Projet d'Analyse de Données avec R

Wilfried TCHATCHOU SINKAM & Joan Cindy MIKONGO OUAMBO

INTRODUCTION ET PROBLÈMATIQUE

La problématique assignée à ce projet consiste à prédire le prix des diamants en fonction de diverses variables. Pour ce faire, nous utiliserons le dataset diamonds et un modèle de régression qui prendra en compte les caractéristiques des diamants telles que le carat, la qualité de la coupe (cut), la couleur (color), la clarté (clarity), la profondeur (depth), la largeur du sommet (table), ainsi que les dimensions physiques (x, y, z).

L'objectif est de développer un modèle capable de capturer les relations entre ces variables et le prix des diamants, afin de pouvoir estimer le prix d'un diamant donné en se basant sur ses caractéristiques. Cette prédiction peut être utilisée par les professionnels de l'industrie diamantaire, tels que les bijoutiers, les négociants en diamants ou les enchérisseurs, pour évaluer la valeur des diamants et prendre des décisions éclairées lors de l'achat ou de la vente de ces pierres précieuses.

Il est important de souligner que la qualité et l'exactitude des prédictions dépendront de la qualité des données d'entraînement, de la sélection appropriée des caractéristiques pertinentes et du choix du modèle de régression approprié.

I- Importation et Description du Dataset

1. Description

Nous avons utilisé les bibliothèques suivantes :

```
library(readr)
library(car)
```

```
Le chargement a nécessité le package : carData
```

```
library(ggplot2)
library(lm.beta)
library(corrplot)
```

```
corrplot 0.92 loaded
```

```
#| echo: false
donnees <- read_csv("diamonds.csv")
```

```
New names:
* `` -> `...1`
```

```
Rows: 53940 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): cut, color, clarity
dbl (8): ...1, carat, depth, table, price, x, y, z

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#View(donnees)
#| echo: false
attach(donnees)
```

Notre Dataset est constitué de 53940 observations. En dehors de la variable d'identification, le dataset est composé de 10 variables ; 07 d'entre elles à savoir carat, depht, table, price, x, y et z sont de nature quantitatives et de type réel; le cut, le color et la clarity sont de nature qualitatives.

Carat: Poids du diamant en carat ; Depth: Profondeur totale pourcentage ; Table: Largeur du sommet du diamant ; Price: Prix du diamant ; X: Longueur du diamant en mm ; Y: Largeur du diamant en mm ; Z: Profondeur du diamant en mm ; Cut: Qualité de la coupure du diamant ; Color: Couleur du diamant ; Clarity: mesure de la clarté du diamant.

2. Vérification des valeurs manquantes

Pour vérifier les valeurs manquantes, nous avons utilisées le code ci-dessous, et nous avons obtenu aucune valeur manquante.

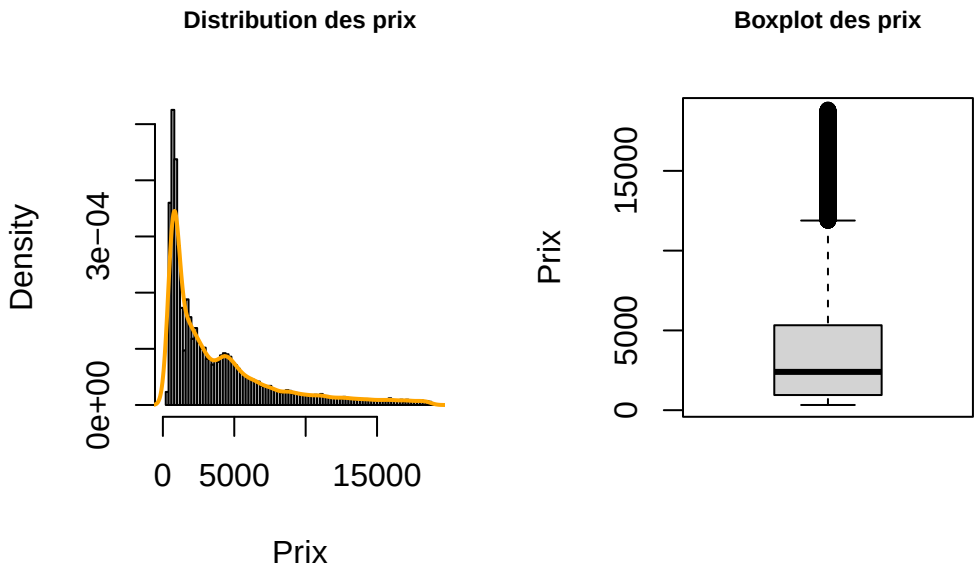
```r
valeur_manquant <- sum(colSums(is.na(donnees)))
if (valeur_manquant>1){
  cat("Les valeurs manquantes sont au nombre de ", valeur_manquant, "\n")
} else {
  cat("Aucune valeur manquante, car il y'a ", valeur_manquant, " valeur manquante!", "\n")
}
```

```
Aucune valeur manquante, car il y'a  0  valeur manquante!
```

