Projet : Logiciels Statistiques

Étude du prix des diamants

Encadrants : M. Jérôme Collet et M. Renaud Mozet.

Rédigé et présenté par : MABROUK Skandar, MOULIN Louise, GRAOUI Mehdi.

1. Table des matières

2. Introduction :

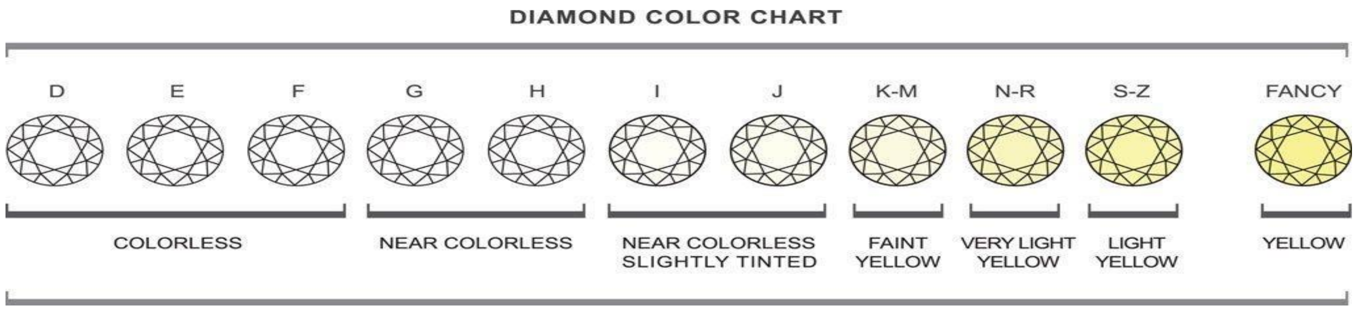
L'objectif est de construire un modèle de valorisation raisonnable pour les diamants basé sur les données relatives à leur poids (en carats), leur couleur (soit D, E, F, G, H ou I) et leur clarté (soit SI, VVS1, VVS2, VS1 ou VS2). Nous allons dans un premier temps analyser les variables du jeu de données ainsi que leurs liaisons avec celles que l'on veut expliquer : le prix (variable « price »). Ceci nous permettra de proposer un modèle de régression simple puis multiple permettant de prédire le prix d'un diamant. Enfin, nous terminerons ce rapport par une Analyse en Composantes Principales suivie d'une Classification Ascendante Hiérarchique.

3. Analyse descriptive :

1. Le Jeu de données

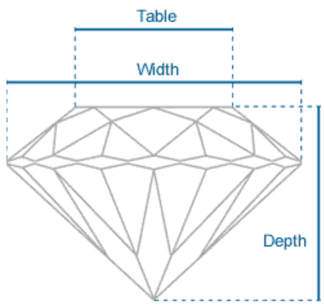
Nous avons un total de 53940 observations correspondant chacune aux caractéristiques d'un diamant et comprenant au total 10 dont 3 qualitatives et 7 quantitatives :

- Le carat correspond à une unité de masse (0.20 gramme pour 1 carat). C'est une variable continue (domaine de définition $\mathbb{R}^{+}$).
- Le prix (Price) qui désigne le prix du diamant.
- La clarté (variable « Clarity ») permet de classer les diamants en fonction de leur degré de pureté. C'est une variable qualitative. Ici nous avons les classements allant du plus pur au moins pur : VVS1 (Very very small inclusion), VVS2, VS1 (Very small inclusion), VS2, SI1 (Small Inclusion), SI2.
- La couleur (variable « color ») : dans le cas du diamant, chaque couleur est associée à une lettre comme figuré ci-dessous. Nous pouvons remarquer que les diamants D, E, F sont dits incolores, c'est-à-dire que posés sur une feuille de papier ils sont parfaitement transparents. Puis plus nous descendons dans l'alphabet, plus le diamant est teinté. Notons que la variable color est donc une variable qualitative.



- La coupe (variable « cut ») indique la qualité de la taille. Un diamant bien taillé amène à une bonne dispersion de la lumière le traversant, une brillance élevée et un bon scintillement. C'est une variable qualitative ordonnée aux modalités suivantes : Assez bonne (Fair), Bonne (Good), Très bonne (Very Good), Excellente (Excellent), Idéale (Ideal).

- La « Depth » n'est pas la hauteur du diamant à proprement parler mais le rapport Depth/Width comme indiqué sur le schéma ci-contre. La donnée Depth est donc un pourcentage ainsi qu'une variable continue($\mathbb{R}^+$)



- La « Table » représente la longueur de la facette plate sur le dessus en déci-millimètres. C'est une variable discrète ($\mathbb{N}$).
- x, y et z représentent respectivement la longueur, la largeur et la profondeur du diamant en millimètres. Ce sont des variables continues ($\mathbb{R}^+$).

2. Importation, Tableau démonstratif, valeurs aberrantes

La première étape est de mettre notre jeu de données sous forme d'un tableau pour plus de lisibilité. On crée une bibliothèque que l'on appelle diamond : libname diamonds "&chemin." ;

Lors de l'importation des données nous avons utilisé l'option « DSD » accompagné du « ~ » après le nom de la variable « Cut » comme elle inclut un séparateur.

```
data diamonds ;
      infile "&chemin./diamonds.txt"
      firstobs=2 lrecl=1500 dsd dlm='20'x;
      input number $ carat cut ~$11. color $
      clarity $ depth table price x y z;
 run ;
```

<u>Tableau démonstratif du data set donnant les 15 premières observations</u>

Obs.	number	carat	cut	color	clarity	depth	table	price	x	y	z
1	1	0.23	"Ideal"	E	SI2	61.5	55	326	3.95	3.98	2.43
2	2	0.21	"Premium"	E	SI1	59.8	61	326	3.89	3.84	2.31
3	3	0.23	"Good"	E	VS1	56.9	65	327	4.05	4.07	2.31
4	4	0.29	"Premium"	I	VS2	62.4	58	334	4.20	4.23	2.63
5	5	0.31	"Good"	J	SI2	63.3	58	335	4.34	4.35	2.75
6	6	0.24	"Very Good"	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	7	0.24	"Very Good"	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	8	0.26	"Very Good"	H	SI1	61.9	55	337	4.07	4.11	2.53
9	9	0.22	"Fair"	E	VS2	65.1	61	337	3.87	3.78	2.49
10	10	0.23	"Very Good"	H	VS1	59.4	61	338	4.00	4.05	2.39
11	11	0.30	"Good"	J	SI1	64.0	55	339	4.25	4.28	2.73
12	12	0.23	"Ideal"	J	VS1	62.8	56	340	3.93	3.90	2.46
13	13	0.22	"Premium"	F	SI1	60.4	61	342	3.88	3.84	2.33
14	14	0.31	"Ideal"	J	SI2	62.2	54	344	4.35	4.37	2.71
15	15	0.20	"Premium"	E	SI2	60.2	62	345	3.79	3.75	2.27

On remarque qu'il y a une colonne supplémentaire « number » qui nous est inutile. En effet, SAS ajoute automatiquement une colonne qui référence le numéro de l'observation. De plus, grâce à la procédure means, nous détectons la présence des valeurs aberrantes. Par exemple, un diamant de 0mm en hauteur, largeur, longueur n'est pas logique.

