Ecole Nationale Supérieure de Statistique et d'Economie Appliquée

Etude Econométrique

PREDICTION DES PRIX DES DIAMANTS

Ammi Yacine | Econométrie | September 25, 2020 | Professeur : Zatout Ali

4eme Année Statistique Appliquée | Groupe : 01

Plan du Travail :

I. Généralités sur les Diamants.

1. Qu'est-ce qu'un Diamant ?

- Le diamant est un minéral transparent composé de cristaux de carbone pur.
- Cette pierre précieuse est connue pour être le minéral le plus dur qui soit.

2. Comment se forme un Diamant ?

- Le diamant se forme naturellement à partir de carbone présent dans les profondeurs de la Terre.
- La plupart des diamants cristallisent à des profondeurs comprises entre 150 et 200 kilomètres.

3. Pourquoi le Diamant est-il si cher ?

- Le diamant est composé de carbone très résistant et très dur, c'est le minéral le plus résistant que nous connaissons à l'heure actuelle.
- De par sa résistance, c'est aussi son brillant qui plaît énormément au grand public.
- C'est une pierre précieuse rare et avec beaucoup de demandes.

4. Comment est calculé le prix d'un Diamant ?

- Le prix n'est pas uniquement fixé par sa rareté, il existe différentes caractéristiques qui donnent le prix d'un diamant.
- Voici les différentes caractéristiques d'un diamant :
  - La couleur du diamant
  - La clarté du diamant
  - La coupe du diamant
  - Le poids du diamant

II. Présentation des données.

1. Présentation générale des données :

L'ensemble de données contient les prix et autres attributs de près de 54 000 diamants. Cet ensemble de données contient des informations sur les prix des diamants, ainsi que divers attributs des diamants, dont certains sont connus pour influencer leur prix (en 2008 « Dollars Américain »): (carat, taille, couleur et clarté), ainsi que certaines mesures physiques (profondeur, table, prix, x, y et z). Les figures ci-dessous montre ce que représentent ces mesures.

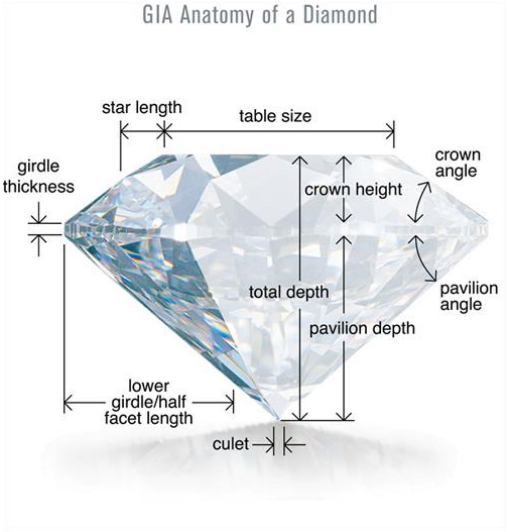
2. La source des données :

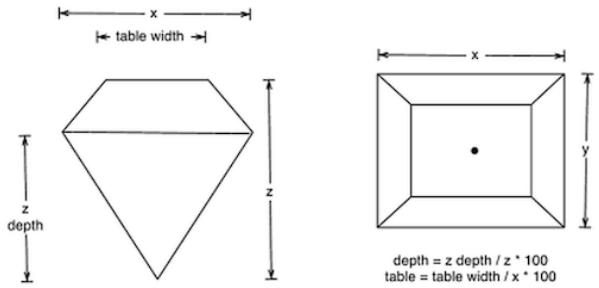
Pour la source, cet ensemble de données a été téléchargé du célèbre site des données « kaggle[1] ».

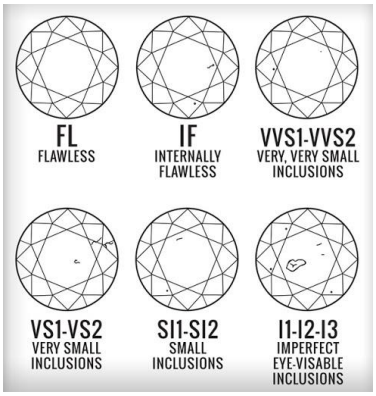
3. Présentation approfondie des données :

[1] https://www.kaggle.com/shivam2503/diamonds

- price : prix en dollars américains (\ $ 326 - \ $ 18,823).
- carat : Le carat est une unité de masse égale à 200 mg et est utilisé pour mesurer les pierres précieuses et les perles (0,2-5,01).
- cut : qualité de coupe de la coupe (Fair, Good, Very Good, Premium, Ideal).
- color : couleur du diamant, de J (pire) à D (meilleur).
- clarity : mesure de la clarté du diamant (I1 (pire), SI2, SI1, VS2, VS1, VVS2, VVS1, IF(meilleur))
- x : longueur en mm (0--10,74).
- y : largeur en mm (0-58,9)
- z : profondeur en mm (0--31,8)
- depth : pourcentage de profondeur totale = z / moyenne (x, y) = 2 * z / (x + y) (43--79).
- table : largeur haut du diamant par rapport au point le plus large (43--95).







III. Rappel théorique.

1. Le modèle linéaire général [2] :

Le modèle linéaire général est une généralisation du modèle de régression simple dans lequel figurent plusieurs variables explicatives :

$$y_t = a_0 + a_1 x_{1t} + a_2 x_{2t} + \ldots + a_k x_{kt} + \varepsilon_t \; pour \; t = 1, \ldots, n$$

Avec :

y_t = variable à expliquer à la date t ;

x_{1t} = variable explicative 1 à la date t ;

x_{2t} = variable explicative 2 à la date t ;

. . .

x_{kt} = variable explicative k à la date t ;

$a_0, a_1, \ldots, a_k$ = paramètres du modèle ;

ε_t = erreur de spécification

n = nombre d'observations.

2. Estimation des coefficients de régression [3] :

Afin d'estimer le vecteur a composé des coefficients $a_0, a_1 \ldots a_k$, nous appliquons la méthode des Moindres Carrés Ordinaires (MCO) qui consiste à minimiser la somme des carrés des erreurs.

Cette méthode consiste à écrire l'ensemble des données sous la forme matricielle, ensuite pour trouver les coefficients des variables indépendantes, il faut appliquer la formule :

$$\hat{a} = (X' X)^{-1} X' Y$$

Attention : En pratique, tous les calculs sont faits à l'aide d'une machine et des logiciels statistiques adéquats.

3. Hypothèses [4] :

- H1 : le modèle est linéaire en x_t (ou en n'importe quelle transformation de x_t).
- H2 : les valeurs x_t sont observées sans erreur (x_t non aléatoire).
- H3 : $E(\varepsilon_t) = 0$, l'espérance mathématique de l'erreur est nulle : en moyenne le modèle est bien spécifié et donc l'erreur moyenne est nulle.
- H4 : $E(\varepsilon^2_t) = \sigma^2_\varepsilon$, la variance de l'erreur est constante : le risque de l'amplitude de l'erreur est le même quelle que soit la période.
- H5 : $E(\varepsilon_t \varepsilon_{t'}) = 0$ si $t \neq t'$, les erreurs sont non corrélées (ou encore indépendantes) : une erreur à l'instant t n'a pas d'influence sur les erreurs suivantes.
- H6 : $Cov(x_t, \varepsilon_t) = 0$, l'erreur est indépendante de la variable explicative.
- H7 : absence de colinéarité entre les variables explicatives, cela implique que la matrice (X' X) est régulière et que la matrice inverse $(X' X)^{-1}$ existe.