II- Distribution des données

1- La variable dépendante: Price

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



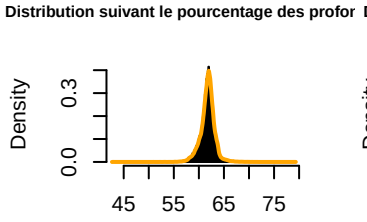
Les prix des diamants sont compris entre 326 et 18823 dollars. La moitié des diamants ont un prix inférieur ou égal à 2401 dollars et le diamant moyen vaut 3933 dollars. L'examen de la distribtion assortie de sa courbe de densité montre que les prix en carat ne suivent pas une loi normale. La présence de valeurs extrêmes est identifiable par des points individuels au-delà de la moustache supérieure. Il s'agit des diamants dont les prix sont significativement plus élevés

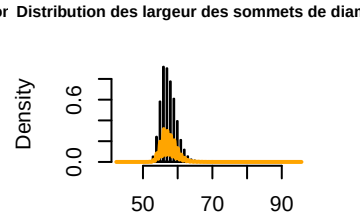
par rapport à ceux des autres diamants. Ainsi, les diamants d'une valeur supérieure à 12000 dollars sont considérés comme étant des valeurs extrêmes.

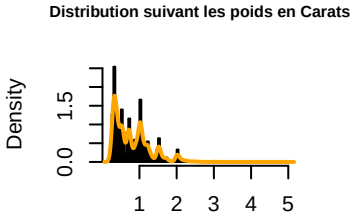
2- Les variables indépendantes quantitatives

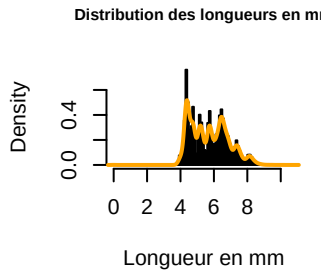
2.1. Distribution

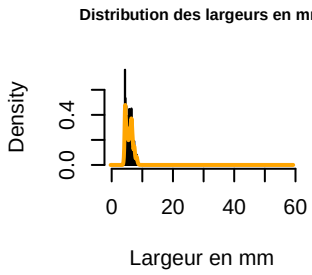
```
     carat             depth           table             x         
 Min.   :0.2000   Min.   :43.00   Min.   :43.00   Min.   : 0.000  
 1st Qu.:0.4000   1st Qu.:61.00   1st Qu.:56.00   1st Qu.: 4.710  
 Median :0.7000   Median :61.80   Median :57.00   Median : 5.700  
 Mean   :0.7979   Mean   :61.75   Mean   :57.46   Mean   : 5.731  
 3rd Qu.:1.0400   3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 6.540  
 Max.   :5.0100   Max.   :79.00   Max.   :95.00   Max.   :10.740  
       y                z         
 Min.   : 0.000   Min.   : 0.000  
 1st Qu.: 4.720   1st Qu.: 2.910  
 Median : 5.710   Median : 3.530  
 Mean   : 5.735   Mean   : 3.539  
 3rd Qu.: 6.540   3rd Qu.: 4.040  
 Max.   :58.900   Max.   :31.800  
```