Variable	N	Moyenne	Ec-type	Minimum	Maximum
carat	53916	0.7976641	0.4737526	0.2000000	5.0100000
depth	53916	61.7496235	1.4322674	43.0000000	79.0000000
table	53916	57.4562430	2.2282310	43.0000000	79.0000000
price	53916	3930.74	3987.04	326.0000000	18823.00
x	53916	5.7315572	1.1193569	3.7300000	10.7400000
y	53916	5.7333806	1.1112267	3.6800000	10.5400000
z	53916	3.5393844	0.6916027	1.0700000	6.9800000

Maj procédure Means 1

Moyennes

La procédure MEANS

Variable	N	Moyenne	Ec-type	Minimum	Maximum
carat	53940	0.7979397	0.4740112	0.2000000	5.0100000
depth	53940	61.7494049	1.4326213	43.0000000	79.0000000
table	53940	57.4571839	2.2344906	43.0000000	95.0000000
price	53940	3932.80	3989.44	326.0000000	18823.00
x	53940	5.7311572	1.1217607	0	10.7400000
y	53940	5.7345260	1.1421347	0	58.9000000
z	53940	3.5387338	0.7056988	0	31.8000000

Procédure Means 1

Pour résoudre ce problème, voici le morceau de code

```
data diamonds2;
      set diamonds;
      drop number ;
      if x=0 or y=0 or z=0 then
delete;
run ;
```

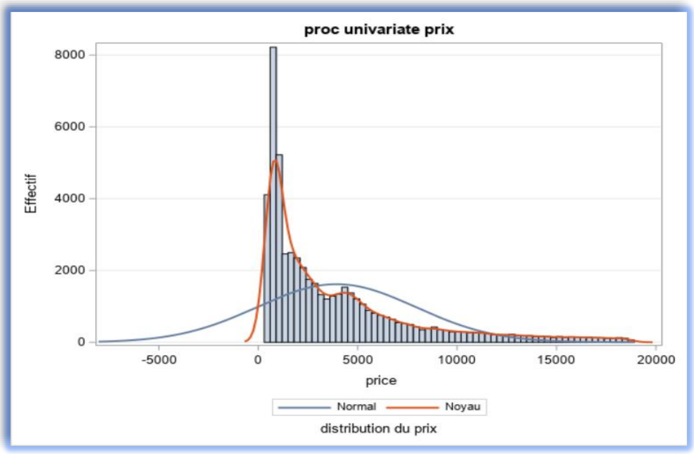
4. Analyse des variables du jeu de données

Cette partie consiste à établir des analyses univariées pour la variable que l'on souhaite expliquer, le prix (« Price »), ainsi que les autres variables du jeu de données. On va étudier si ces variables suivent une loi théorique usuelle.

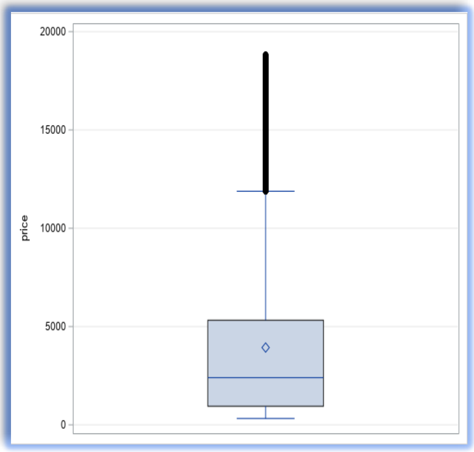
3. Le Prix :

Nous constatons que la variable prix n'est pas distribué normalement. On peut le remarquer graphiquement sur la figure Price normality 1 mais aussi quand on effectue le test de Kolmogorov-Smirnov ou Anderson-Darling. En effet, on pose l'hypothèse H0 :« La distribution suit une loi normale » et on remarque que la p-value est inférieur à $\alpha = 0.05$. On décide donc de rejeter cette hypothèse.

La boite à moustache (Price boxplot 1) nous confirme aussi que le prix ne suit pas une loi normale. La majorité des diamants se situent dans la fourchette de prix suivante : [3000$;5323.5$]. Sachant que le min est de 326$ et le max est d'environ 18823$ avec une moyenne de 3932.80$ la dispersion donnée est de 15898405.4, ce qui est trop important pour que cela nous donne une loi normale.



Price normality 1



Price boxplot 1

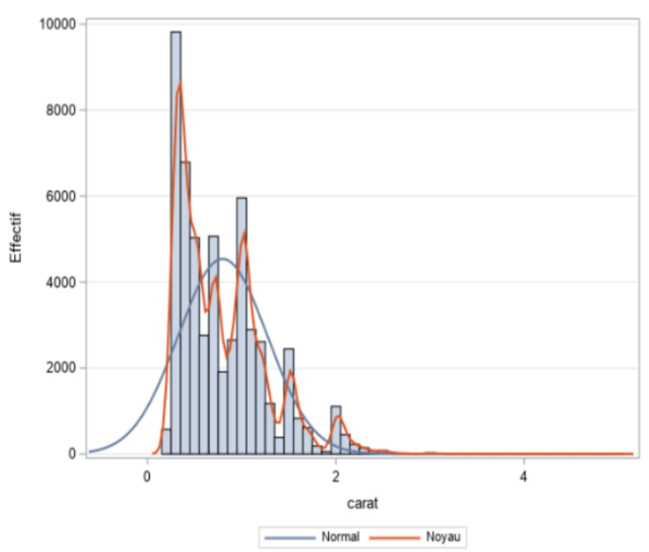
4. Carat :

La variable carat ne suit pas non plus une loi normale. On a beaucoup de petits diamants (de l'ordre du 0,2-0,4 carat) ceci s'expliquant par une forte demande pour de plus petits diamants que les bijoutiers peuvent utiliser pour des boucles d'oreilles ou pour orner une autre pierre. On constate aussi qu'une part importante des diamants sont de 1 carat, correspondant aux bagues de fiançailles mais aussi pour 2 carats. Ainsi nous pourrions modéliser la distribution des carats en trois lois normales centrées respectivement en 0.5, 1.5, et 2.

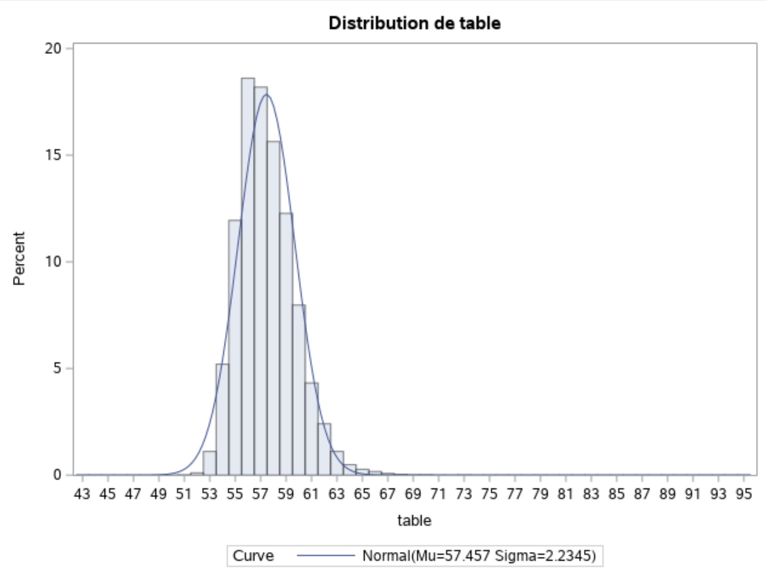
Tests de tendance centrale : Mu0=0				
Test	Statistique		p-value	
t de Student	t	390.9515	Pr > |t|	<.0001
Signe	M	26960	Pr >= |M|	<.0001
Rang signé	S	7.2686E8	Pr >= |S|	<.0001