---

[2] Régis Bourbonnais, Économétrie : Cours et exercices corrigés, 9e édition, Dunod, 2015, pp-47-48

[3] ibid, pp-49

[4] ibid, pp-18

4. Comment s'occuper des variables indépendantes qualitatives [5] :

En règle générale, dans la plupart des recherches économiques, un modèle de régression contient des variables explicatives qui sont quantitatives et d'autres qui sont qualitatives. Les modèles contenant un mélange de variables quantitatives et qualitatives sont appelés analyse des modèles de covariance (ANCOVA). Les modèles ANCOVA sont une extension du Modèle ANOVA en ce qu'ils fournissent une méthode de contrôle statistique des effets régresseurs quantitatifs, appelés covariables ou variables de contrôle, dans un modèle qui comprend des régresseurs à la fois quantitatifs et qualitatifs, ou fictifs.

Les variables fictives peuvent être intégrées aux modèles de régression aussi facilement que les variables quantitatives.

Une façon de « quantifier » ces attributs sont en construisant des variables artificielles qui prennent des valeurs de 1 ou 0, 1 indiquant la présence (ou possession) de cet attribut et 0 indiquant l'absence de cet attribut.

Une variable indépendante fictive (également appelée variable explicative fictive) qui, pour certaines observations, a une valeur de 0, fera que le coefficient de cette variable n'aura aucun rôle à influencer la variable dépendante, tandis que lorsque la variable fictive prend la valeur 1, son coefficient agit pour modifier l'interception.

IV. Partie pratique

1. Plan de l'étude :

Etant donné que les variables indépendantes de l'ensemble de données est composé de variables quantitatives et qualitatives à la fois, le choix du modèle se déroulera en deux phases :

- Première Phase : Modèle sans variables qualitatives
- Deuxième Phase : Modèle avec les variables qualitatives (facultatif / supplémentaire)

L'étude économétrique sera réalisée intégralement en utilisant le langage de programmation « Python » et a l'aide des bibliothèques dédiés aux modèles statistiques.

Remarque : Le code utilisé dans l'exécution des différentes étapes de l'étude économétrique sera fourni séparément dans un fichier PDF optionnel.

2. Préparation des données :

2.1. Importation des Données :

[5] Damodar N. Gujarati et Dawn C. Porter, Basic Econometrics, Fifth Edition, 2008, pp-234

Out[2]:

	Unnamed: 0	carat	cut	color	clarity	depth	table	price	x	y	z
0	1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...	...
53935	53936	0.72	Ideal	D	SI1	60.8	57.0	2757	5.75	5.76	3.50
53936	53937	0.72	Good	D	SI1	63.1	55.0	2757	5.69	5.75	3.61
53937	53938	0.70	Very Good	D	SI1	62.8	60.0	2757	5.66	5.68	3.56
53938	53939	0.86	Premium	H	SI2	61.0	58.0	2757	6.15	6.12	3.74
53939	53940	0.75	Ideal	D	SI2	62.2	55.0	2757	5.83	5.87	3.64

53940 rows × 11 columns

2.2. Inspection des données :

Out[4]:

	carat	depth	table	price	x	y	z
count	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000
mean	0.797940	61.749405	57.457184	3932.799722	5.731157	5.734526	3.538734
std	0.474011	1.432621	2.234491	3989.439738	1.121761	1.142135	0.705699
min	0.200000	43.000000	43.000000	326.000000	0.000000	0.000000	0.000000
25%	0.400000	61.000000	56.000000	950.000000	4.710000	4.720000	2.910000
50%	0.700000	61.800000	57.000000	2401.000000	5.700000	5.710000	3.530000
75%	1.040000	62.500000	59.000000	5324.250000	6.540000	6.540000	4.040000
max	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

On remarque que les valeurs minimales de "x", "y" et "z" sont 0.oooooooo => ces valeurs sont erronées car ce n'est pas possible d'avoir ces valeurs, Il faut donc les éliminer.

2.3. Suppression des valeurs erronées :

Out[5]:

	carat	depth	table	price	x	y	z
count	53920.000000	53920.000000	53920.000000	53920.000000	53920.000000	53920.000000	53920.000000
mean	0.797698	61.749514	57.456834	3930.993231	5.731627	5.734887	3.540046
std	0.473795	1.432331	2.234064	3987.280446	1.119423	1.140126	0.702530
min	0.200000	43.000000	43.000000	326.000000	3.730000	3.680000	1.070000
25%	0.400000	61.000000	56.000000	949.000000	4.710000	4.720000	2.910000
50%	0.700000	61.800000	57.000000	2401.000000	5.700000	5.710000	3.530000
75%	1.040000	62.500000	59.000000	5323.250000	6.540000	6.540000	4.040000
max	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

3. Première Phase : Modèle sans variables qualitatives.

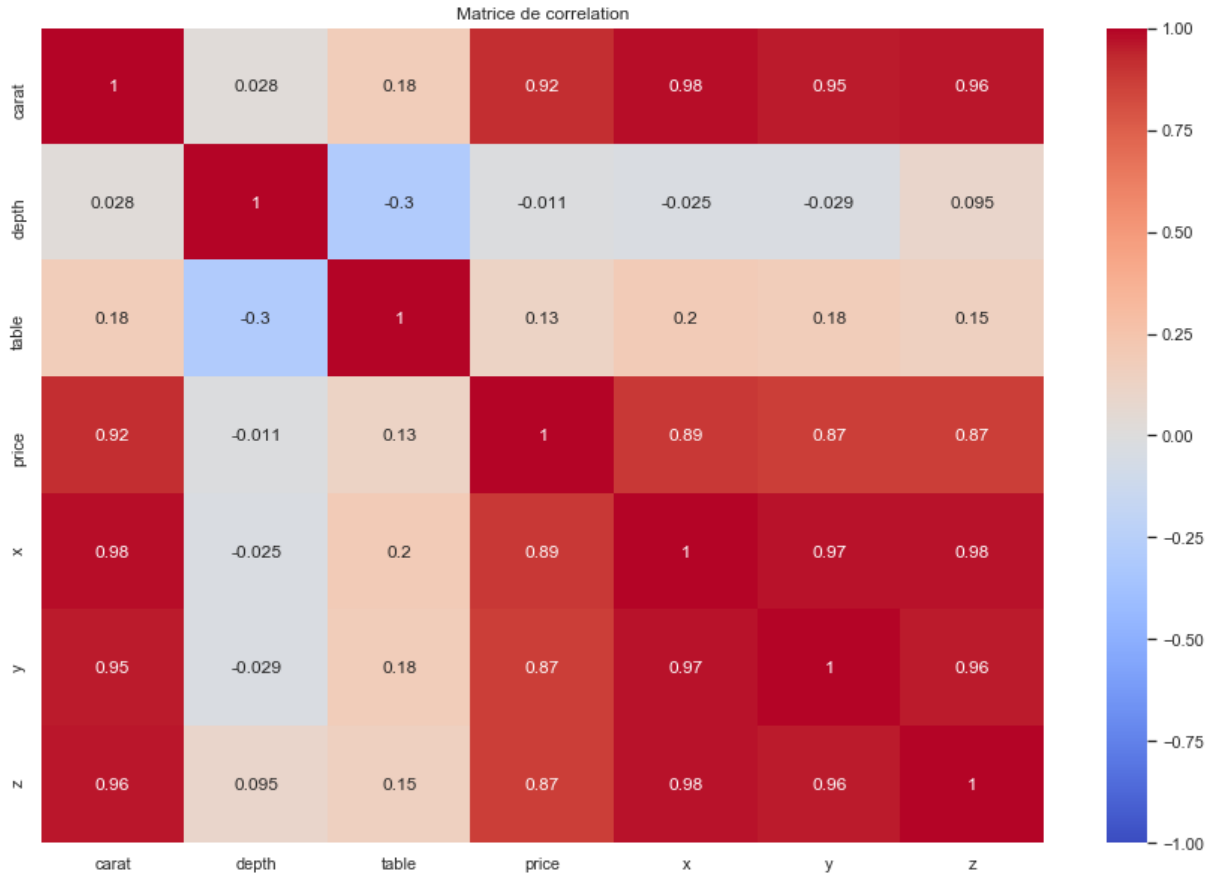
3.1. Enlèvement des variables qualitatives :

