ETUDE DU PRIX DES DIAMANTS

TABLE OF CONTENTS

Université de Paris

Projet encadré par Mr Collet, et présenté par Sarah Berkani et Lucie Guillaumin.

INTRODUCTION

L'objectif est 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 de clarté (soit SI, VVS1, VVS2, VS1 ou VS2). La valeur relative des différentes qualités est particulièrement intéressante, ainsi que l'impact potentiel de l'organisme de certification (soit GIA, IGI ou DRH). Par ailleurs, pour estimer correctement le modèle, il faudra transformer la variable, ce qui implique de la transformer en sens inverse pour juger de la qualité du modèle.

DESCRIPTION DES VARIABLES

- **carat** : poids en carats du diamant
- **cut** : qualité de la taille du diamant (Fair, Good, Very Good, Premium, Ideal)
- **color** : couleur du diamant, D étant le meilleur et J le pire
- **clarity** : clarté, degré d'évidence des inclusions dans le diamant : (dans l'ordre du meilleur au pire, FL = sans défaut, I3= inclusions de niveau 3) FL,IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3
- **depth** : profondeur en pourcentage : la hauteur d'un diamant, mesurée de la collette à la table, divisée par le diamètre moyen de sa gaine
- **table** : table en pourcentage : la largeur de la table du diamant exprimée en pourcentage de son diamètre moyen
- **price** : le prix du diamant
- **x**, **y** et **z** : longueur, largeur et profondeur respective du diamant en mm

VISUALISATION DE NOTRE JEU DE DONNEES

Obs.	carat	cut	color	clarity	depth	table	price	x	y	z
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	0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Goo	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Goo	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Goo	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 Goo	H	VS1	59.4	61	338	4.00	4.05	2.39

Visualisation des 10 premières observations de notre jeu de données.

On remarque tout de suite que nous avons sept variables quantitatives (**carat**, **depth**, **table**, **price**, **x**, **y** et **z**) et trois variables catégorielles (**cut**, **color** et **clarity**).

ANALYSE DESCRIPTIVE

ANALYSE DES VARIABLES QUANTITATIVES

Dans cette partie, nous nous intéressons aux variables quantitatives de notre jeu de données.
On commence alors par afficher les moyennes, les écarts-types, les minimums et les maximums de chacune de nos variables quantitatives que nous avons définies plus haut.
Ensuite, nous regardons les différentes corrélations (Pearson) de nos variables.
En effet, après étude de ce tableau nous allons pouvoir déterminer quelles variables influent entre elles.
On remarque que le prix est lié au carat du diamant, ce qui nous paraît logique. De plus, la longueur, la largeur ainsi que la profondeur du diamant sont liées entre elles, mais aussi au prix du diamant.

Statistiques simples

Variable	N	Moyenne	Ec-type	Somme	Minimum	Maximum
carat	53940	0.79794	0.47401	43041	0.20000	5.01000
depth	53940	61.74940	1.43262	3330763	43.00000	79.00000
table	53940	57.45718	2.23449	3099241	43.00000	95.00000
price	53940	3933	3989	212135217	326.00000	18823
x	53940	5.73116	1.12176	309139	0	10.74000
y	53940	5.73453	1.14213	309320	0	58.90000
z	53940	3.53873	0.70570	190879	0	31.80000

Visualisation des statistiques de base.

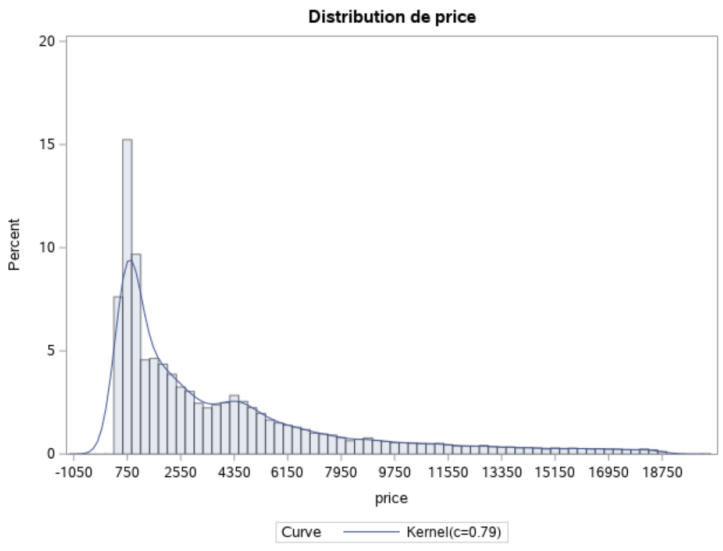
Coefficients de corrélation de Pearson, N = 53940