Tests de normalité				
Test	Statistique		p-value	
Kolmogorov-Smirnov	D	0.122799	Pr > D	<0.0100
Cramer-von Mises	W-Sq	228.5219	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1528.724	Pr > A-Sq	<0.0050

Quantiles (Définition 5)	
Niveau	Quantile
100Max 100%	5.01
99%	2.18
95%	1.70
90%	1.51
75% Q3	1.04
50% Médiane	0.70
25% Q1	0.40
10%	0.31
5%	0.30
1%	0.24
0% Min	0.20



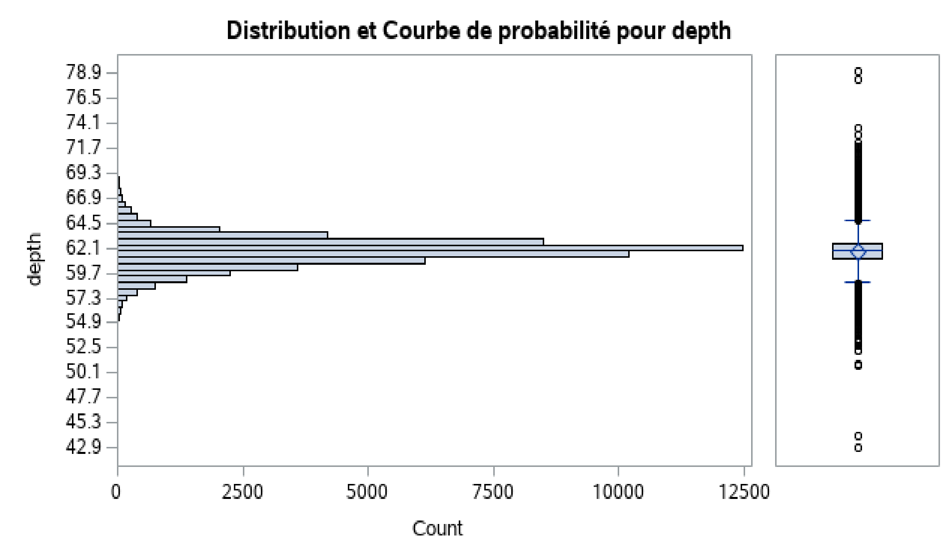
5. La Table



Bien que graphiquement la variable « Table » semble suivre une loi normale, le test de Kolmogorov-Smirnov nous indique qu'il n'en est rien.

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistique		p-value	
Kolmogorov-Smirnov	D	0.132245	Pr > D	<0.010
Cramer-von Mises	W-Sq	125.852994	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	695.641346	Pr > A-Sq	<0.005

6. La Depth :



La boite à moustache de la distribution de probabilité pour la variable « Depth » nous montre 4 valeurs extrêmes. On constate aussi que cette variable ne suit pas non plus une loi normale.

Tests de normalité				
Test	Statistique		p-value	
Kolmogorov-Smirnov	D	0.075871	Pr > D	<0.0100
Cramer-von Mises	W-Sq	88.14724	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	502.8321	Pr > A-Sq	<0.0050

Pour les variables « x », « y » et « z », elles aussi semblent suivre une loi normale en regardant leurs distributions. Mais comme les autres variables ci-dessus, les tests de normalité indiquent le contraire. Ainsi, aucune des variables quantitatives de notre jeu de données suivent une loi normale.

5. Analyse de la liaison entre la variable Prix et les autres variables du jeu de données

En utilisant la procédure corr qui étudie la corrélation entre la variable prix et les autres on obtient des liens significatifs entre la variable prix et l'ensemble {carat, x, y, z}.

Comme le montre le tableau suivant, les coefficients qui sont très proches de 1 sont les plus significatifs et démontrent un lien majeur avec la variable prix. Pour les variables qualitatives on étudiera leur lien par la suite.

Coefficients de corrélation de Pearson, N = 53916	
	price
carat	0.92158
depth	-0.01056
table	0.12684
x	0.88721
y	0.88881
z	0.88210

Pearson test 1 : Plus la valeur absolue du coefficient est importante, plus la relation linéaire entre les variables est forte. Pour la corrélation de Pearson, une valeur absolue de 1 indique une relation linéaire parfaite.

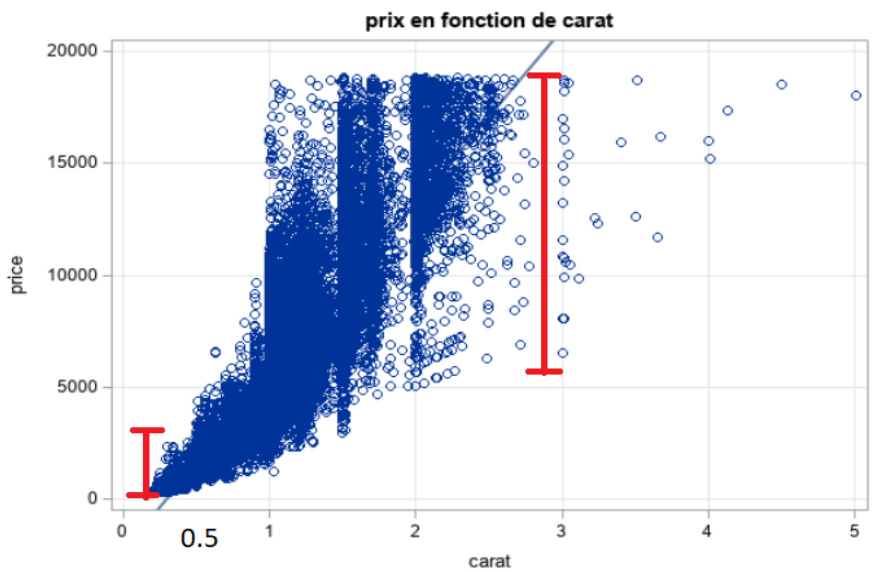
Une corrélation proche de 0 indique l'absence de relation linéaire entre les variables.

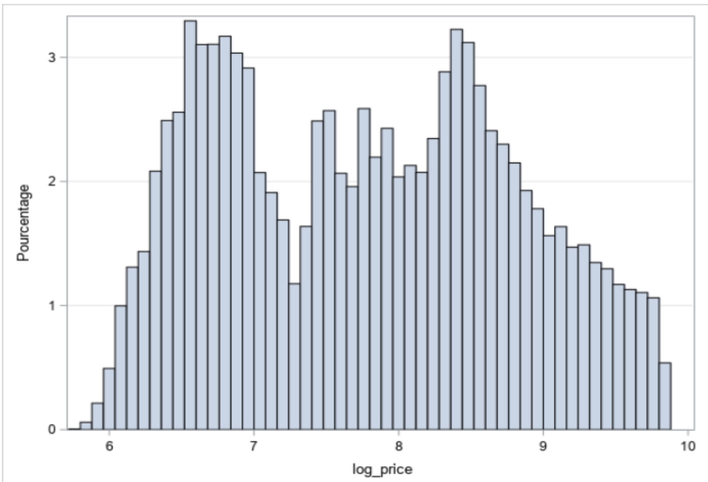
Nous pouvons donc en conclure que le prix du diamant va surtout dépendre de son carat. Les variables « x », « y » et « z » étant liées au carat par la relation : $m_{diamant} = \rho xyz$

Avec ρ : la masse volumique du diamant. Il est donc normal d'avoir là aussi un coefficient de corrélation proche de 1.