Out[6]:

	carat	depth	table	price	x	y	z
0	0.23	61.5	55.0	326	3.95	3.98	2.43
1	0.21	59.8	61.0	326	3.89	3.84	2.31
2	0.23	56.9	65.0	327	4.05	4.07	2.31
3	0.29	62.4	58.0	334	4.20	4.23	2.63
4	0.31	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...
53935	0.72	60.8	57.0	2757	5.75	5.76	3.50
53936	0.72	63.1	55.0	2757	5.69	5.75	3.61
53937	0.70	62.8	60.0	2757	5.66	5.68	3.56
53938	0.86	61.0	58.0	2757	6.15	6.12	3.74
53939	0.75	62.2	55.0	2757	5.83	5.87	3.64

53920 rows × 7 columns

3.2. Inspection des corrélations entre toutes les variables pour vérifier la "multicolinéarité" a l'aide d'une matrice des corrélations



On remarque que les variables "x", "y" et "z" sont toutes positivement et fortement corrélés entre elles même et avec la variable "carat", Donc il serait peut-être mieux d'enlever ces trois variables car elles ne rapportent pas plus d'explication dans notre modèle et pour essentiellement éviter les problèmes de multicolinéarité.

3.3. Estimation du Modèle avec les variables "x", "y" et "z" :

Out[8]:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.860
Model:	OLS	**Adj. R-squared:**	0.860
Method:	Least Squares	**F-statistic:**	5.505e+04
Date:	Tue, 29 Sep 2020	**Prob (F-statistic):**	0.00
Time:	23:03:23	**Log-Likelihood:**	-4.7061e+05
No. Observations:	53920	**AIC:**	9.412e+05
Df Residuals:	53913	**BIC:**	9.413e+05
Df Model:	6		
Covariance Type:	nonrobust		

R_carré_ajusté

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.132e+04	456.639	46.682	0.000	2.04e+04	2.22e+04
carat	1.099e+04	66.978	164.062	0.000	1.09e+04	1.11e+04
depth	-203.0201	5.658	-35.882	0.000	-214.110	-191.930
table	-101.9364	3.079	-33.108	0.000	-107.971	-95.902
x	-1412.9648	45.781	-30.863	0.000	-1502.697	-1323.233
y	88.1662	25.683	3.433	0.001	37.828	138.504
z	-46.5781	50.091	-0.930	0.352	-144.757	51.601

Statistique de student

Omnibus:	14157.382	**Durbin-Watson:**	1.270
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	409714.584
Skew:	0.652	**Prob(JB):**	0.00
Kurtosis:	16.441	**Cond. No.**	6.04e+03

Le model est puissant avec un R_Carré = 0.86, et on remarque que la variable "z" avec une P_value = 0.352 > 0.05 ce qui implique que la variable "z" n'est pas significative.

3.4. Estimation du Modèle sans les variables "x", "y" et "z" :

Out[9]:

OLS Regression Results

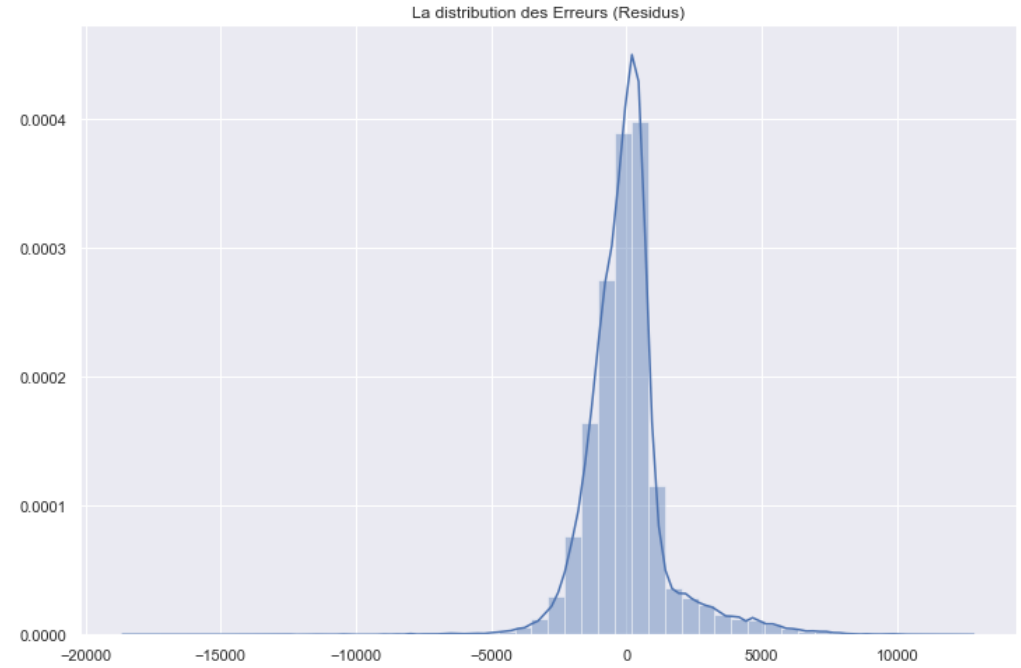
Dep. Variable:	price	**R-squared:**	0.854
Model:	OLS	**Adj. R-squared:**	0.854
Method:	Least Squares	**F-statistic:**	1.049e+05
Date:	Tue, 29 Sep 2020	**Prob (F-statistic):**	0.00
Time:	23:03:24	**Log-Likelihood:**	-4.7174e+05
No. Observations:	53920	**AIC:**	9.435e+05
Df Residuals:	53916	**BIC:**	9.435e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.301e+04	390.842	33.284	0.000	1.22e+04	1.38e+04
carat	7858.0632	14.152	555.253	0.000	7830.325	7885.802
depth	-151.4453	4.819	-31.426	0.000	-160.891	-142.000
table	-104.3284	3.141	-33.219	0.000	-110.484	-98.173