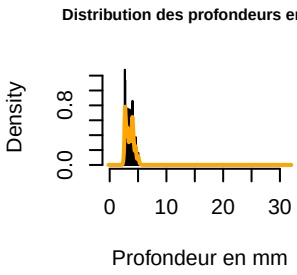






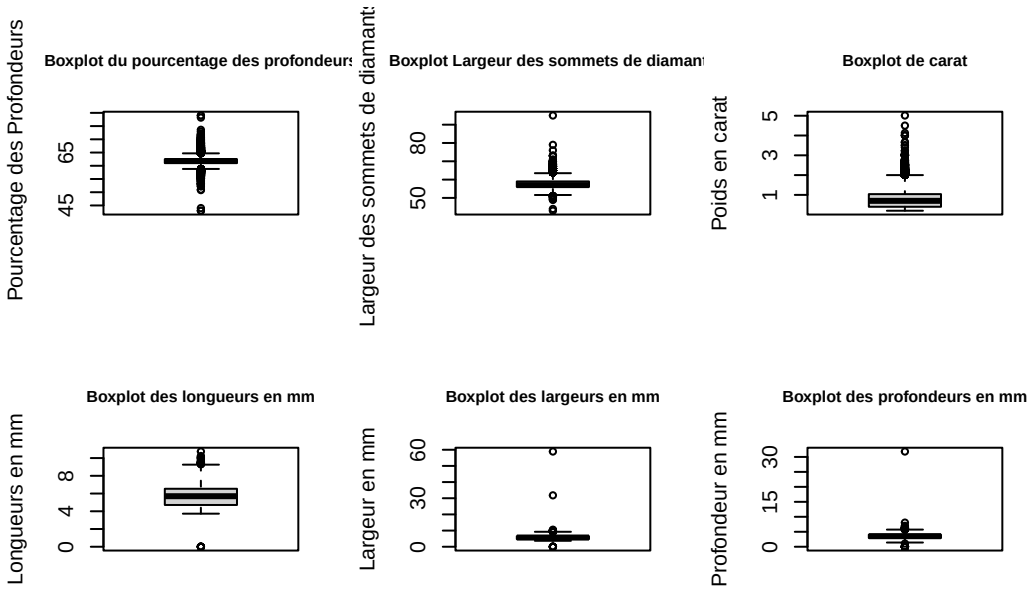






- Le diamant moyen pèse 0.79 Carat, un sommet de 57.46 mm, un pourcentage de profondeur égal à 61.75%, une longueur de 5.731 mm, une largeur de 5.73 mm et une profondeur de 3.53.
- L'examen des distribtions assorties de leurs courbes de densité montre qu'en dehors du pourcentage des profondeurs (Depth), les autres variables ne suivent pas une loi normale.

2.2. Valeurs aberrantes



La présence de valeurs extrêmes est identifiable par des points individuels au-delà des moustaches. Il s'agit des diamants dont les caractéristiques sont significativement plus élevées ou moins élevées par rapport à celles des autres diamants.

3- Les variables indépendantes quanlitatives

- Les couleurs vont de D à J, qui représentent la couleur la plus meilleure à la plus pire. Ainsie, la plus grande partie des diamants (20.93%) ont une couleur jugée de qualité moyenne (couleur G). Seuls 5.21% des diamants ont une couleur considérée comme mauvaise (couleur J).
- 39.95% des diamants ont des coupures considérées comme idéales, 25.57% sont de qualité premium, et seul 2.98% sont considérées comme ayant des coupures passables.
- Les mesures de clarté varie entre I1 et IF. I1 représente la pire des clartés où des défauts internes sont visibles à l'oeil nu (1.37%) tandis que IF représente la meilleure des clartés où aucune inclusion ou défaut interne n'est visible sous une loupe grossissante de 10x (3.32%), c'est la plus haute qualité de clarté.
- VVS1/VVS2 : De très petites inclusions difficiles à voir sous une loupe grossissante de 10x.
- VS1/VS2 : De petites inclusions visibles sous une loupe grossissante de 10x, mais difficilement visibles à l'œil nu.
- SI1/SI2 : Des inclusions visibles sous une loupe grossissante de 10x, et parfois visibles à l'œil nu.