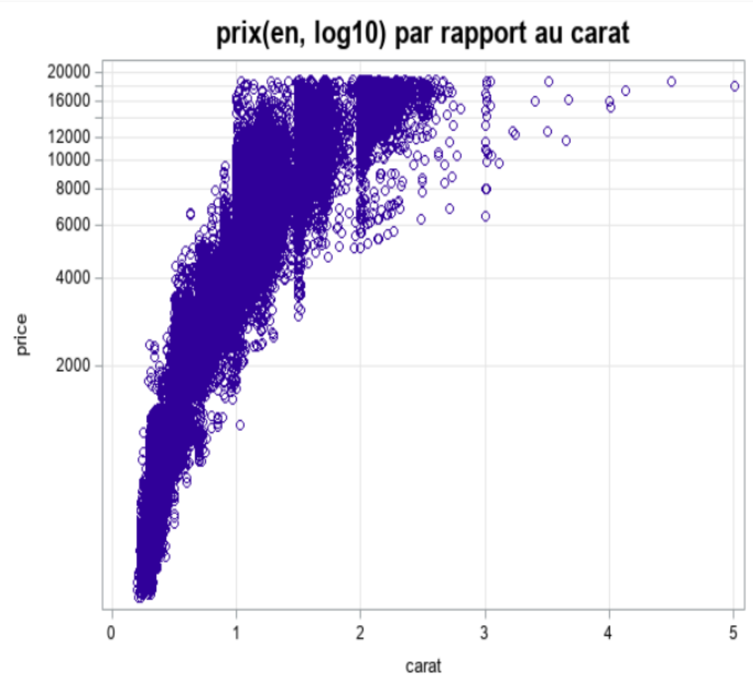
7. Prix et Carat

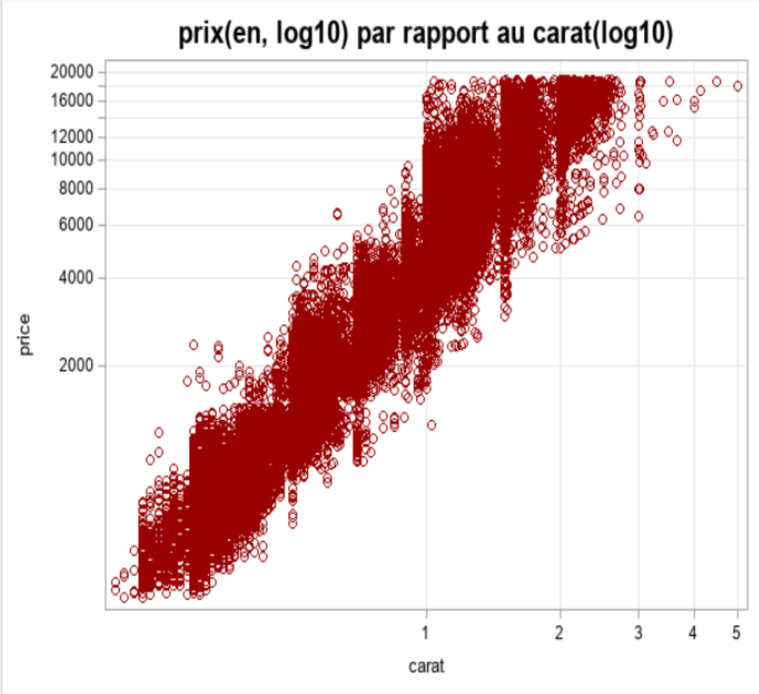
Comme le montre le graphique suivant, le prix augmente exponentiellement avec le carat. On observe aussi une dispersion plus importante du prix lorsque le poids du diamant augmente. Ceci peut s'expliquer par la rareté d'un diamant plus « gros ». En effet, ceux-ci étant plus rare on a une fourchette de prix plus large.





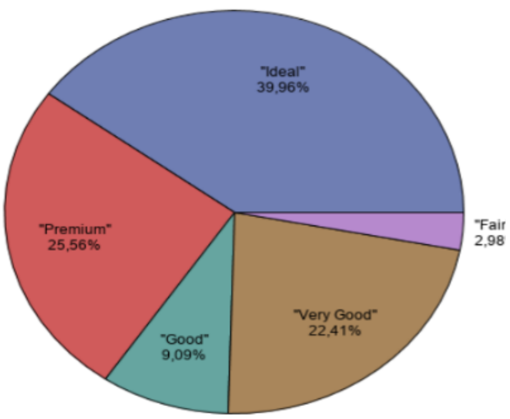
Nous avons vu précédemment que le prix ne suivait pas une loi normale. En revanche, si on met cette variable en échelle log on constate que celle-ci est plus proches d'une courbe en cloche de distribution normale. Nous pouvons même voir une bimodalité, ce qui est cohérent avec notre hypothèse qu'un diamant est acheté soit pour faire une bague de fiançailles quand il est de plus d'un carat soit utilisé pour décorer d'autres bijoux lorsqu'il est plus petit.



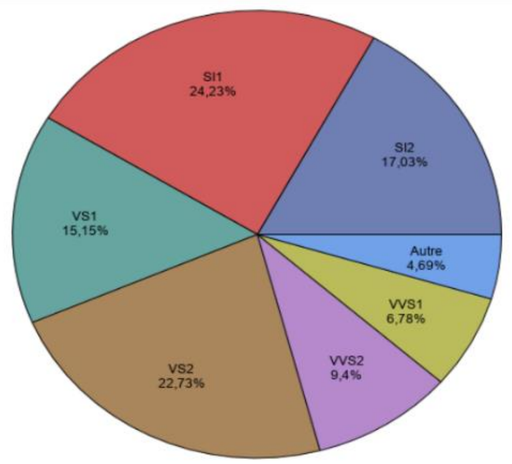


8. Prix et répartition avec les variables qualitatives : (coupe, clarté et couleur)

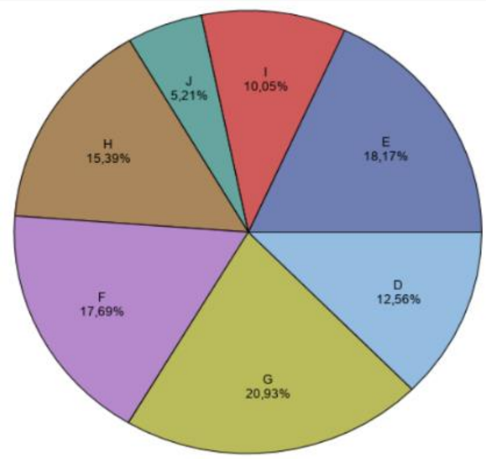
La majorité des diamants sont « idéal », « Premium » et « Very good » en termes de coupe : en termes de clarté ils sont « SI1 » « VS2 » «VS1 » « SI2 », et en termes de couleur « G » « F » « D » « E » « H ».