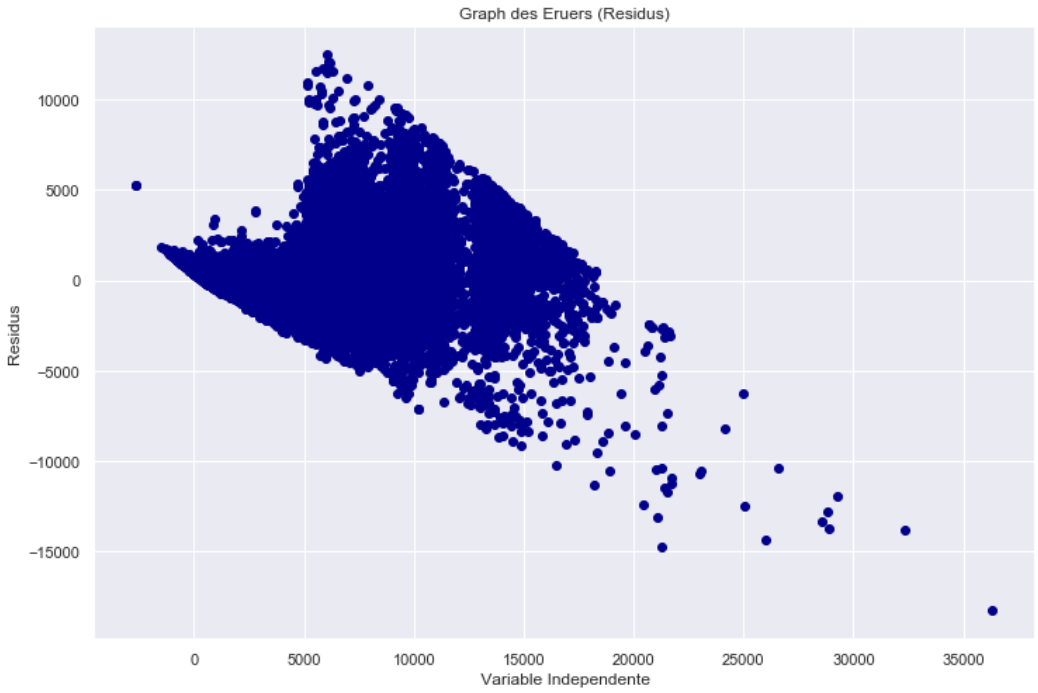
Omnibus:	14098.572	**Durbin-Watson:**	1.015
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	154432.961
Skew:	0.945	**Prob(JB):**	0.00
Kurtosis:	11.073	**Cond. No.**	5.02e+03

Malgré la suppression de trois variables "x", "y" et "z", La puissance explicative de notre modèle na pas diminuée significativement avec le R_carré précédant = 0.86 contre un R_carré actuellement = 0.854, d'où on garde ce dernier modèle.

3.5. Examination des Erreurs (Residus) :



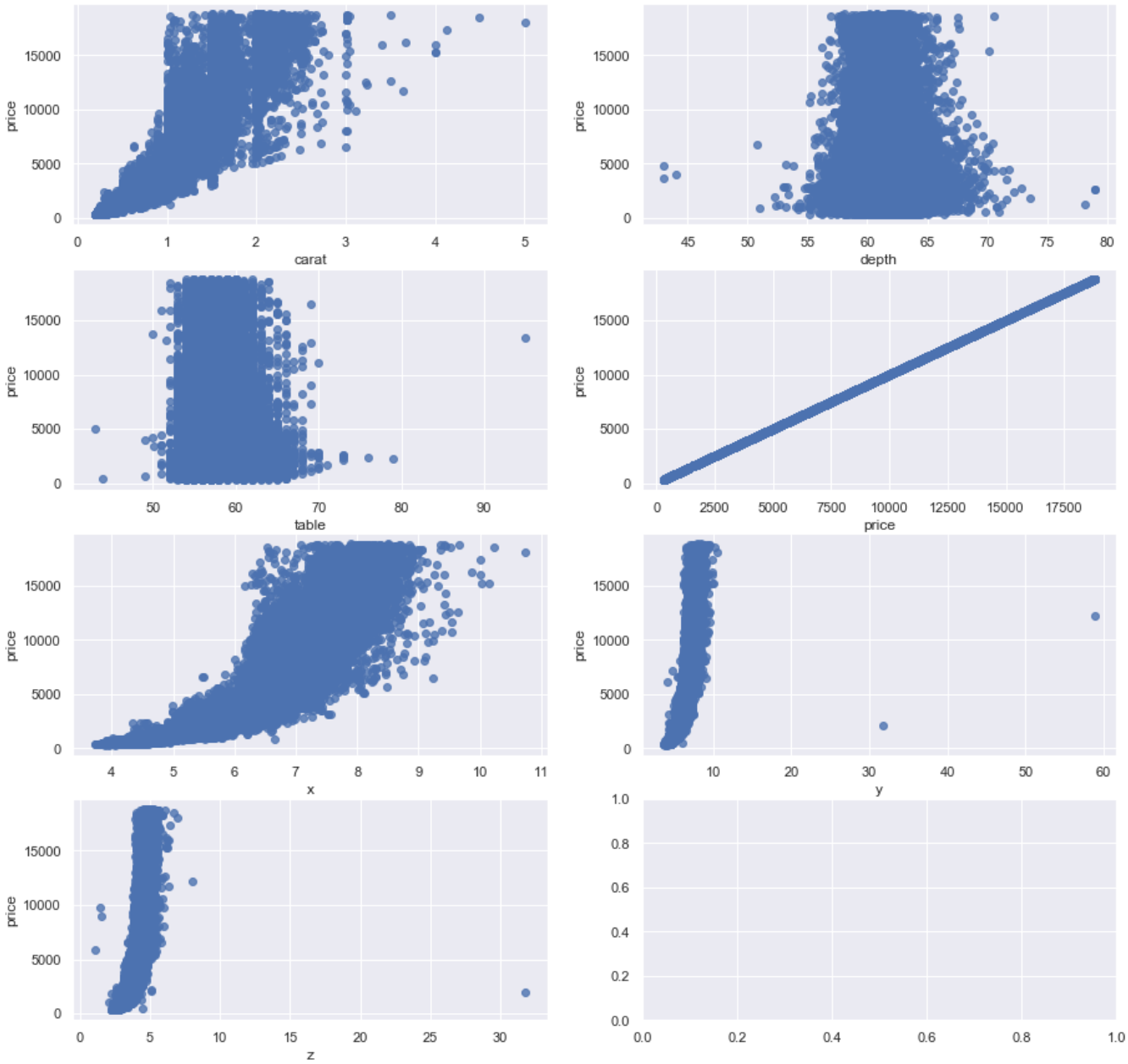
Selon le graph, la distribution des Erreurs qui est supposé suivre une loi Normale n'est pas parfaitement Normale



On remarque que les erreurs ne sont pas distribuées d'une façon aléatoire, en plus, le graph montre qu'il existe peut-être une heteroscedasticité dans les erreurs, la cause la plus probable est qu'une ou plusieurs variables sont heteroscedastiques par rapport a la variable dépendante "price" ou même que leurs relations ne sont pas linéaires.