	carat	depth	table	price	x	y	z
carat	1.00000	0.02822	0.18162	0.92159	0.97509	0.95172	0.95339
depth	0.02822	1.00000	-0.29578	-0.01065	-0.02529	-0.02934	0.09492
table	0.18162	-0.29578	1.00000	0.12713	0.19534	0.18376	0.15093
price	0.92159	-0.01065	0.12713	1.00000	0.88444	0.86542	0.86125
x	0.97509	-0.02529	0.19534	0.88444	1.00000	0.97470	0.97077
y	0.95172	-0.02934	0.18376	0.86542	0.97470	1.00000	0.95201
z	0.95339	0.09492	0.15093	0.86125	0.97077	0.95201	1.00000

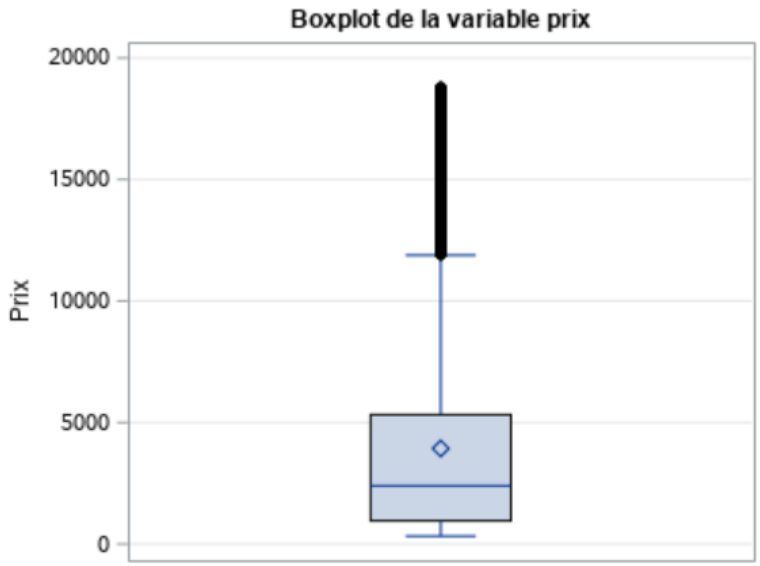
Visualisation des différentes corrélations.

LA VARIABLE *PRICE*

Le but de notre étude est d'étudier le prix des diamants en fonction de ses caractéristiques.
On s'intéresse donc dans cette partie à cette variable.
On affiche dans la figure suivante la distribution du prix. On remarque qu'elle suit une loi normale.
On affiche ensuite le boxplot de cette variable où l'on remarque que sa moyenne est de 3 933. Cela signifie que dans notre jeu de données, le prix moyen d'un diamant est de 3 933$.



Distribution de la variable price.



Boxplot de la variable price.

On note aussi que la valeur minimale d'un diamant est de 326$ et la valeur maximale de 18 823$. On voit aussi que la plupart des valeurs sont contenues entre 4000 et 5000 dollars.

Dans notre étude, il est important de savoir quelles sont les variables qui sont corrélées avec notre variable d'intérêt : la variable prix, afin de construire des modèles intéressants.
On finit donc par calculer les corrélations entre le prix et le poids, la profondeur en pourcentage, la table en pourcentage; la longueur, la largeur et la profondeur en mm du diamant.
On constate que les variables **carat**, **x**, **y** et **z** sont fortement corrélées avec la variable **price**.

Coefficients de corrélation de Pearson, N = 53940	price
carat	0.92159
depth	-0.01065
table	0.12713
x	0.88444
y	0.86542
z	0.86125

Corrélation entre la variable price et les autres variables quantitatives.

COMPARAISON D'ECHANTILLONS

LES PRIX DES DIAMANTS SONT-ILS IDENTIQUES EN FONCTION DES DIFFERENTES COULEURS ?

