe beaucoup plus proche d'une droite. Nous allons donc garder ce modèle.

c. Coefficients du modèle

```
coef(Reg_pcr2,ncomp=3)
```

```
## , , 3 comps
##
##              log(data$price)
## data$carat        0.24236597
## data$fdepth      -0.01971249
## data$table       -0.01926520
## data$length       0.24584869
## data$width        0.24633490
## data$depth        0.24315436
```

Les coefficients de la régression permettent de quantifier l'impact de chaque variable sur la variable à expliquer Y(**log(price)**) lorsque l'on prend en compte 3 composantes principales.

Le prix d'un diamant est donc fortement positivement lié à sa largeur, sa longueur, son poids et sa profondeur. Par exemple, si l'on augmente d'une unité la variable carat, le log du prix du diamant augmente 0.24, toutes choses égales par ailleurs.

En revanche, le prix est négativement corrélé aux variables **fdepth** et **table**. Si l'on augmente d'une unité la variable **table** alors le log du prix du diamant diminue de 0.02, TCEPA.

d. Conclusion

Cette méthode nous a donc permis de trouver un modèle satisfaisant pour expliquer non pas le prix mais le log du prix des diamants de la base de données.
Avec cette méthode, on voit que ce sont les dimensions et le poids d'un diamant qui influence le plus positivement sont prix. Des valeurs de table ou de fdepth trop grandes auront en revanche tendance à le faire diminuer.
Cependant, la régression sur composantes principales est une méthode maximisant uniquement la variance des X (variables explicatives) et non la variance de Y. Nous allons donc pour finir notre étude, utiliser une autre méthode de régression sur variables latentes, la PLS.

VIII. Partial Least Squares Regression

Cette méthode de régression a l'avantage de pouvoir utiliser et des variables quantitatives et des variables qualitatives comme variables explicatives. De plus, cette méthode maximise non plus uniquement la variance de Y ou la variance des X mais la covariance entre X et Y. Elle est également très pertinente en cas de problème de degrés de liberté et de multicolinéarité.

Nous allons donc effectuer une régression PLS en utilisant uniquement les variables explicatives quantitatives, pour pouvoir comparer les résultats obtenus aux résultats de la pcr effectuée précédemment.

a. Régression PLS

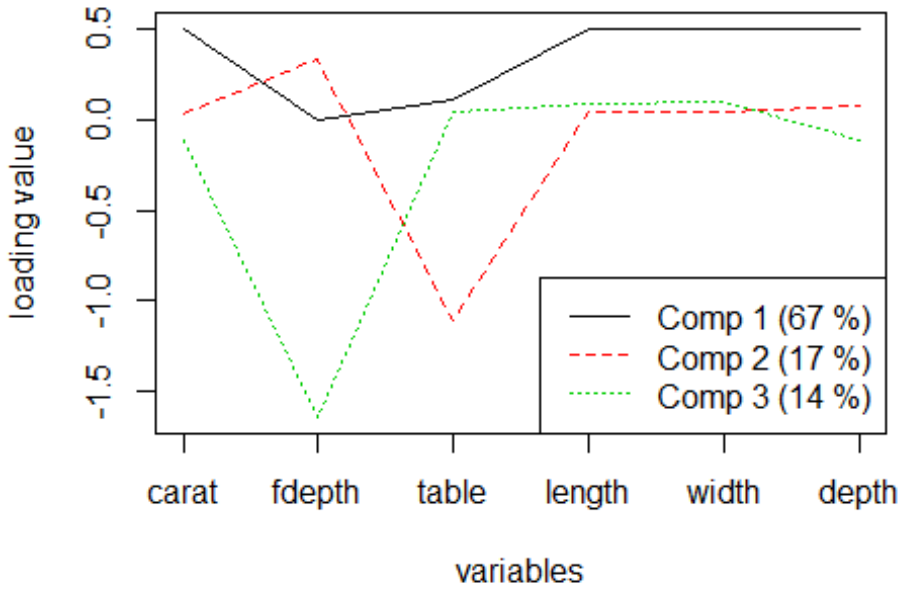
```
Reg_pls=plsr(log(price) ~ carat + fdepth + table + length + width + depth, da
ta=data, scale=TRUE, validation="LOO")
summary(Reg_pls)
```

```
## Data:    X dimension: 5388 6
##  Y dimension: 5388 1
## Fit method: kernelpls
## Number of components considered: 6
##
## VALIDATION: RMSEP
## Cross-validated using 5388 leave-one-out segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           1.014   0.3014   0.2942   0.2877   0.2545   0.2543   0.2543
## adjCV        1.014   0.3014   0.2942   0.2877   0.2545   0.2543   0.2543
##
## TRAINING: % variance explained
##             1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X             66.62     83.9    98.09    99.98    99.99   100.00
## log(price)    91.17     91.6    91.96    93.71    93.72    93.73
```