On peut en déduire que ce modèle n'est pas bon.

3.6. Vérification de la linéarité et de l'homoscedasticité de chaque variable avec la variable dépendante "Price" :



On remarque que les variables "carat" et "x" sont heteroscedastiques par rapport a la variable dépendante "price" et que la relation entre "price" et "carat" et entre "price" et "x" ne sont pas linéaires

Pour enlever l'heteroscedasticité et assurer la linéarité, on applique la transformation logarithmique pour ces variables.

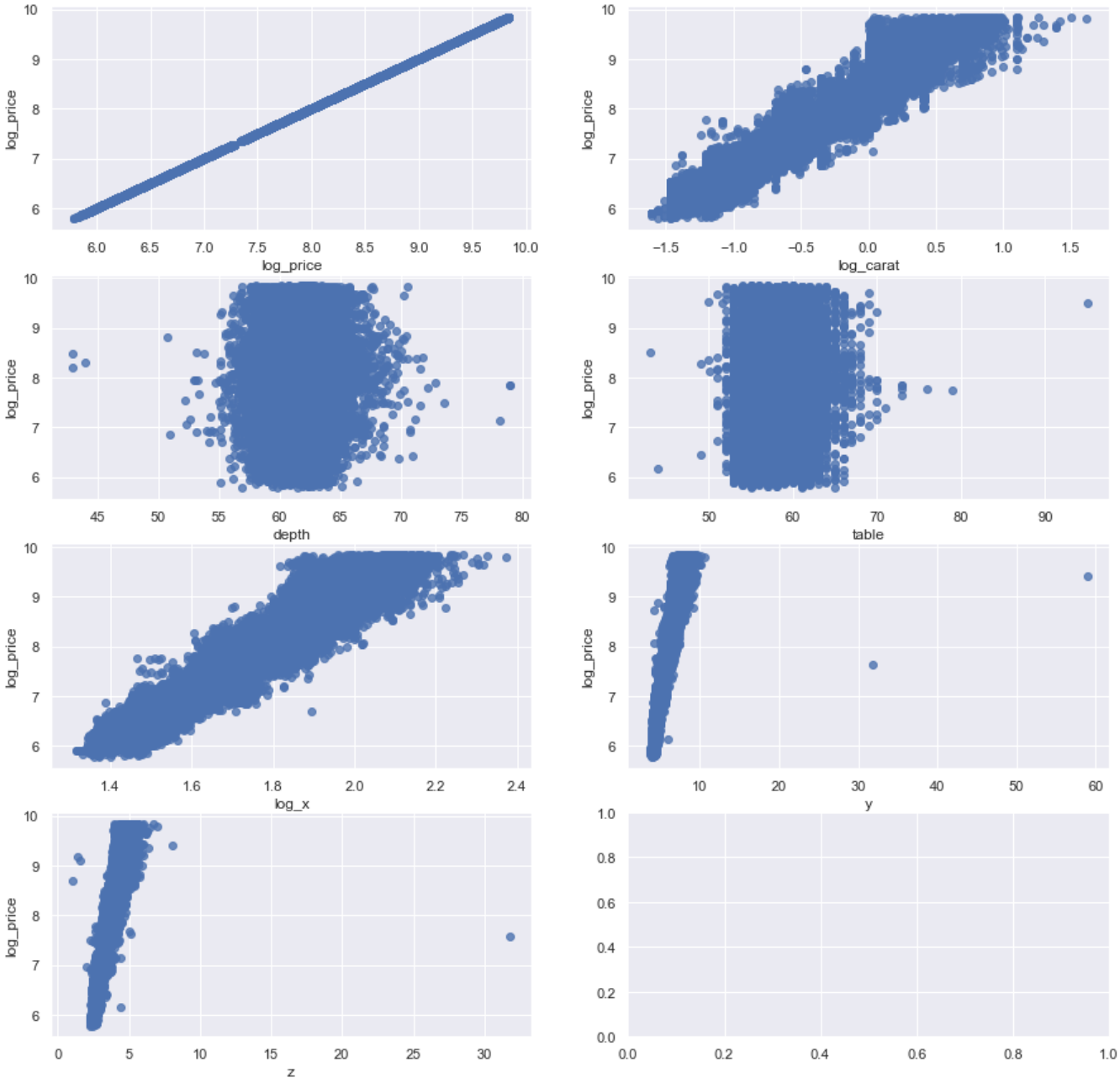
3.7. Application de la transformation logarithmique :

Out[13]:

	log_price	log_carat	depth	table	log_x	y	z
0	5.786897	-1.469676	61.5	55.0	1.373716	3.98	2.43
1	5.786897	-1.560648	59.8	61.0	1.358409	3.84	2.31
2	5.789960	-1.469676	56.9	65.0	1.398717	4.07	2.31
3	5.811141	-1.237874	62.4	58.0	1.435085	4.23	2.63
4	5.814131	-1.171183	63.3	58.0	1.467874	4.35	2.75
...	...	...	...	...	...	...	...
53935	7.921898	-0.328504	60.8	57.0	1.749200	5.76	3.50
53936	7.921898	-0.328504	63.1	55.0	1.738710	5.75	3.61
53937	7.921898	-0.356675	62.8	60.0	1.733424	5.68	3.56
53938	7.921898	-0.150823	61.0	58.0	1.816452	6.12	3.74
53939	7.921898	-0.287682	62.2	55.0	1.763017	5.87	3.64