On souhaite savoir si les prix des diamants sont identiques en fonction des couleurs possibles.
On rappelle qu'il existe sept couleurs de diamants différentes alors de D (pour la meilleure) à J (pour la pire).

Pour commencer, on regarde la distribution de la variable qui représente la couleur.

color	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
D	6775	12.56	6775	12.56
E	9797	18.16	16572	30.72
F	9542	17.69	26114	48.41
G	11292	20.93	37406	69.35
H	8304	15.39	45710	84.74
I	5422	10.05	51132	94.79
J	2808	5.21	53940	100.00

Fréquence de la variable color.

Ensuite, nous choisissons de regarder si les distributions pour les prix d'un diamant de couleur D est en moyenne égale aux prix d'un diamant de couleur J.
Pour cela, nous créons un data frame contenant uniquement les couleurs D et J. On vérifie ensuite pour chacune des couleurs, si la variable prix suit une loi normale. On rappelle que ce n'est pas obligatoire. En effet, nous sommes dans le cas où nous possédons beaucoup d'observations, il n'est donc pas nécessaire de vérifier la normalité.

On peut maintenant effectuer un test de Student qui teste si les moyennes du prix entre nos deux échantillons sont égales ou non.

Méthode	Variances	DDL	Valeur du test t	Pr > |t|
Pooled	Egal	9581	-25.89	<.0001
Satterthwaite	Non égal	4197.9	-23.12	<.0001

Résultat du test de Student.

On obtient une p-valeur strictement inférieure à $\alpha = 0.05$. On rejette donc H_0 et on choisit H_1 : les moyennes des deux échantillons sont différentes.
Dans la réalité, cela veut dire que le prix est calculé en fonction de la couleur du diamant. Ce qui paraît tout à fait logique.

LES PRIX DES DIAMANTS SONT-ILS IDENTIQUES EN FONCTION DE LEUR QUALITE ?

On utilise la même procédure que précédemment pour savoir si le prix d'un diamant varie en fonction de sa qualité. On imagine déjà que cela va être le cas.
On commence par afficher la distribution de la variable qui représente la qualité.

cut	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
Fair	1610	2.98	1610	2.98
Good	4906	9.10	6516	12.08
Ideal	21551	39.95	28067	52.03
Premium	13791	25.57	41858	77.60
Very Goo	12082	22.40	53940	100.00

Fréquence de la variable cut.

Ici, on prend un échantillon avec des diamants convenables et un avec des diamants de très bonne qualité. On effectue alors un test de Student.

Méthode	Variances	DDL	Valeur du test t	Pr > |t|
Pooled	Egal	13690	3.65	0.0003
Satterthwaite	Non égal	2168	3.94	<.0001

Résultat du test de Student.

On obtient encore une fois une p-valeur strictement inférieure à $\alpha = 0.05$. On rejette donc H_0 et on accepte H_1 : les moyennes des deux échantillons sont différentes.
Dans la réalité, cela veut dire que le prix d'un diamant est lié à sa qualité.

ANALYSE EN COMPOSANTES PRINCIPALES

L'analyse en composantes principales permet de réduire la dimension de notre jeu de données afin de faire une analyse statistique descriptive meilleure.
L'ACP sert à décrire un jeu de données comportant de nombreux individus et se fait uniquement sur des variables quantitatives. L'analyse permet d'extraire l'information pertinente et la synthétise sous forme de composantes principales, nouveaux axes pour décrire le jeu de données. Elle permet aussi de quantifier les corrélations entre les variables du jeu de données. De plus, on peut observer des groupes de variables ayant des tendances identiques sont identifiés sur le cercle des corrélations.