Là encore, 6 composantes minimisent l'indicateur PRESS mais 4 composantes seulement donnent quasiment les mêmes résultats. Pour pouvoir comparer cette régression à la pcr précédente, nous allons également conserver 3 axes. Les 3 premiers axes expliquent 98% de la variance des X et quasiment 92% de la variance de Y (log(price)). Ne garder que 3 axes est donc bien justifié ici.

b. Signification des axes

```
plot(Reg_pls, "loadings", comps = 1:3, legendpos = "bottomright",
                  labels = "names", xlab = "variables")
```




```
Yloadings(Reg_pls)[,1:3]
```

```
##     Comp 1     Comp 2     Comp 3
## 0.48444425 0.07548966 0.10993715
```

Nous retrouvons pour la dimension 1 exactement la même définition que lors de nos différentes pcr. L'axe 1 met toujours en opposition les diamants ayant de fortes valeurs pour les variables **carat**, **length**, **width** et **depth** aux diamants ayant de faibles valeurs pour ces variables.

En revanche, la définition de l'axe 2 change. Lors de notre pcr, l'axe 2 mettait en opposition les diamants ayant une forte valeur pour la variable **table** et une faible valeur pour la variable **fdepth** à ceux ayant les caractéristiques inverses. Ici, la variable **table** est fortement négativement corrélée à l'axe 2 et la variable **fdepth** est plus faiblement positivement corrélée à l'axe 2. On a donc inversé le sens de la dimension 2 par rapport à la pcr.

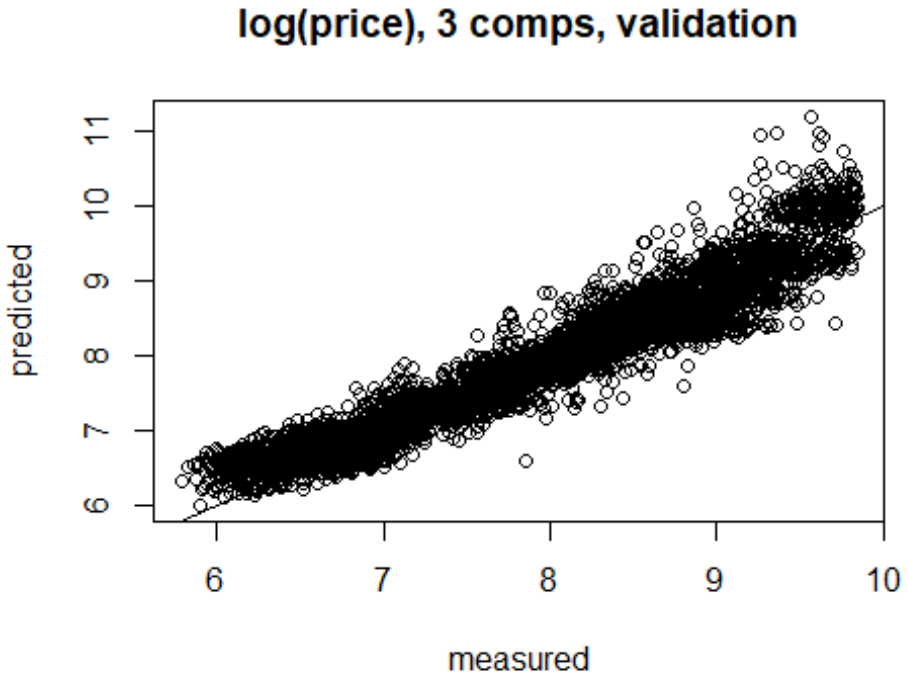
Le 3ème axe est quant à lui uniquement définit par la variable **fdepth**, négativement corrélée à l'axe. La dimension 3 met donc en opposition les diamants ayant une faible valeur de **fdepth** à ceux ayant une forte valeur.