Coupe 1

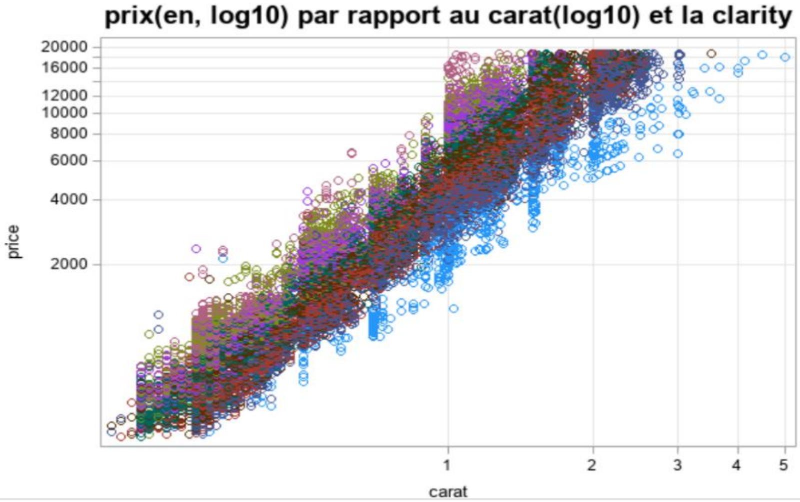


Clarté 1

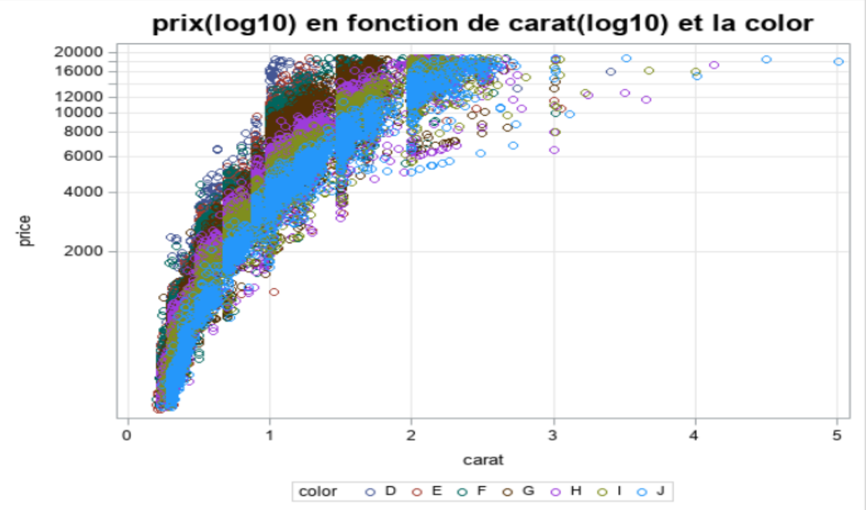


couleur 1

9. Prix, carat et coupe : la coupe joue-t-elle un rôle sur le prix d'un diamant ?



La clarté semble expliquer une grande partie de la variation du prix. Cela se voit d'autant plus après avoir ajouté un code couleur à notre graphique. En effet, si nous prenons un poids en carat constant on observe très clairement que plus le prix augmente plus la qualité de la clarté aussi (vvs1 et 2 étant plus haut tandis que I1 est tout en bas).



La couleur semble aussi expliquer une partie de la variance du prix. Tout comme nous l'avons vu avec la variable de clarté, si l'on se place à un poids en carat constant on voit que les couleurs D, E, F rendent le diamant plus couteux que s'il était H, I, J. Ainsi plus un diamant est transparent plus il est cher.

6. Construction de modèles de régressions simples

La formule d'un modèle linéaire simple est la suivante $Y = \beta_0 + \beta_1 X + \varepsilon$

- Y est la valeur prédite
- β_0 est l'intersection (Intercept)
- β_1 est le coefficient de régression
- X est la variable indépendante qui influence Y
- ε est l'erreur de l'estimation.

10. Modèle n°1 : Prix en fonction de carat

Le modèle qu'on obtient est le suivant : Prix = -2255.85 + 7755.88 * carat.

Modèle : MODEL1
Variable dépendante : price

Nb d'observations lues	53916
Nb d'obs. utilisées	53916

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	7.279056E11	7.279056E11	303854	<.0001
Erreur	53914	1.291553E11	2395580		
Total sommes corrigées	53915	8.570609E11			

Root MSE	1547.76609	R carré	0.8493
Moyenne dépendante	3930.73509	R car. ajust.	0.8493
Coeff Var	39.37600		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	-2255.85320	13.05348	-172.82	<.0001
carat	1	7755.88113	14.07016	551.23	<.0001

1. Interprétation du test de la signification globale de la régression

La statistique F = (MSR / MSE) = (carrée Moyen de régression / erreur quadratique moyenne) = 303853 indique que globalement le modèle avec le régresseur carat améliore la prévision du prix, par rapport à la moyenne seule dans le modèle

2. Interprétation des estimations des paramètres

L’estimateur de β_0 a pour valeur -2255,8532. Son écart type vaut 13,05348. La statistique de Test de student t-value = -2255,8532/13,05348 = -172,82 et sa p-valeur associée est bien inférieur au seuil 0,05. Donc on rejette l’hypthèse nulle « $\beta_0 = 0$ » avec une grande confiance. Même raisonnement pour l’estimateur de β_1 qui a pour valeur 7755.88113.

Le R² qui varie entre 0 et 1, mesure la proportion de variation totale de Y autour de la moyenne expliquée par la régression, c’est-à-dire prise en compte par le modèle. Plus R² se rapproche de la valeur 1, meilleure est l'adéquation du modèle aux données. Dans notre cas, on obtient un R² = 0.85 et RMSE (L’erreur quadratique moyenne) = 1547.7

Remarque ! Dans le cas de la régression simple la statistique de test de l’estimateur de β_1 est lié à F=(t-valeur) ².

```
proc reg data=WORK.DIAMONDS2 alpha=0.05 ;
model price=carat /;
        run; quit;
```

11. Modèle n°2 : log (Prix) en fonction de log(carat)

Root MSE	0.26260	R carré	0.9330
Moyenne dépendante	7.78634	R car. ajust.	0.9330
Coeff Var	3.37263		

Paramètres estimés					
Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	8.44875	0.00137	6189.12	<.0001
log_carat	1	1.67594	0.00193	866.48	<.0001

Log(prix) = 8.44 + 1.67 *log(carat).

On constate que le coefficient β_1 vaut 1.67594. Une interprétation de β_1 dans le cadre d’une régression log-log nous dit que si on augmente le carat d’un diamant de 1% on attend que le prix soit augmenté d’environ 1.68 %. Notez que le modèle estime le log (prix) et non le prix du diamant. Pour convertir le log (prix) estimé en prix, il doit y avoir une transformation. Cette dernière traite le log (prix) comme un exposant de la base e : $e^{\log(\text{prix})} = \text{prix}$. D’où le modèle : prix = exp (8.44 + 1.67 *log(carat)).

12. Conclusion

Le modèle linéaire simple le plus valorisant est log(prix) en fonction de log(carat) en effet il est caractérisé par un RMSE faible, un R carrée plus élevé 0.933 (93 % de la variabilité de prix est expliquée par le modèle), ainsi sa valeur t-test est élevée, les p-valeurs sont très faibles cela explique que nous n’avons pas trouvé ses résultats par hasard et qu’elles sont significatives.

7. Construction d’un modèle de régression multiple

Au vu des résultats précédent, on cherche maintenant à définir un modèle linéaire afin de prédire le prix d’un diamant. On souhaite expliquer la variable **prix** (Y) qui peut être modélisé par l’équation suivante :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_7 X_7$$

Avec :

- β_0 qui correspond à l'ordonnée à l'origine du modèle.
- les β_i , i ∈ [1,7] qui sont des coefficients associés à la i -ème variable explicative.