III- Analyse bivariée

1.Test d'indépendance entre la variable à expliquer (Price) et les variables qualitatives

1.1. Entre Price et Cut

```
               Df    Sum Sq   Mean Sq F value Pr(>F)
cut             4 1.104e+10 2.760e+09   175.7 <2e-16 ***
Residuals   53935 8.474e+11 1.571e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value <2e-16, donc nous pouvons rejeter l'hypothèse nulle que toutes les catégories de 'cut' ont la même moyenne. En termes simples, il y a une différence significative sur la variable dépendante en fonction des différents niveaux de 'cut'. Le F-value élevé (175.7) confirme également que la variable 'cut' a un effet significatif et fort sur la variable dépendante.

1.2. Entre Price et Color

```
               Df    Sum Sq   Mean Sq F value Pr(>F)
color           6 2.685e+10 4.475e+09   290.2 <2e-16 ***
Residuals   53933 8.316e+11 1.542e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La p-value pour la variable 'color' est inférieure à 0.001 (indiquée par ***), ce qui signifie que nous pouvons rejeter l'hypothèse nulle que toutes les catégories de 'color' ont la même moyenne. En termes simples, il y a une différence significative sur la variable dépendante en fonction des différents niveaux de 'color'. Le F-value élevé (290.2) confirme également que la variable 'color' a un effet significatif et fort sur la variable dépendante.

1.3. Entre Price et Clarity

```
               Df    Sum Sq   Mean Sq F value Pr(>F)
clarity         7 2.331e+10 3.330e+09     215 <2e-16 ***
Residuals   53932 8.352e+11 1.549e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La p-value pour la variable 'clarity' est inférieure à 0.001 (indiquée par ***), ce qui signifie que nous pouvons rejeter l'hypothèse nulle que toutes les catégories de 'clarity' ont la même moyenne. En termes simples, il y a une différence significative sur la variable dépendante en fonction des différents niveaux de 'clarity'. Le F-value élevé (215) confirme également que la variable 'clarity' a un effet significatif et fort sur la variable dépendante.

2. Test d'indépendance entre les variables qualitatives

2.1. Entre Cut et Clarity

```
    Pearson's Chi-squared test

data:  table(cut, clarity)
X-squared = 4391.4, df = 28, p-value < 2.2e-16
```

La p-value très basse nous permet de rejeter l'hypothèse nulle d'indépendance entre les variables 'cut' et 'clarity'. En d'autres termes, il existe une association statistiquement significative entre la qualité de la coupure ('cut') et la clarté ('clarity') des données analysées. Les deux variables ne sont pas indépendantes, et les variations dans l'une sont associées à des variations dans l'autre.

2.2. Entre Cut et Color

```
    Pearson's Chi-squared test

data:  table(cut, color)
X-squared = 310.32, df = 24, p-value < 2.2e-16
```

La p-value très basse nous permet de rejeter l'hypothèse nulle d'indépendance entre les variables 'cut' et 'color'. En d'autres termes, il existe une association statistiquement significative entre la qualité de la coupure ('cut') et la couleur ('color') des données analysées. Les deux variables ne sont pas indépendantes, et les variations dans l'une sont associées à des variations dans l'autre.

2.3. Entre Color et Clarity

```
    Pearson's Chi-squared test

data:  table(color, clarity)
X-squared = 2047.1, df = 42, p-value < 2.2e-16
```

La p-value très basse nous permet de rejeter l'hypothèse nulle d'indépendance entre les variables 'color' et 'clarity'. En d'autres termes, il existe une association statistiquement significative entre la couleur ('color') et la clarté ('clarity') des données analysées. Les deux variables ne sont pas indépendantes, et les variations dans l'une sont associées à des variations dans l'autre.

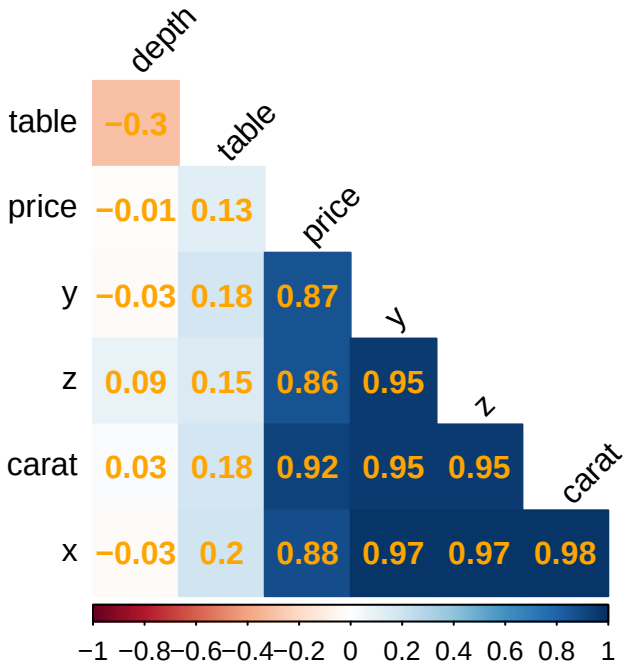
3. Test d'indépendance entre la variable à expliquer (Price) et les variables quantitatives

3.1. La covariance