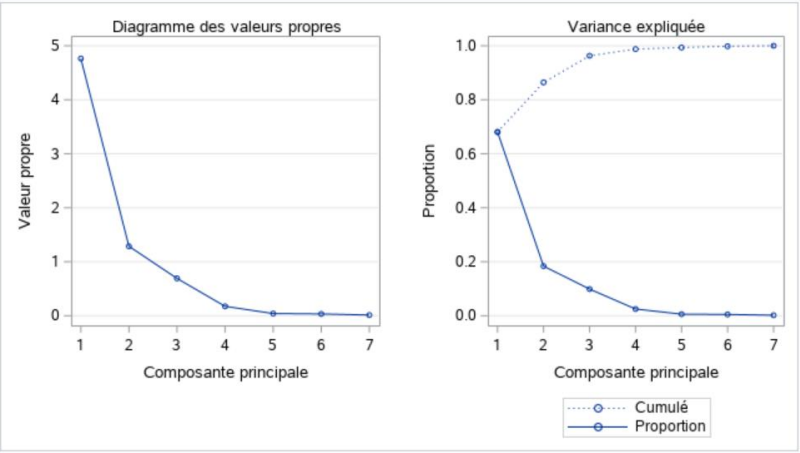
On trouve ci-dessous les résultats de notre ACP.
On notera que l'on n'affiche pas la matrice de corrélation car nous en avons déjà parlé précédemment.

Vecteurs propres							
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
price	0.425519	0.035258	-.105449	0.849778	0.053772	0.273309	0.082814
carat	0.452445	0.034696	-.005495	0.068359	-.133999	-.768151	-.425880
depth	-.000916	0.730680	0.672829	0.047248	0.088738	-.014450	0.055600
table	0.099516	-.675067	0.728069	0.059541	0.010376	0.025268	0.002049
x	0.453213	-.003513	-.039509	-.242995	-.088980	-.198461	0.828658
y	0.447265	-.002158	-.054189	-.328461	0.774058	0.215267	-.208857
z	0.445954	0.089035	0.039603	-.317007	-.603397	0.498670	-.279958

Matrice des vecteurs propres.

Valeurs propres de la matrice de corrélation				
	Valeur propre	Différence	Proportion	Cumulé
1	4.76391480	3.47804673	0.6806	0.6806
2	1.28586808	0.59505681	0.1837	0.8643
3	0.69081126	0.51705793	0.0987	0.9629
4	0.17375333	0.13344611	0.0248	0.9878
5	0.04030722	0.00736063	0.0058	0.9935
6	0.03294659	0.02054788	0.0047	0.9982
7	0.01239871		0.0018	1.0000

Matrice des valeurs propres.



Proportions des variances expliquées par les axes de l'ACP.

A l'aide de la figure concernant la matrice des valeurs propres, le critère de Keiser (que nous utilisons lors d'une ACP normée, ce qui est le cas de la notre) nous permet de retenir les axes dont les valeurs propres sont supérieures à 1.

La première valeur propre $vp_1 = 4.76 > 1$. On choisit donc de retenir le premier axe.
De plus, la seconde valeur propre $vp_2 = 1.29 > 1$, on choisit donc de retenir le deuxième axe.

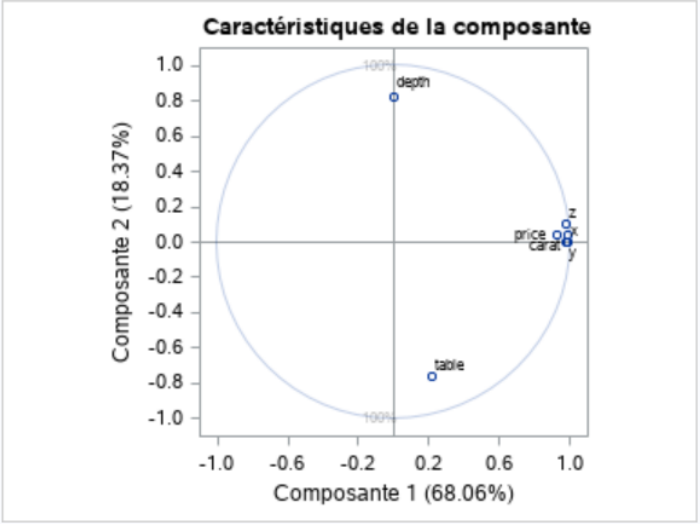
D'après le graphique des proportions, on observe un décrochement au deuxième axe, puis une décroissance régulière à partir du deuxième axe : seuls les deux premiers axes présentent un éventuel intérêt. On appelle ce critère le critère de "coude".

