RAPPORT DE PROJET

Data visualisation

Préparé par **:**
Saifeddin El hairech

Sous la direction de:

Youssef karim El alaoui

Année Universitaire

2020-2021

I. La Data :

Notre dataset contient des informations sur le prix des diamants, et aussi sur des nombreux autres attribues, parmi lesquels il y'en a ceux qui sont connu d'être capable d'influencer le prix (carat, cut, clarity et color), et aussi quelque paramètres physiques (depth, table, Price, x, y, z).

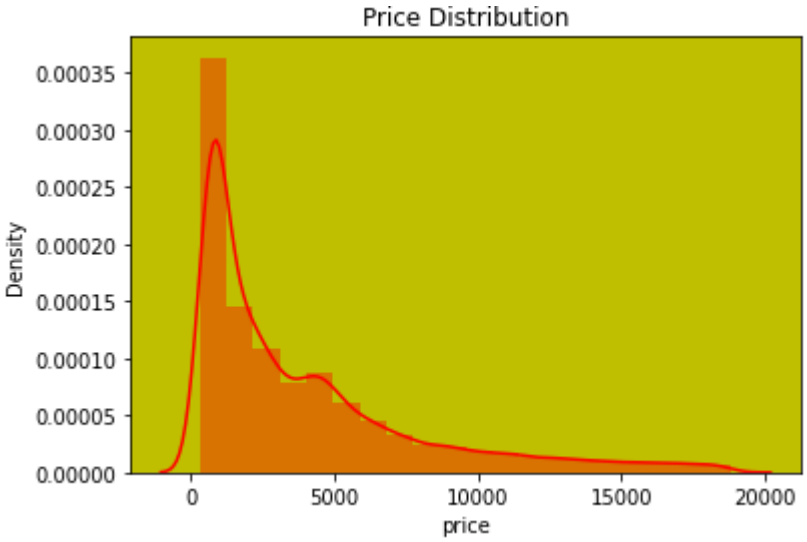
Ce dataset contient envers 54000 observations (informations sur 53940 diamants)

- L'exploration des variables (individuellement):

On commence avec une étude uni variée :

Price :

C'est une variable quantitatif de valeurs continue, il signifie le prix des diamants, l'étude de sa distribution nous donne :



On voit qu'il y'en a beaucoup moins de data pour les valeurs de 'Price' plus de 5000 par rapport à ceux qui sont moins de 5000 §. Beaucoup de diamants ont un prix inférieur à 2500 §.Ceci peut être justifié par le fait que les diamants de prix moins que 5000 § sont plus demander et moins couteuse et alors sont les plus fréquent.

Sommaire de statistique descriptif de ce variable :

```
count    53940.000000
mean      3932.799722
std       3989.439738
min        326.000000
25%        950.000000
50%       2401.000000
75%       5324.250000
max      18823.000000
Name: price, dtype: float64
```

count	mean	std	min	25%	50%	75%	max	var
53940	3932.799722	3989.439738	326	950	2401	5324.25	18823	15915629.42430147

On déduit de ce tableau que:

- le prix minimal est 326 et le prix maximal est 18823

- 25% des diamants ont des prix inférieurs à 950 §

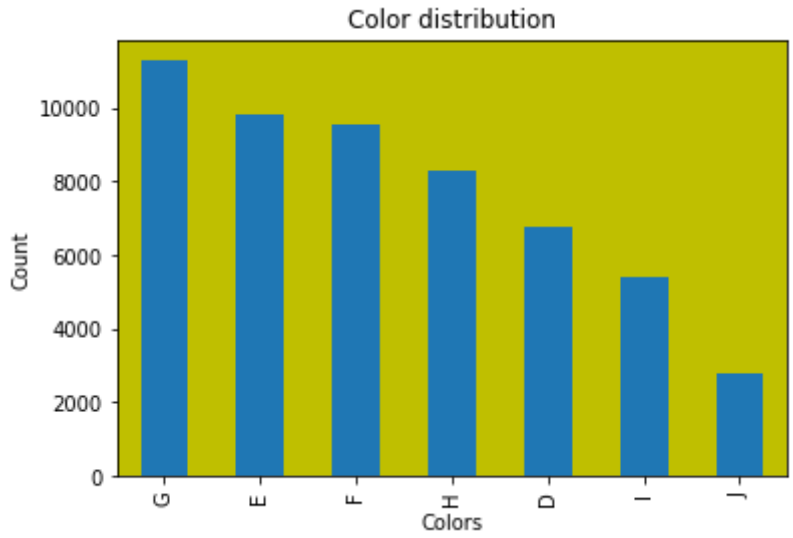
-50% des diamants ont des prix inférieurs à 2401 §