Les loadings montrent encore une fois que la variable cible Y est plus corrélée avec la dimension 1 qu'avec toutes les autres dimensions. Cependant, le coefficient pour l'axe 2 est

cette fois-ci positif et le coefficient pour l'axe 3 est plus élevé que pour la PCR. La variable **fdepth** donne donc l'impression de jouer un rôle plus important dans l'explication du prix des diamants ici.

c. Qualité de prédiction du modèle

```
predplot(Reg_pls,ncomp=3, line=T)
```



On obtient également ici une courbe quasi-linéaire. Le modèle est donc satisfaisant. Les 2 méthodes, PCR et PLS, donnent donc à peu de choses près les mêmes résultats pour ce modèle.

d. Coefficients du modèle

```
coef(Reg_pls,ncomp=3)
```

```
## , , 3 comps
##
##          log(price)
## carat   0.11084836
## fdepth -0.06194383
## table  -0.04295844
## length  0.28956264
## width   0.29548142
## depth   0.28526265
```

On remarque tout d'abord que les signes des coefficients sont les mêmes que pour la PCR mais la valeur du coefficient pour carat est beaucoup plus faible (divisé par 2). Cela voudrait donc dire que le poids d'un diamant peut ne pas être aussi important dans l'explication de son prix qu'on pourrait penser à première vue.

Si l'on compare ces résultats à ceux de la PCR, on pourrait donc dire qu'un diamant ayant un poids important mais des dimensions plus faible aura le même prix qu'un diamant de poids plus faible mais ayant des dimensions plus importantes. En revanche si ce dernier diamant à de fortes valeurs pour fdepth et table, alors son prix diminuera plus rapidement que pour le premier diamant.

En ne regardant que les résultats que l'on obtient ici, on peut dire que si l'on augmente la profondeur du diamant d'une unité (le mm ici en l'occurrence) alors le log du prix augmente de 0.29, TCEPA. Enfin, si l'on augmente d'une unité le variable **fdepth** alors le log du prix du diamant diminue de 0.06, TCEPA.

e. Conclusion

On retrouve avec cette régression PLS les résultats les plus souvent revenus dans toute notre étude. Ce sont les variables **carat**, **length**, **width** et **depth** qui définissent le premier axe et qui sont le plus positivement corrélées à la variable à expliquer. Néanmoins, cette régression a aussi permis de montrer que la variable **fdepth** peut jouer un rôle plus important qu'on pourrait le penser. Il en résulte que les coefficients de régression obtenus sont différents de ceux de la PCR. Nous avons interprété ces résultats dans la section précédente.

IX. Conclusion générale

Nous avons à travers cette étude, partiellement réussi à expliquer le prix d'un diamant. En effet, même si grâce aux résultats de l'analyse descriptive des données, nous avons pu voir que parmi les variables qualitatives, c'est la variable **cut** qui influence le plus positivement le prix des diamants, nous n'avons pas introduit ces variables dans nos différentes régressions par soucis pratique. Il serait donc pertinent dans une autre étude de ne considérer que les variables qualitatives et d'utiliser des méthodes comme l'analyse des correspondances pour essayer d'expliquer le prix d'un diamant. Nous avons choisi de ne pas le faire ici pour ne pas alourdir le rapport et pour se concentrer sur les méthodes de régression vues en cours.

Par suite l'analyse des correspondances multiples n'a pas apporté beaucoup plus d'informations sur les variables qualitatives à part que celles-ci sont légèrement corrélées aux variables quantitatives et notamment aux variables quantifiant les dimensions d'un diamant et que peu de modalités ont un effet significatif dans l'explication de la variable **price**.

L'analyse en composantes principales nous a elle permit de montrer que les diamants peuvent être séparés en 4 groupes, définis par les différentes variables quantitatives. Il est également apparu lors de cette analyse que ce sont les variables **carat**, **length**, **width** et **depth** qui sont le plus positivement corrélées à la variable **price** et qui influencent donc le plus positivement le prix d'un diamant.

Malheureusement, nous n'avons pu réaliser de régression multiple sur nos données car celles-ci ne vérifient aucun des postulats nécessaires.

La régression PCR a elle aussi mis en avant la grande part explicative des variables susmentionnées dans l'explication du prix d'un diamant. De plus, cette régression a permis de montrer qu'il était beaucoup plus judicieux et pertinent d'expliquer le log du prix d'un diamant par les variables quantitatives à notre disposition au lieu du prix lui-même.

Pour finir, la régression PLS, bien qu'elle confirme les résultats déjà énoncés, a mis en lumière que la variable **carat** peut ne pas être si importante que ça comparée aux dimensions du diamant et que la variable **fdepth** peut avoir une influence négative sur le prix assez importante. Nous nous retrouvons donc avec 2 fonctions possibles pour expliquer et prédire le prix d'un diamant à la suite de cette étude.