Pour conclure, on choisit de retenir les deux premiers axes qui conservent environ 86% de l'inertie totale. En effet le premier axe conserve à lui seul plus de la moitié (environ 68%) de l'inertie du nuage, et le second axe conserve une part importante de l'inertie totale qui est de 18%.
La chute est importante dès le troisième axe qui ne conserve moins de 1% de l'inertie totale.
On décide donc de ne retenir que les deux premiers axes.

Le sens de contribution d'une variable s'établit grâce au signe de la coordonnée.
A l'aide de la figure concernant la matrice des vecteurs propres, on remarque que toutes les variables ont une coordonnée positive mise à part la variable **depth** et **table** qui possède une coordonnée très petite. Ainsi les variables sauf **depth** et **table** contribuent dans le premier axe.

On peut donc afficher le cercle des corrélations dans la figure ci-contre.

D'après le cercle, on observe que mise à part les variables **depth** et **table** qui ne sont pas proche du cercle. Quant aux autres variables, elles sont du même côté de l'axe et très proche du cercle : elles contribuent toutes dans le même sens à la formation de l'axe. On appelle ceci par un effet de taille.



Cercle des corrélations pour les deux premiers axes.

CLASSIFICATION ASCENDANTE HIERARCHIQUE

BOOTSTRAP

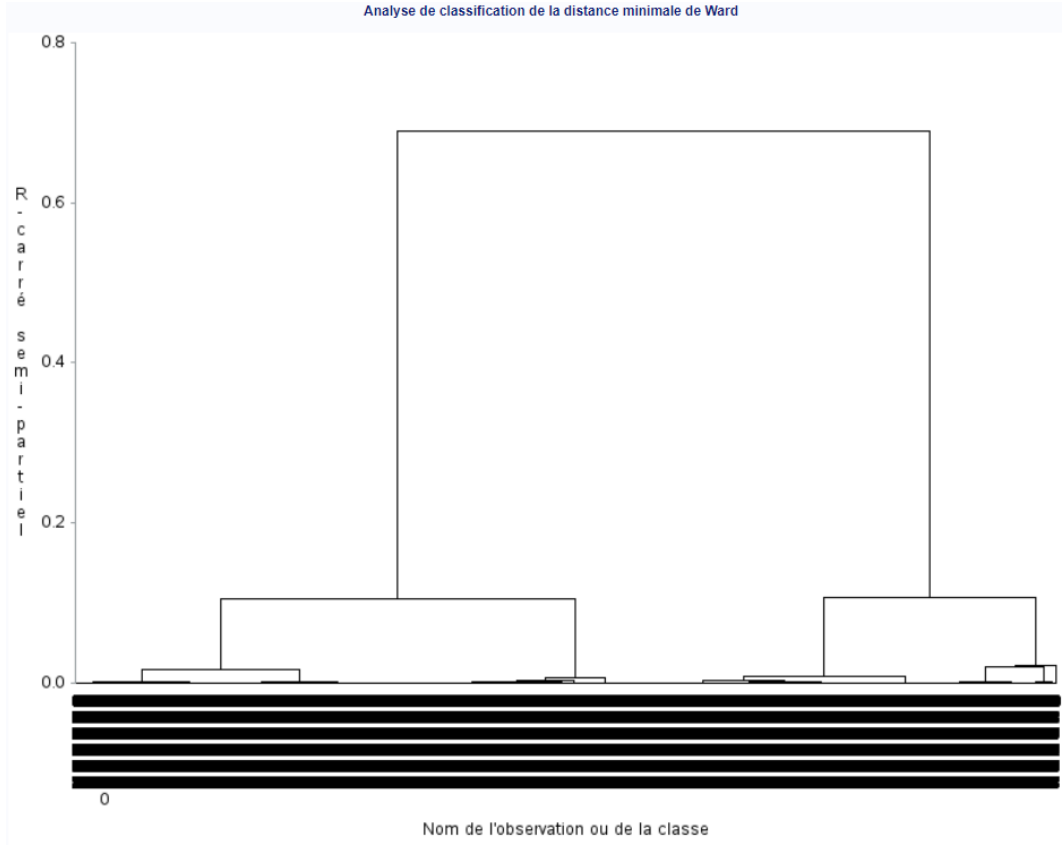
Avant de rentrer dans le sujet de la classification ascendante hiérarchique, nous allons utiliser le Bootstrap.
En effet, nous avons un jeu de données avec presque 54 000 données. La CAH sera donc obtenu après un long temps de calculs.
On préfère alors tirer au hasard un échantillons contenant 10 000 diamants.
Voici un aperçu du nouveau jeu de données obtenu :