```
       carat         depth         table         price             x
1.742765e+03 -6.085371e+01  1.133318e+03  1.591563e+07  3.958021e+03
           y             z
3.943271e+03  2.424713e+03
```

La covariance respective entre le prix, le poids en carat, la largeur du sommet (table) et les caractéristiques physiques x, y, z du diamant est positive; donc ces variables évoluent dans le même sens. Autrement dit, une augmentation de chacune d'elle entraine une augmentation du prix du diamant. Par contre, la covariance entre le prix et la profondeur du diamant est négative, donc les deux évoluent en sens inverse.

3.2. La corrélation



Le prix est fortement correlé au carat et aux caractéristiques physiques x, y et z du diamant. Par contre, il est faiblement correlé à la profondeur et à la largeur du somment du diamant. Par ailleurs, les variables x, y, z et carat sont également fortement correlées entre elles.

IV- MODÈLE DE REGRESSION

D'entrée de jeu, nous allons diviser notre dataset en 2 parties: train et test.

1. Modèle de régression simple

```
Call:
lm(formula = price ~ carat, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6358 -0.2033 -0.0043  0.1355  3.1817

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.744e-18  1.815e-03     0.0        1
carat        9.214e-01  1.815e-03   507.7   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3886 on 45847 degrees of freedom
Multiple R-squared:  0.849, Adjusted R-squared:  0.849
F-statistic: 2.578e+05 on 1 and 45847 DF,  p-value: < 2.2e-16
```

- Le coefficient de carat est estimé à 9.215e-01, ce qui signifie qu'une augmentation d'une unité de carat est associée à une augmentation de 9.215e-01 unités dans le prix estimé.
- Le R^2 ajusté, qui prend en compte le nombre de variables explicatives, révèle que le modèle explique 84.91% du prix du diamant.
- La p-value associée à la statistique F est très faible ($< 2.2e\text{-}16$), ce qui indique que le modèle dans son ensemble est statistiquement significatif. Le modèle est donc pertinent.

2. Modèle Saturé

```
Call:
lm(formula = price ~ carat + depth + table + x + y + z, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7271 -0.1539 -0.0130  0.0867  3.1903

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.518e-16  1.752e-03   0.000    1.000
carat        1.286e+00  8.304e-03 154.857  < 2e-16 ***
depth       -7.339e-02  2.130e-03 -34.455  < 2e-16 ***
table       -5.715e-02  1.879e-03 -30.411  < 2e-16 ***
x           -4.605e-01  1.709e-02 -26.947  < 2e-16 ***
y            9.085e-02  1.406e-02   6.462 1.04e-10 ***
z            9.638e-03  8.182e-03   1.178    0.239
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3751 on 45842 degrees of freedom
Multiple R-squared:  0.8593,    Adjusted R-squared:  0.8593
F-statistic: 4.667e+04 on 6 and 45842 DF,  p-value: < 2.2e-16