-75% des diamants ont des prix inférieur à 5324 §

-les valeurs de la variable 'Price' varie (variance très important)

-Color :

'Color' est une variable qualitatif, il signifie le grade du couleur des diamants et sa qualité ; l'étude de sa distribution nous donne :



On voit donc la distribution de la data concernant le couleur ; G est le grade couleur le plus fréquent dans la data puis E et F et ainsi de suite. G est donc le grade de couleur le plus utilisé dans les diamants de notre dataset.

Sommaire de statistique descriptif de ce variable :

```
count     53940
unique        7
top           G
freq      11292
Name: color, dtype: object
```

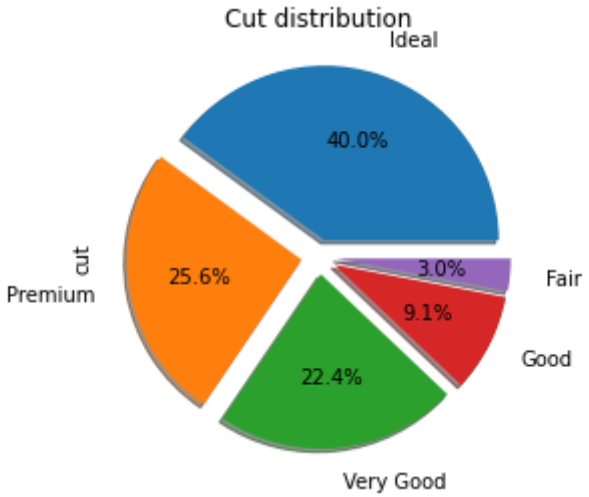
count	unique	top	freq
53940	7	G	11292

On déduit de ce tableau que :

- que g est le plus fréquent parmi les grades de couleurs
- la fréquence de g est 11292

-Cut :

‘Cut’ est une variable qualitatif, il signifie la qualité de découpage d’un diamant, l’étude de sa distribution donne :



On voit que la valeur ‘Ideal’ du variable Cut est la plus fréquente dans notre data suivi par ‘premium’ et ‘very good’, c’est trois valeurs constituent plus de 88% de notre data. Cela est justifier par le fait que les diamants les plus demander sont ceux avec la plus bonne qualité.

Sommaire de statistique descriptif de ce variable :

```
count     53940
unique        5
top       Ideal
freq      21551
Name: cut, dtype: object
```

count	unique	top	freq
26408	5	Ideal	8796

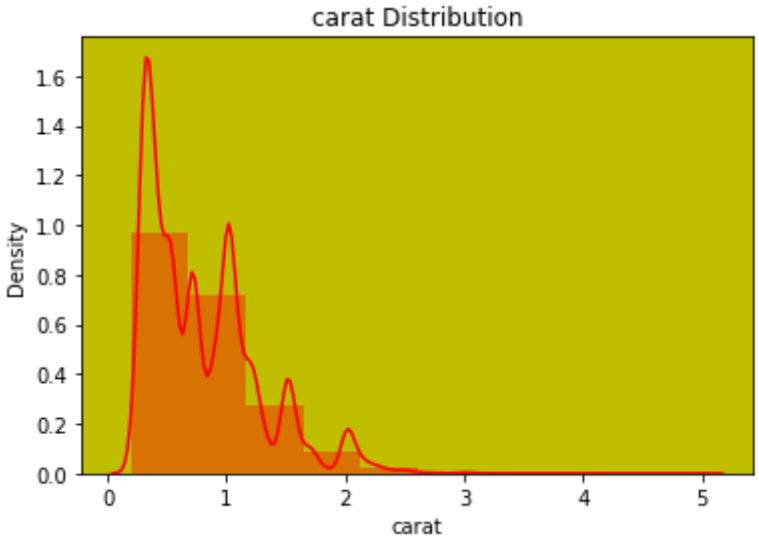
On déduit du tableau que :

- les diamants avec valeur ‘Ideal’ de variable ‘cut’ sont les plus fréquents.