Obs.	Replicate	carat	cut	color	clarity	depth	table	price	x	y	z
1	1	0.77	Premium	F	VS2	61.8	56	2889	5.94	5.90	3.66
2	1	0.79	Ideal	G	SI1	62.3	57	2898	5.90	5.85	3.66
3	1	0.31	Ideal	E	SI2	61.7	56	558	4.40	4.35	2.70
4	1	0.73	Ideal	E	VS2	62.2	59	3198	5.74	5.77	3.58
5	1	0.77	Ideal	F	VS1	62.5	57	3357	5.88	5.90	3.68
6	1	0.90	Very Goo	D	SI2	60.3	63	3473	6.22	6.12	3.72
7	1	0.31	Good	D	SI1	63.1	57	571	4.30	4.32	2.72
8	1	1.08	Ideal	H	SI2	62.0	57	3629	6.65	6.55	4.06
9	1	1.13	Ideal	E	I1	62.0	55	3797	6.70	6.66	4.14
10	1	1.00	Ideal	J	SI1	59.2	62	3929	6.47	6.50	3.84

Visualisation des 10 premières lignes de notre jeu de données par Bootstrap.

CAH

La Classification Ascendante Hiérarchique est une technique statistique visant à identifier des groupes d'observations ayant des caractéristiques similaires. On souhaite que les individus dans un même groupe soit le plus semblables possibles tandis que les classes sont le plus dissemblables.
Nous effectuons ainsi une CAH afin de regrouper nos 10 000 diamants suivant quatre variables : **carat**, **x**, **y** et **z**.
On affiche maintenant l'arbre hiérarchique :



Arbre hiérarchique

Comme dit précédemment, les individus qui ont des caractéristiques similaires sont réunis dans un même groupe (cluster, classe); les individus présentant des caractéristiques dissemblables (éloignées) sont associés à des groupes différents.
On remarque que le saut maximal est celui entre le passage de deux classes à une classe.
L'arbre hiérarchique nous suggère donc de classer les individus dans deux classes différentes.

MODELE LINEAIRE

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 **price** (Y_i) qui peut être modéliser par une loi Normale de paramètre μ et σ^2, comme vu précédemment c'est une variable continue symétriques par rapport à la moyenne, qui définit le prix d'un diamant.
Nous étudions donc le modèle suivant :

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_7 X_{7,i}$$

Avec :

- β_0 qui correspond à l'ordonnée à l'origine du modèle
- les β_k, $k \in [1,7]$ qui sont associés à la i_ème variable explicative
- les $X_{k,i}$ qui correspondent aux variables de notre étude : ce sont les variables explicatives de notre modèle. On précise que nous avons utilisés 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 les variables quantitatives suivantes **carat**, **x**, **y** et **z** (elles sont fortement corrélées avec la variable **price**).

D'après les résultats obtenus dans la figure ci-dessous (gauche), on a :

- dans la colonne Estimation, les estimations de nos 25 coefficients. En effet, pour chacune des modalités de nos variables quantitatives on obtient un coefficient.
- dans la colonne $Pr > |t|$, les pvaleurs du test de Student.
  Ce test permet de savoir si nos coefficients sont significatifs ou non, c'est à dire si la variable explicative associée joue un rôle dans la détermination du prix d'un diamant.

En prenant un seuil de 5%, on remarque que toutes les p-valeurs des coefficients, à l'exception de celle de **y**, sont strictement inférieures à 0.05 : dans ce cas on accepte H_1. De plus, on rappelle que pour les modalités 'Very Good' de **cut**, 'J' de **color** et 'VVS2' de **clarity** il n'y a pas de valeurs car ce sont les valeurs de références pour des variables qualitatives.

En conclusion, lorsque l'on a accepté H_1 les coefficients sont significatifs et donc les variables explicatives associées influent pas sur le prix du diamant.
On peut donc dire que les variables explicatives qui influent sur notre modèle sont définies dans la figure 15.
De plus, on constate que $R^2 = 0.9195$. Le modèle linéaire obtenu est donc très suffisant.