    carat     depth     table         x         y         z
```

```
22.471945  1.478636  1.151058 95.183888 64.411981 21.815641
```

- Toute chose étant égale par ailleurs, le carat, le depth, la table et la caractéristique x ont un effet significatif sur le prix (p-value<2e-16).
- Le modèle explique environ 85.86% de la variabilité du prix des diamants, ce qui est très élevé. Cela suggère que le modèle est globalement efficace pour prédire le prix.
- La p-value associée très faible indique que le modèle est statistiquement significatif et pertinent.
- Le modèle est statistiquement robuste et efficace pour prédire le prix des diamants en se basant sur les variables incluses. Cependant, l'interprétation de certains coefficients (notamment pour les dimensions x, y, et z) nécessite une attention particulière, car ils sont affectés par la multicollinéarité.

3. Estimation de la performance du modèle: le BIC

Etant donné que nous avions des variables fortement correlées entre elles dans le modèle précédent, nous allons procéder à une sélection de variables avec le critère Bayesian Information Criterion (BIC). Il est parcimonieux et permet d'obtenir un modèle autant performant que les autres modèles mais avec le moins de variables.

```
Start:  AIC=-89849.94
price ~ carat + depth + table + x + y + z

        Df Sum of Sq    RSS    AIC
- z      1       0.2 6449.8 -89859
<none>               6449.6 -89850
- y      1       5.9 6455.5 -89819
- x      1     102.2 6551.8 -89140
- table  1     130.1 6579.7 -88945
- depth  1     167.0 6616.7 -88688
- carat  1    3373.9 9823.5 -70569

Step:  AIC=-89859.28
price ~ carat + depth + table + x + y

        Df Sum of Sq    RSS    AIC
<none>               6449.8 -89859
+ z      1       0.2 6449.6 -89850
- y      1       6.2 6456.0 -89826
- x      1     115.0 6564.8 -89060
```

```
- table  1     130.3 6580.1 -88953
- depth  1     204.0 6653.8 -88442
- carat  1    3375.2 9825.0 -70573


Call:
lm(formula = price ~ carat + depth + table + x + y, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7250 -0.1539 -0.0130  0.0867  3.1903

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.533e-16  1.752e-03   0.000        1
carat        1.286e+00  8.303e-03 154.886  < 2e-16 ***
depth       -7.225e-02  1.897e-03 -38.079  < 2e-16 ***
table       -5.719e-02  1.879e-03 -30.431  < 2e-16 ***
x           -4.530e-01  1.585e-02 -28.586  < 2e-16 ***
y            9.258e-02  1.398e-02   6.621 3.61e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3751 on 45843 degrees of freedom
Multiple R-squared:  0.8593,    Adjusted R-squared:  0.8593
F-statistic: 5.601e+04 on 5 and 45843 DF,  p-value: < 2.2e-16
```

Le critère BIC suggère que les variables les plus pertinentes de ce modèle sont carat, depth, table, et x. Le modèle constitué de ces 4 variables est significatif (p-value: $<$ 2.2e-16) et explique 85.86% de la variation du prix des diamants, tout comme le modèle saturé.

4. Prédiction

Pour la prédiction, nous allons utiliser le modèle final qui a été fourni grâce à la sélection effectuée avec le critère BIC. Et nous avons obtenu les prédictions suivantes pour les 06 pemiers tests :

```
       1         2         3         4         5         6
1139.1889 5559.4994 1127.5325 3749.5117  694.9732  596.2107
```

5. Axes d'amélioration

Pour améliorer ce modèle, on pourrait envisager d'exclure ou de transformer certaines variables, d'ajouter des interactions si pertinentes, ou d'explorer des modèles non linéaires ou des méthodes d'ensemble si la complexité des relations le justifie.

CONCLUSION GÉNÉRALE

Le modèle de régression linéaire ajusté sur le dataset des diamants démontre une capacité élevée à prédire le prix des diamants, avec un R^2 ajusté de 0.8593, indiquant que le modèle explique environ 86% de la variabilité du prix à partir des variables sélectionnées telles que le carat, la profondeur, la table et la dimension x. Les coefficients significatifs pour toutes ces variables suggèrent des relations fortes et statistiquement significatives avec le prix. Les erreurs résiduelles sont relativement faibles, renforçant la fiabilité des prédictions du modèle. Ce modèle peut donc servir efficacement les professionnels de l'industrie diamantaire pour évaluer de manière précise la valeur des diamants.