53920 rows × 7 columns

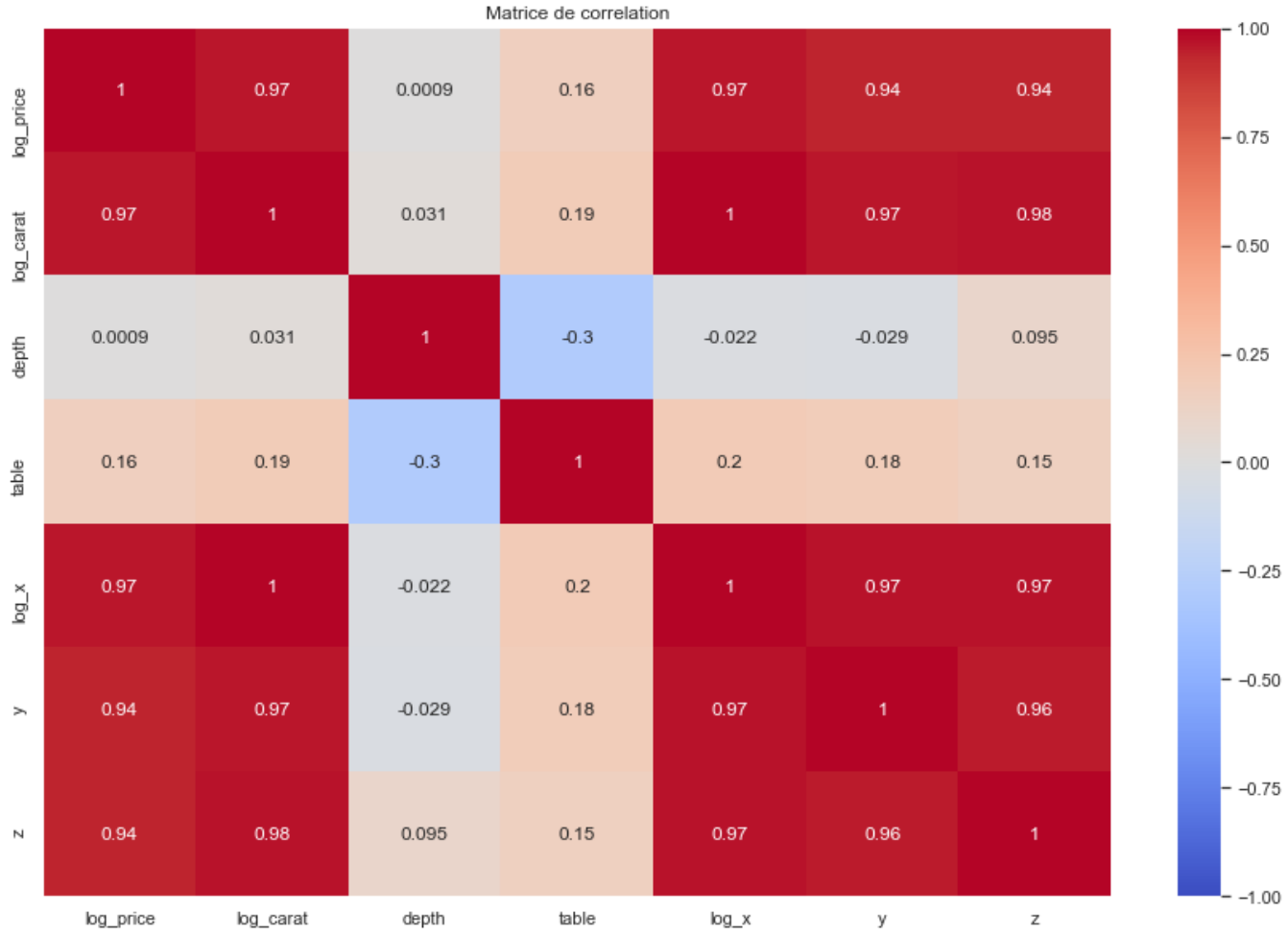
3.8. Vérification de la linéarité de chaque variable avec la variable dépendante "log_price" :



La linéarité est vérifiée

L'homoscedasticité est vérifiée

3.9. Inspection des corrélations entre toutes les variables pour vérifier la "multicolinéarité" a l'aide d'une matrice des corrélations



On remarque que les variables "log_x", "y" et "z" sont toutes positivement et fortement corrélés entre elles même et avec la variable "carat", Donc il serait peut-être mieux d'enlever ces trois variables car elles ne rapportent pas plus d'explication dans notre modèle et pour essentiellement éviter les problèmes de multicolinéarité.

3.10. Estimation du Modèle sans les variables "log_x", "y" et "z" :

Out[17]:

OLS Regression Results

Dep. Variable:	log_price	**R-squared:**	0.935
Model:	OLS	**Adj. R-squared:**	0.935
Method:	Least Squares	**F-statistic:**	2.597e+05
Date:	Tue, 29 Sep 2020	**Prob (F-statistic):**	0.00
Time:	23:03:38	**Log-Likelihood:**	-3481.7
No. Observations:	53920	**AIC:**	6971.
Df Residuals:	53916	**BIC:**	7007.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.3096	0.067	169.672	0.000	11.179	11.440
log_carat	1.6916	0.002	869.559	0.000	1.688	1.695
depth	-0.0290	0.001	-35.526	0.000	-0.031	-0.027
table	-0.0185	0.001	-34.780	0.000	-0.020	-0.017

Omnibus:	814.942	**Durbin-Watson:**	1.248
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1478.824
Skew:	0.087	**Prob(JB):**	0.00
Kurtosis:	3.792	**Cond. No.**	5.06e+03

R_carré ajusté

Statistique de Fisher (Significativité Globale)

Statistique de student (Significativité de chaque variable)

Les coefficients de chaque variable du modèle

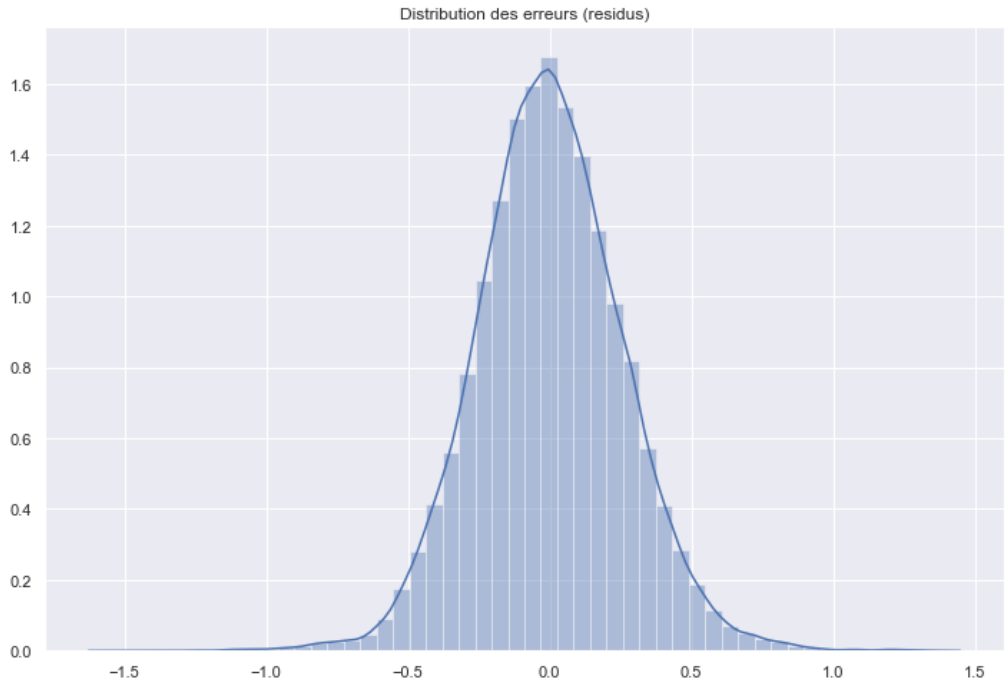
Test de Durbin-Watson

- Le modèle est puissant avec un R_Carré ajusté = 0.935
- Toutes les variables sont significatives avec une p_value < 0.05
- Le modèle est globalement significatif étant donné que la p_value de la f_statistic (statistique de Fisher de la significativité globale) = 0.00 < 0.05
- Le test de Durbin-Watson a une valeur de 1.248 => il existe peut-être une autocorrélation positive des erreurs