Paramètre	Estimation		Erreur type	Valeur du test t	Pr > |t|
Constante	-172.26430	B	82.07096388	-2.10	0.0358
cut Fair	-866.79484	B	30.86009420	-28.09	<.0001
cut Good	-195.74029	B	19.22920788	-10.18	<.0001
cut Ideal	160.87384	B	12.99702035	12.38	<.0001
cut Premium	41.06124	B	14.30099569	2.87	0.0041
cut Very Goo	0.00000	B	.	.	.
color D	2374.88286	B	26.17999357	90.71	<.0001
color E	2165.85841	B	24.96696545	86.75	<.0001
color F	2101.27844	B	24.86068647	84.52	<.0001
color G	1888.69140	B	24.35877005	77.54	<.0001
color H	1388.80330	B	24.93904067	55.69	<.0001
color I	902.77163	B	26.38909768	34.21	<.0001
color J	0.00000	B	.	.	.
clarity I1	-4978.31460	B	45.90482290	-108.45	<.0001
clarity IF	407.97154	B	31.28840665	13.04	<.0001
clarity SI1	-1299.91601	B	19.23761929	-67.57	<.0001
clarity SI2	-2258.64227	B	20.81377218	-108.52	<.0001
clarity VS1	-375.21190	B	20.37703516	-18.41	<.0001
clarity VS2	-691.85974	B	19.12721939	-36.17	<.0001
clarity VVS1	60.74155	B	24.64465085	2.46	0.0137
clarity VVS2	0.00000	B	.	.	.
carat	11129.29110		47.71306592	233.25	<.0001
x	-869.65395		30.61845294	-28.40	<.0001
y	32.20063		19.29587464	1.67	0.0952
z	-227.69544		29.72099239	-7.66	<.0001

Résultats du modèle linéaire sans ajustement.

Paramètres estimés				
Paramètre	DDL	Estimation	Erreur type	Valeur du test t
Intercept	1	-169.135086	82.050898	-2.06
cut Fair	1	-870.206270	30.792821	-28.26
cut Good	1	-196.359263	19.225948	-10.21
cut Ideal	1	160.373967	12.993783	12.34
cut Premium	1	39.530868	14.271798	2.77
cut Very Goo	0	0	.	.
color D	1	2374.985564	26.180355	90.72
color E	1	2166.033145	24.967159	86.76
color F	1	2101.420846	24.860951	84.53
color G	1	1888.739952	24.359156	77.54
color H	1	1388.896997	24.939390	55.69
color I	1	902.767118	26.389534	34.21
color J	0	0	.	.
clarity I1	1	-4979.563755	45.899479	-108.49
clarity IF	1	408.206152	31.288609	13.05
clarity SI1	1	-1300.146003	19.237444	-67.58
clarity SI2	1	-2258.807301	20.813882	-108.52
clarity VS1	1	-375.181166	20.377364	-18.41
clarity VS2	1	-692.061081	19.127155	-36.18
clarity VVS1	1	60.731456	24.645058	2.46
clarity VVS2	0	0	.	.
carat	1	11130	47.708821	233.30
x	1	-841.194934	25.429915	-33.08
z	1	-222.526843	29.559660	-7.53

Résultats du modèle linéaire après sélection des variables qui influent sur le prix d'un diamant.

CONCLUSION

Le but de ce projet était d'étudier le prix d'un diamant en fonction de différentes caractéristiques.

A l'aide de l'analyse descriptive du jeu de données, nous avons étudié et analysé la variable concernant le prix afin de bien les maîtriser. Nous avons constaté que le prix du diamant est lié d'une part au carat du diamant mais aussi à la longueur, la largeur et la profondeur de ce dernier. Celui-ci est en moyenne égale à 3933 $, mais il peut aller de 326$ jusqu'à 18823$. Par la suite, nous avons découvert que le prix dépend aussi de la couleur ainsi que de la qualité du diamant.

L'analyse en composante principale nous a permis de confirmer le fait que les variables concernant le prix, le carat, la longueur, la largeur et la profondeur d'un diamant sont très corrélés donc très liés.