• les X_i, i ∈ [1,7] qui correspondent aux variables de notre étude : ce sont les variables explicatives de notre modèle. On précise que nous avons utilisé les variables qualitatives **cut** qui possède cinq modalités, **color** qui en possède sept et **clarity** qui en possède huit. De plus, on a pris en compte toutes les variables quantitatives suivantes **carat**, **x**, **y**, **z**, **depth** et **table.**

Mais avant de cela…

13. Modèle 1(Sans utiliser les variables qualitatives)

Nous allons d'abord effectuer une régression multiple seulement sur les variables quantitatives pour notre curiosité. Ainsi, nous pourrons comparer par la suite si l'ajout des variables qualitatives au modèle est pertinent.

```
proc glmselect data=diamonds2;
class cut color clarity / param=glm;
model price=depth table x y z carat / showpvalues selection=none;
run;
```

Root MSE	1493.73339	R carré	0.8597
Moyenne dépendante	3930.99323	R car. ajust.	0.8597
Coeff Var	37.99888		

Root MSE	1493.87127	R carré	0.8596
Moyenne dépendante	3930.99323	R car. ajust.	0.8596
Coeff Var	38.00239		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	21317	456.63887	46.68	<.0001
carat	1	10989	66.97835	164.06	<.0001
x	1	-1412.96480	45.78136	-30.86	<.0001
y	1	88.16615	25.68272	3.43	0.0006
z	1	-46.57809	50.09115	-0.93	0.3524
depth	1	-203.02009	5.65793	-35.88	<.0001
table	1	-101.93636	3.07894	-33.11	<.0001

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	21547	422.34755	51.02	<.0001
carat	1	10992	66.94810	164.19	<.0001
x	1	-1355.40029	28.32523	-47.85	<.0001
depth	1	-206.30077	4.85712	-42.47	<.0001
table	1	-102.28632	3.07634	-33.25	<.0001

Nous sommes partis initialement du tableau de gauche, et après avoir éliminé au fur et à mesure les variables qui sont non significatives (>=0,05) d'après notre test de Student ($Pr > |t|$), nous arrivons au tableau de droite dont toutes variables sont significatives (<0,05) et qui explique notre prix d'une façon la plus naïve possible avec un R carré = 0,86.

14. Modèle 2(Sans utiliser les variables qualitatives)

Root MSE	0.25790	R carré	0.9354
Moyenne dépendante	7.78639	R car. ajust.	0.9354
Coeff Var	3.31216		

Root MSE	0.25791	R carré	0.9354
Moyenne dépendante	7.78639	R car. ajust.	0.9354
Coeff Var	3.31228		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	7.97095	0.40272	19.79	<.0001
logcarat	1	1.20005	0.05498	21.83	<.0001
logx	1	0.83595	0.15688	5.33	<.0001
logy	1	0.47624	0.08219	5.79	<.0001
logz	1	0.17236	0.07796	2.21	0.0271
depth	1	-0.02008	0.00183	-10.95	<.0001
table	1	-0.01657	0.00056947	-29.09	<.0001

Paramètres estimés

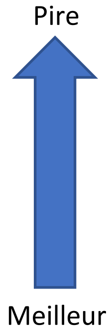
Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	7.85174	0.39910	19.67	<.0001
logcarat	1	1.21812	0.05437	22.40	<.0001
logx	1	0.89982	0.15421	5.84	<.0001
logy	1	0.53007	0.07850	6.75	<.0001
depth	1	-0.01776	0.00150	-11.81	<.0001
table	1	-0.01667	0.00056774	-29.35	<.0001

Sous le même principe que le modèle 1, mais cette fois on explique le logarithme du prix, en fonction du logaritme du carat, logarithme (x,y,z), depth et table. Et on voit qu'on a seulement la variable log(z) qui est non sinificative. On voit aussi que l'on a gagné en précision grâce à l'augmentation du R carré = 0,94 et de la diminution de la racine MSE = 0,26. Le deuxiéme modèle est donc meilleur que le premier.

15. Modèle 3(en utilisant les variables qualitatives)

Pour ce modèle, nous avons décidé d'ajouter les variables qualitatives clarity, color et cut en tant que variables quantitatives. En effet, nous connaissons déjà l'ordre de ces variables (pire-meilleur) et nous avons décidé de donner un poids plus important aux meilleurs niveaux dans les nouvelles variables : claritynum, colornum et cutnum respectivement pour clarity, color et cut.

Cutnum/Claritynum/Colornum	Clarity	Color	Cut
1	I1	J	Fair
2	SI2	I	Good
3	SI1	H	Very Good
4	VS2	G	Premium
5	VS1	F	Ideal
6	VVS2	E	
7	VVS1	D	
8	IF		



Root MSE	0.14578	R carré	0.9794
Moyenne dépendante	7.78639	R car. ajust.	0.9794
Coeff Var	1.87219		

Paramètres estimés					
Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	4.60001	0.22792	20.18	<.0001
logcarat	1	1.40474	0.03119	45.04	<.0001
logx	1	1.43748	0.08924	16.11	<.0001
logy	1	-0.09710	0.04654	-2.09	0.0369
logz	1	0.08385	0.04407	1.90	0.0571
depth	1	0.00523	0.00105	4.99	<.0001
table	1	0.00074764	0.00036798	2.03	0.0422
cutnum	1	0.02925	0.00068872	42.47	<.0001
colornum	1	0.07786	0.00038655	201.42	<.0001
claritynum	1	0.12317	0.00042345	290.86	<.0001

Root MSE	0.14588	R carré	0.9793
Moyenne dépendante	7.78639	R car. ajust.	0.9793
Coeff Var	1.87348		

Paramètres estimés					
Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	5.98058	0.09626	62.13	<.0001
logcarat	1	1.59256	0.01737	91.69	<.0001
logx	1	0.85812	0.05220	16.44	<.0001
cutnum	1	0.02884	0.00060374	47.77	<.0001
colornum	1	0.07773	0.00038652	201.10	<.0001
claritynum	1	0.12293	0.00042178	291.45	<.0001

Toujours sur le même principe, nous partons du tablau de gauche pour arriver à celui de droite en suprimmant les variables non sinificatives. Nous avons ici un R carré = 0,98 et Racine MSE = 0,14.

Tout ceci nous prouve ici que nous avons un modèle multiple très perfomant, meilleur que les deux autres précédents par rappot aux critéres du R carré et de la racine du MSE. On peut donc déduire le prix Y du diamant grace à la formule suivante :

$$Y = \exp(5{,}98058 + 1{,}59256 * X1 + 0{,}85812 * X2 + 0{,}02884 * X3 + 0{,}07773 * X4 + 0{,}12293 * X5)$$

Où : X1 = logarithme du carat

X2 = logatithme de x

X3,X4,X5 = cutnum, colornum, claritynum ∈ [1,7] (se référer au tableau plus haut)

8. Exporter les données de SAS vers R :

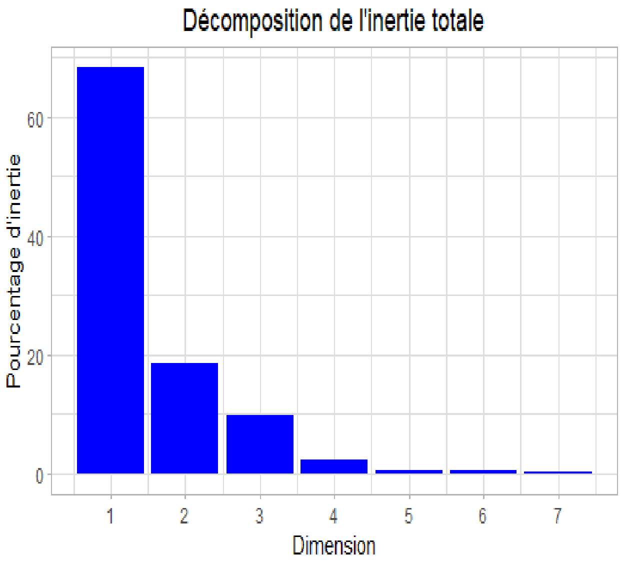
On a exporté les données de SAS vers R en utilisant la procédure proc export, nous avons décidé que le format sera csv.

```
proc export data=diamonds2 dbms=xlsx
outfile="C:\Users\skani\Desktop\diam.csv"
dbms=csv
replace;
run;
```

9. ACP/CAH

16. Analyse en Composantes Principales

L'analyse en composantes principales est une technique utile pour l'analyse statistique exploratoire des données. Elle est particulièrement utile dans le cas d'ensembles de données massives comportant de nombreux individus et variables quantitatives. Le but de l'ACP est d'identifier les directions (ou composantes principales) selon lesquelles la variation de données est maximale, c'est-à-dire réduire la dimensionnalité d'une donnée à quelques axes principaux qui peuvent être visualisées graphiquement, avec une perte minimale d'infirmations.