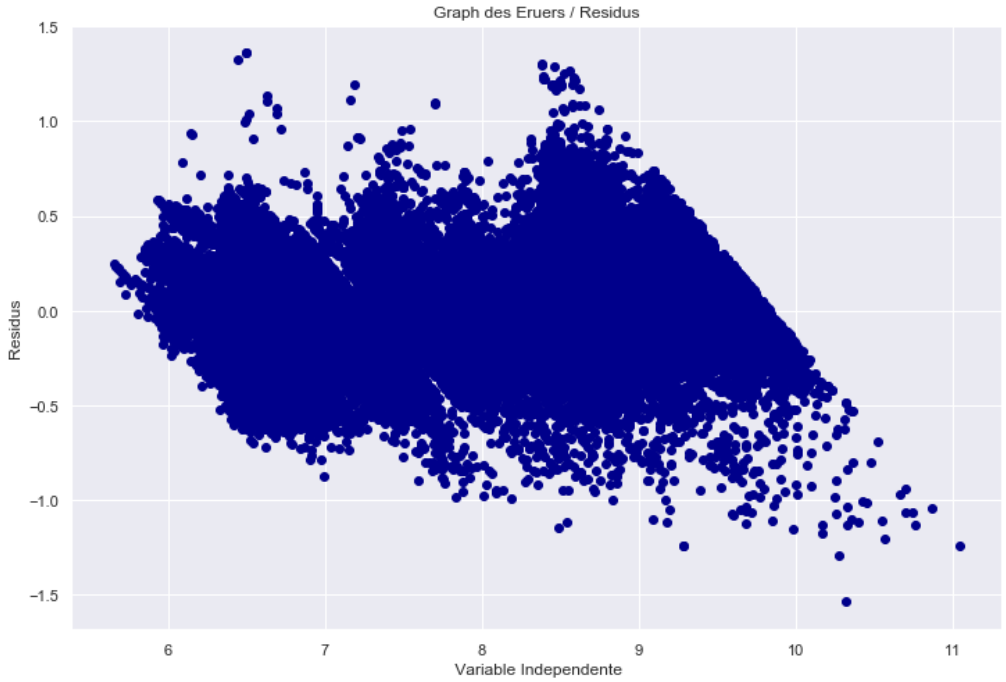
Le modèle :

log_price = 11.3 + 1.69 * log_carat - 0.029 * depth - 0.0185 * table

3.11. Vérification de la distribution des Erreurs (Résidus) :



La distribution des erreurs suit une loi Normale avec une moyenne = 0



L'homoscedasticité est vérifiée.

4. Deuxième Phase : Modèle avec les variables qualitatives.

4.1. Transformation des variables qualitatives en variables Fictives / Muettes :

Out[20]:

	log_price	log_carat	depth	table	clarity_IF	clarity_SI1	clarity_SI2	clarity_VS1	clarity_VS2	clarity_VVS1	...	cut_Good	cut_Ideal	cut_Premium	cut_V G
0	5.786897	-1.469676	61.5	55.0	0	0	1	0	0	0	...	0	1	0	
1	5.786897	-1.560648	59.8	61.0	0	1	0	0	0	0	...	0	0	1	
2	5.789960	-1.469676	56.9	65.0	0	0	0	1	0	0	...	1	0	0	
3	5.811141	-1.237874	62.4	58.0	0	0	0	0	1	0	...	0	0	1	
4	5.814131	-1.171183	63.3	58.0	0	0	1	0	0	0	...	1	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
53935	7.921898	-0.328504	60.8	57.0	0	1	0	0	0	0	...	0	1	0	
53936	7.921898	-0.328504	63.1	55.0	0	1	0	0	0	0	...	1	0	0	
53937	7.921898	-0.356675	62.8	60.0	0	1	0	0	0	0	...	0	0	0	
53938	7.921898	-0.150823	61.0	58.0	0	0	1	0	0	0	...	0	0	1	
53939	7.921898	-0.287682	62.2	55.0	0	0	1	0	0	0	...	0	1	0	

53920 rows × 21 columns

4.2. Estimation du modèle :

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.9362	0.043	183.219	0.000	7.851	8.021
clarity_IF	1.1130	0.006	184.079	0.000	1.101	1.125
clarity_SI1	0.5926	0.005	114.838	0.000	0.582	0.603
clarity_SI2	0.4275	0.005	82.365	0.000	0.417	0.438
clarity_VS1	0.8118	0.005	154.003	0.000	0.801	0.822
clarity_VS2	0.7417	0.005	142.906	0.000	0.732	0.752
clarity_VVS1	1.0181	0.006	182.159	0.000	1.007	1.029
clarity_VVS2	0.9468	0.005	174.288	0.000	0.936	0.957
color_E	-0.0542	0.002	-25.600	0.000	-0.058	-0.050
color_F	-0.0946	0.002	-44.141	0.000	-0.099	-0.090
color_G	-0.1601	0.002	-76.348	0.000	-0.164	-0.156
color_H	-0.2509	0.002	-112.728	0.000	-0.255	-0.247
color_I	-0.3724	0.002	-149.362	0.000	-0.377	-0.367
color_J	-0.5108	0.003	-166.125	0.000	-0.517	-0.505
cut_Good	0.0786	0.004	19.757	0.000	0.071	0.086
cut_Ideal	0.1582	0.004	40.003	0.000	0.150	0.166
cut_Premium	0.1368	0.004	35.859	0.000	0.129	0.144
cut_Very Good	0.1149	0.004	30.137	0.000	0.107	0.122
depth	-0.0009	0.000	-1.931	0.054	-0.002	1.39e-05
log_carat	1.8838	0.001	1659.875	0.000	1.882	1.886
table	-0.0004	0.000	-1.017	0.309	-0.001	0.000

Toutes les variables du modèle sont significatives sauf "depth" avec une p_value = 0.054 > 0.05 et la variable "table" avec une p_value = 0.309 > 0.05.