Quant à la classification, elle nous indique que l'on peut distinguer deux groupes de diamants qui ont des caractéristiques similaires (dans chacun de leurs groupes) mais assez différents pour former deux classes différentes.

Grâce aux informations que nous avons obtenues, nous avons créé un modèle linéaire afin de prédire le prix d'un diamant. Nous avons constaté que le meilleur modèle linéaire est de prédire le prix d'un diamant en fonction des variables concernant la qualité, la couleur, la clarté, la longueur et la profondeur de celui-ci.

CODE ANNEXE

```
%let chemin=/home/u50424209/Projet_SAS/;
libname diamonds "&chemin.";

/*Création de macro variables */
%let var_quant = carat depth table price x y z;
%let var_qual = cut color clarity;

/*Chargement du jeu de données */
data diamonds;
 infile "&chemin./diamonds2.txt" firstobs=2 dlm = '09'x;
 input nb $ carat cut $ color $ clarity $ depth table
price x y z ;
run;

/*On enlève la colonne nb */
data diamonds2;
 set diamonds;
 drop nb;
run;

/*Affichage des 10 premières valeurs de notre jeu de
données */
proc print data = diamonds2 (obs = 10);
run;

/*Analyse des variables quantitatives */
proc corr data = diamonds2 nosimple noprob;
run;

/*Variable price corrélation avec les autres var quant
*/
proc corr data=WORK.DIAMONDS2 pearson nosimple noprob
plots(maxpoints=none)=matrix;
 var price;
 with carat depth table x y z;
run;

/*Distribution de la variable price */
ods graphics / reset width=4in height=3in imagemap;
proc univariate data=diamonds normal;
 var price ;
 histogram price / kernel;
 axis1 label=("Prix");
 axis2 label=("Pourcentage");
 title "Distribution du prix";
run;

/*Boxplot de la variable prix */
ods graphics / reset width=4in height=3in imagemap;
proc sgplot data=WORK.DIAMONDS2;
 title height=10pt "Boxplot de la variable prix";
 vbox price / boxwidth=0.3;
 yaxis grid label="Prix";
run;
```

```
/*Distribution des variables color et cut */
proc freq data=diamonds2;
 table color cut;
run;

/*Création du data frame pour les couleurs D et J
*/
data dia_colDJ;
 set diamonds2;
 where color = "D" or color = "J";
run;

/*Vérifications de la normalité de la variable
prix pour la couleur */
proc univariate data=dia_colDJ normal;
 var price;
 class color;   histogram; run;

/*Test de student sur le data frame dia_colDJ */
proc ttest data=dia_colDJ;
 class color;
 var price;
run;

/*Création du data frame pour les qualité Fair et
Very Good */
data dia_cutFVG;
 set diamonds2;
 where cut = "Fair" or cut = "Very Goo";
run;

/*Test de student sur le data frame dia_cutFVG */
proc ttest data=dia_cutFVG;
 class cut;
 var price;
run;

/*ACP On ne précise pas le nombre d'axes à
calculer : SAS calcul par défaut le nombre de
variables(7).
Nos données sont réduites par défaut.*/
ods graphics / reset width=4in height=3in
imagemap;
proc princomp data= diamonds2 out= diamonds_acp
outstat= mon_acp plots=pattern(circles=1.0);
 var &var_quant.;
run;
```

```
/*Bootstrap :
On récupère un échantillon de 10 000 données pour effctuer la CAH*/
proc surveyselect data=diamonds2 out=sample_dia noprint
 seed=175
 method=srs
 sampsize=100
 rep=100;
run;


proc print data = sample_dia (obs = 10);
run;


/*CAH */
ods graphics / reset width=4in height=3in imagemap;
proc cluster data = sample_dia method = ward standard;
 var carat x y z;
proc tree; run;

/*Modèle linéaire
Ici on commence par récupérer les estimations de nos coefficients*/
proc glm data=diamonds2;
 class &var_qual. ;
 model price= &var_qual. carat x y z/ solution;
run;

/*On utilise glmselect pour avoir les résultats avec les meilleurs coefficients */
proc glmselect data=diamonds2;
 class &var_qual. ;
 model price= &var_qual. carat x y z;
run;
```