3. Distribution de l'inertie :

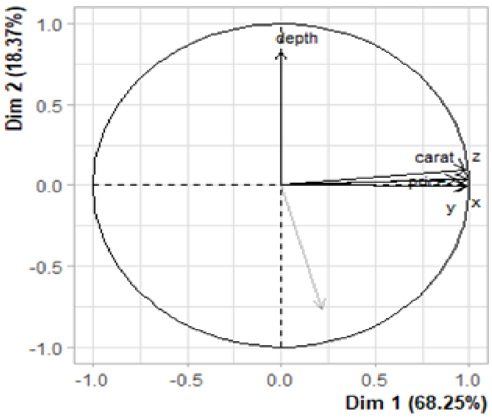
L'inertie des axes factoriels indique d'une part si les variables sont structurées et suggère d'autre part le nombre judicieux de composantes principales à étudier. Les 2 premiers axes de l'analyse (qui sont orthogonaux) expriment 86.62% de l'inertie totale du jeu de données ; cela signifie que 86.62% de la variabilité totale du nuage des individus (ou des variables) est représentée dans ce plan. C'est un pourcentage élevé, et le premier plan représente donc bien la variabilité contenue dans une très large part du jeu de données actif. Cette valeur est nettement supérieure à la valeur de référence de 29.04% (qui représente le quantile 0.95 de distributions aléatoires). Cette observation suggère que seuls les deux premiers axes sont porteurs d'une véritable information et qu'il n'est pas nécessaire d'étudier les autres dimensions.

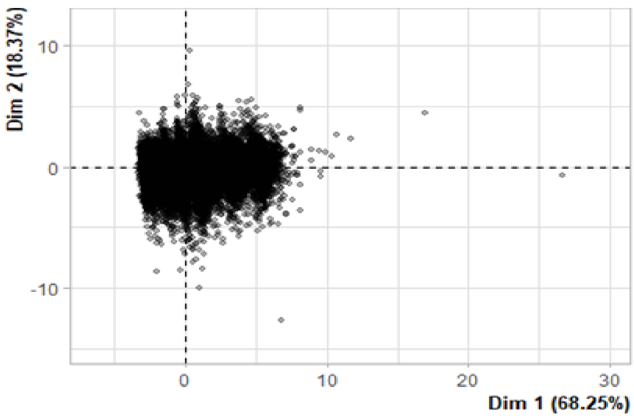
4. Valeurs Propres

On utilise le critère "absolu" : ne retenir que les axe dont les valeurs propres sont supérieures à 1 (c'est le critère de Kaiser). En effet, les deux premières valeurs propres sont 4.76 et 1.29 on les retient en négligeant les autres valeurs propres.

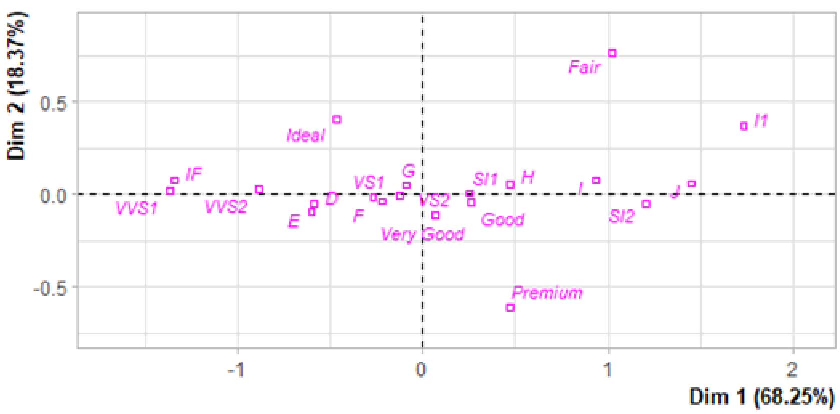
```
library(Factoshiny)
res.PCA<-PCA(diaments,quali.sup=c(3,4,5),graph=FALSE)
plot.PCA(res.PCA,choix='var',title="Graphe des variables de l'ACP")
plot.PCA(res.PCA,title="Graphe des individus de l'ACP")
```

5. Description du plan





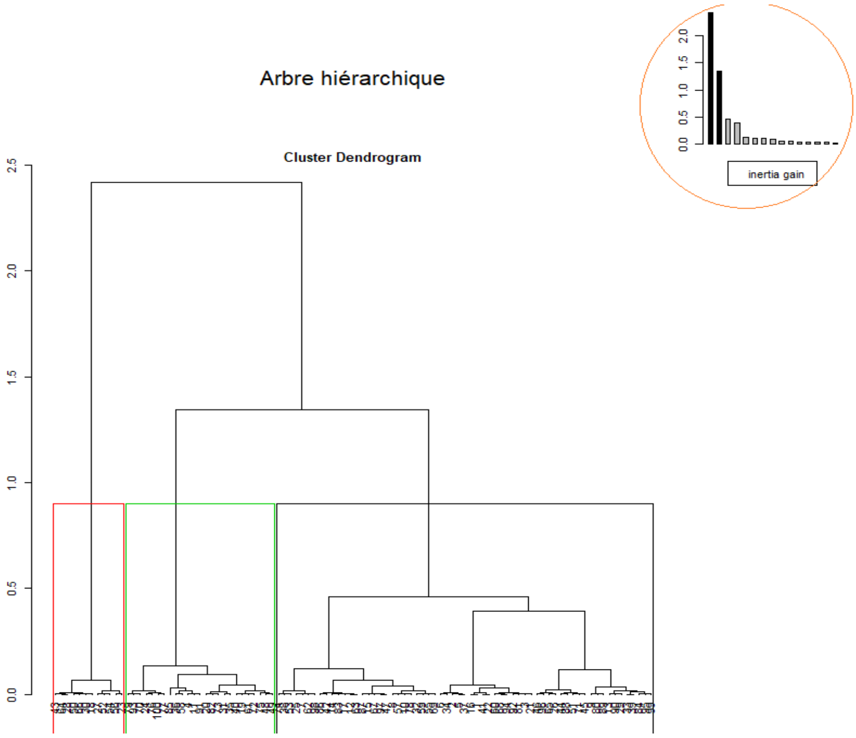
<u>La dimension 1 :</u> oppose des individus caractérisés par une coordonnée fortement positive sur l'axe (à droite du graphe) à des individus caractérisés par une coordonnée fortement négative sur l'axe (à gauche du graphe). La plupart de nos variables possèdent des fortes coordonnées positives tels que y, x, z, carat et price alors que table et Depth (coordonnée négative) ont une faible à très faible coordonnée respectivement.
→ Notons que les variables price, carat, x , y ,et z (les modalités supplémentaire liées à la couleur d'un diamant -un plus-) sont extrêmement corrélées à cette dimension et pourraient donc résumer à elles seulless la dimension 1.
<u>La dimension 2 :</u> De même, on constate que sur cet axe, toutes les variables sauf Depth possèdent une coordonnée négative.
<u>Informations sur les variables qualitatives supplémentaires</u> : on va considérer les modalités de ces variables : on effectue une projection au barycentre des individus qui prennent cette modalité.
<u>Ps</u> : Les variables supplémentaires -à titres illustratives- ne servent pas à construire les axes càd qu'elles ne servent pas à calculer les distances entre individues.