4.3. Estimation du modèle sans les variables "depth" et "table" :

Dep. Variable:	log_price	R-squared:	0.983
Model:	OLS	Adj. R-squared:	0.983
Method:	Least Squares	F-statistic:	1.693e+05
Date:	Tue, 29 Sep 2020	Prob (F-statistic):	0.00
Time:	23:03:42	Log-Likelihood:	31961.
No. Observations:	53920	AIC:	-6.388e+04
Df Residuals:	53901	BIC:	-6.372e+04
Df Model:	18		
Covariance Type:	nonrobust		

R_carré ajusté

Statistique de Fisher(Significativité Globale)

Statistique de student (Significativité de chaque variable)

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.8568	0.006	1382.891	0.000	7.846	7.868
clarity_IF	1.1137	0.006	184.434	0.000	1.102	1.125
clarity_SI1	0.5929	0.005	114.942	0.000	0.583	0.603
clarity_SI2	0.4279	0.005	82.487	0.000	0.418	0.438
clarity_VS1	0.8122	0.005	154.231	0.000	0.802	0.823
clarity_VS2	0.7421	0.005	143.067	0.000	0.732	0.752
clarity_VVS1	1.0186	0.006	182.410	0.000	1.008	1.030
clarity_VVS2	0.9472	0.005	174.523	0.000	0.937	0.958
color_E	-0.0542	0.002	-25.600	0.000	-0.058	-0.050
color_F	-0.0945	0.002	-44.135	0.000	-0.099	-0.090
color_G	-0.1602	0.002	-76.405	0.000	-0.164	-0.156
color_H	-0.2511	0.002	-112.834	0.000	-0.255	-0.247
color_I	-0.3725	0.002	-149.483	0.000	-0.377	-0.368
color_J	-0.5110	0.003	-166.237	0.000	-0.517	-0.505
cut_Good	0.0801	0.004	20.592	0.000	0.073	0.088
cut_Ideal	0.1613	0.004	45.442	0.000	0.154	0.168
cut_Premium	0.1394	0.004	38.942	0.000	0.132	0.146
cut_Very Good	0.1172	0.004	32.392	0.000	0.110	0.124
log_carat	1.8837	0.001	1668.528	0.000	1.882	1.886

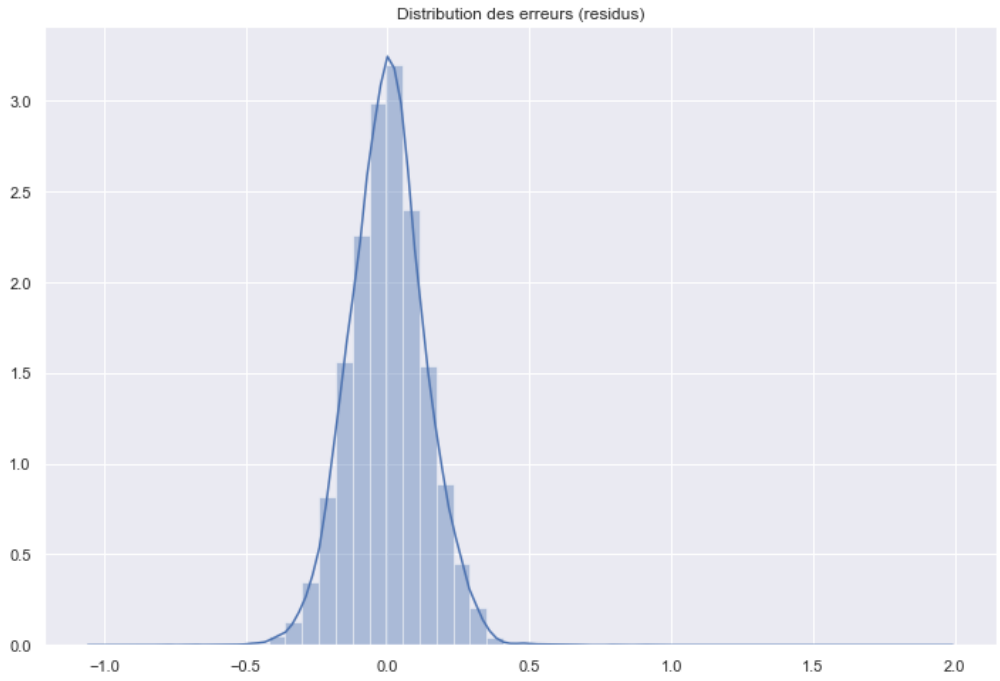
Les coefficients de chaque variable du modèle

Omnibus:	3725.321	Durbin-Watson:	1.245
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15922.894
Skew:	0.210	Prob(JB):	0.00
Kurtosis:	5.629	Cond. No.	33.4

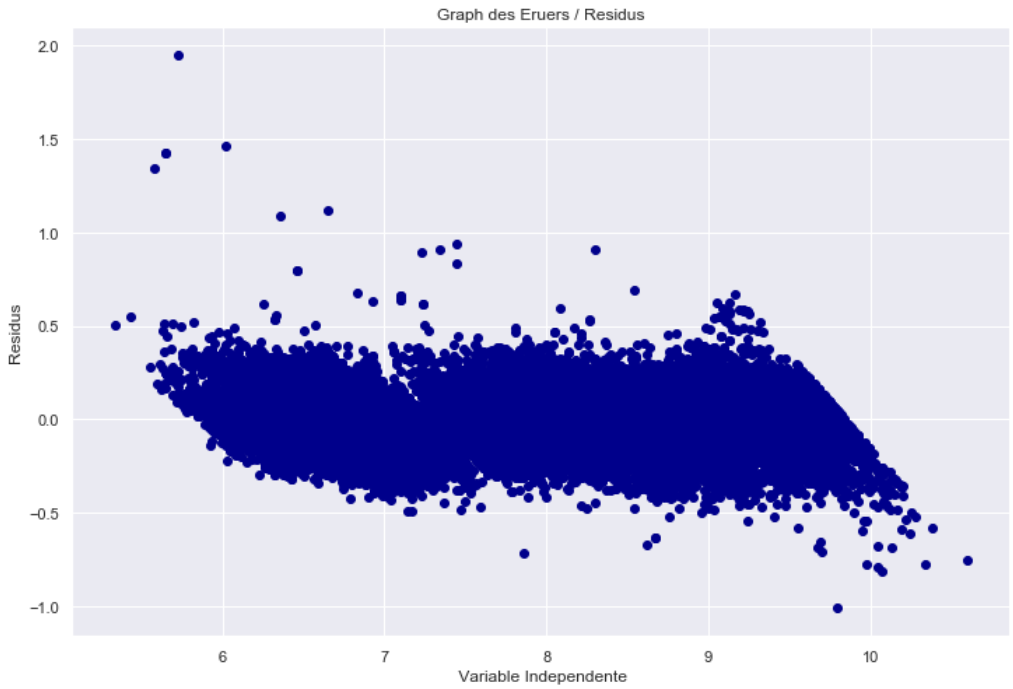
Test de Durbin-Watson

- Le modèle est très puissant avec un R_carré ajusté = 0.983, Donc on garde ce modèle
- Toutes les variables sont significatives avec une p_value < 0.05
- Le modèle est globalement significatif étant donné que la p_value de la f_statistic (statistique de Fisher) = 0.00 < 0.05.
- Le test de Durbin-Watson a une valeur de 1.245 => il existe peut-être une autocorrélation positive des erreurs

4.4. Examination des Erreurs (Résidus)



La distribution des erreurs suit une loi Normale avec une moyenne = 0
=> La Normalité est vérifiée



Le graph ne montre pas des signes d'heteroscedasticité,
=> Les erreurs sont homoscedastiques.

V. Bibliographie.

- Kaggle, https://www.kaggle.com/shivam2503/diamonds

- Régis Bourbonnais, Économétrie : Cours et exercices corrigés, 9e édition, Dunod, 2015
- Damodar N. Gujarati et Dawn C. Porter, Basic Econometrics, Fifth Edition, 2008