17. Classification Ascendante Hiérarchique

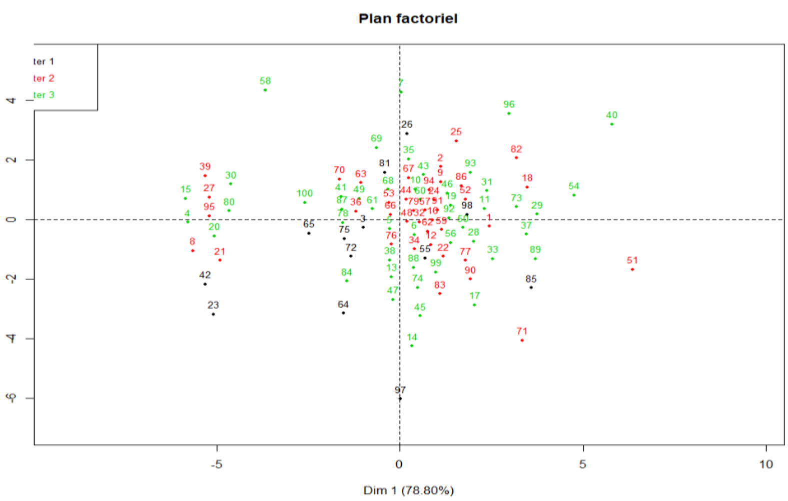
```
res.HCPC<-HCPC(res.PCA,nb.clust=3,kk=100,consol=FALSE,graph=FALSE)
plot.HCPC(res.HCPC,choice='tree',title='Arbre hiérarchique')
plot.HCPC(res.HCPC,choice='map',draw.tree=FALSE,title='Plan factoriel')
plot.HCPC(res.HCPC,choice='3D.map',ind.names=FALSE,centers.plot=FALSE,angle=60,title
='Arbre hiérarchique sur le plan factoriel')
```

La CAH organise les observations, définies par un certain nombre de variables, elles-mêmes divisées en modalités, en les regroupant de façon hiérarchique. On l'applique à un objet résultat de L'ACP précédemment, en utilisant les coordonnées des individus en conservant les 2 premières dimensions.

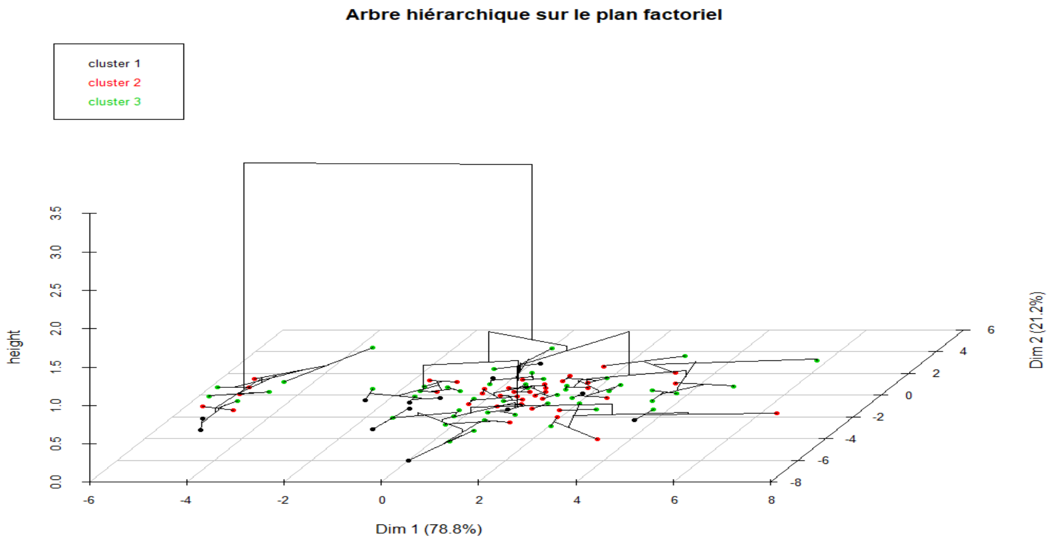


On choisit d'utiliser le prétraitement par K-means avant la classification, l'idée est alors de créer une partition grossière avec beaucoup de classes, une centaine par exemple, construire ensuite l'arbre à partir des 100 barycentres des classes, pondérés par l'effectif de la classe. Le haut de l'arbre hiérarchique est stable par rapport à une classification construites à partir de tous les individus, mais la classification va être beaucoup rapide. On utilisera la distance euclidienne, en effet notre résultat est issu d'une analyse factorielle.

Dans le premier graphe il y a un diagramme avec les gains d'inertie qui indique qu'il y a une forte perte d'inertie si on passe de deux classes à une seule, donc il ne faut pas utiliser une seule classe, même chose si on passe de 3 à 2 classes. Cependant, si on passe de 4 à 3 classes la perte d'inertie est beaucoup plus faible, on peut donc garder 3 classes ici. Un autre critère est de regarder la forme de l'arbre, on constate que l'allure de découpage convient et on garde 3 classes. Le second graphe correspond au graphe des individus sur le plan principal de l'ACP donc sur les deux premières dimensions, ces individus sont colorés en fonction de leur appartenance aux différentes classes.



On constate l'existence de quelques valeurs aberrantes, qui sont des données très éloignés des autres, peuvent influencer nos résultats. Cependant ces valeurs ne sont pas nombreuses tel que l'individu 58, 40 et 97. Mais quelques individus sur les 100 sont mal représentés par le premier plan principal et qui sont éloignés du point moyen, cela ne nous oblige pas d'ajouter un axe supplémentaire. En ajoutant de la couleur, à notre graphique on voit qu'il y a une différenciation entre les individus de la classe rouge et verte donc une forte distance entre les deux classes.



10. Conclusion générale

L'analyse descriptive du jeu de données nous a permis de voir qu'il y a une forte corrélation entre le prix et les variables suivantes : carat, x, y, z, cut, color et clarity. Cependant, nous n'avons pas pu déceler de distribution normale, ce qui n'est pas un problème dans notre cas au vu de la grande masse de données qui nous permet l'emploi de méthodes non-paramétriques performantes. Les derniers modèles de régression linéaire simple et multiple que nous avons sélectionné sont très significatifs et ils sont les meilleurs modèles que nous avons trouvés. Nous avons pu étudier la différence entre le prix estimé et le prix réel, et nous avons trouvé une belle courbe en cloche avec plus de 60% des valeurs estimés avec plus ou moins 500€ du prix réel pour la simple et plus de 75% pour la régression multiple.

Également, nous avons réalisé une ACP, qui nous a permis de confirmer la corrélation entre le prix, le carat , la longueur , la largeur et la profondeur d'un diamant. Ensuite, une CAH, celle-ci nous a permis de regrouper nos individus dans différentes classes.