- le fréquence de ‘Ideal’ est 8796

-Carat :

‘Carat’ est une variable quantitatif à valeurs continue, il représente l’unité de mesure du poids des diamants, l’étude de sa distribution donne :



On voit qu'il y'en considérablement moins de data pour les diamants de Carat supérieur à 1 par rapport à celle avec des valeurs de Carat inférieur à 1. Beaucoup de diamants ont de 0.3 vers 0.7 carats. La raison peut être que Les diamants de carat inferieurs à 1 sont très demandés de plus il est couteux de produire des diamants avec des valeurs de carat supérieur à 1et donc le bijoutier choisi ces poids.

Sommaire de statistique descriptif de ce variable :

```
count    53940.000000
mean         0.797940
std          0.474011
min          0.200000
25%          0.400000
50%          0.700000
75%          1.040000
max          5.010000
Name: carat, dtype: float64
```

Count	Mean	std	min	25%	50%	75%	Max	Var
53940	0.79794	0.474011	0.2	0.4	0.7	1.04	5.01	0.22468665982274233

On déduit du tableau que :

- le poids minimal est 0.2 alors que le poids maximal est 5.01, le poids moyen est 0.797 et le 'sample standard diviation' est 0.474.

- 25% des diamants ont des valeurs de carat inférieur à 0.4

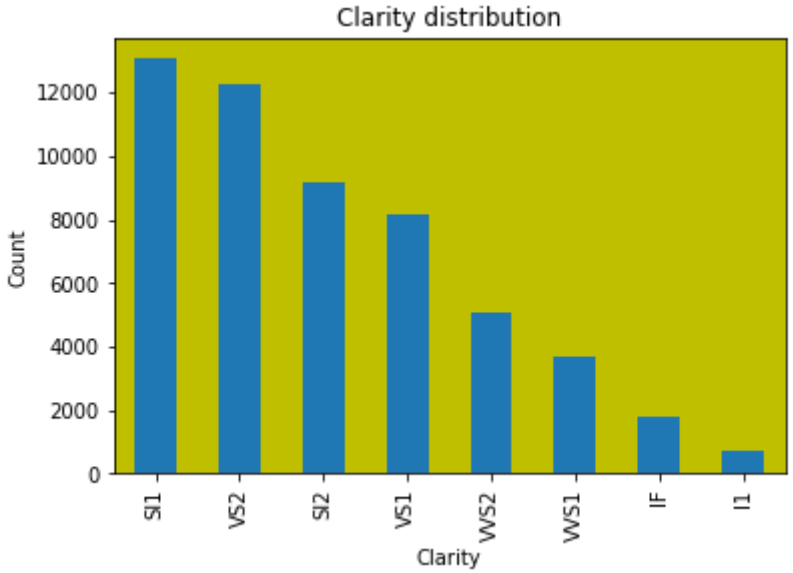
-50% des diamants ont des valeurs de carat inférieur à 0.7

-75% des diamants ont des valeurs de carat inférieur à 1.04

-les valeurs de la variable ‘Carat’ ne sont pas très varié (variance non significatif)

-Clarity :

‘Clarity’ est une variable qualitatif, il représente la pureté et les inclusions d’un diamant l’étude de sa distribution donne :



On voit que le type ‘SI1’ (slightly included) est le plus fréquent dans notre dataset, suivi par ‘VS2’ (very slightly included) et ‘SI2’. On remarque qu’on n’a pas beaucoup de data concernant ‘VVS2’ qui est considérer quasi-parfait ca est probablement dû au fait qu’il est difficile de produire ce type de ‘clarity’.

Sommaire de statistique descriptif de ce variable :

```
count     53940
unique        8
top         SI1
freq      13065
Name: clarity, dtype: object
```

Count	unique	top	freq

53940	8	SI1	13065

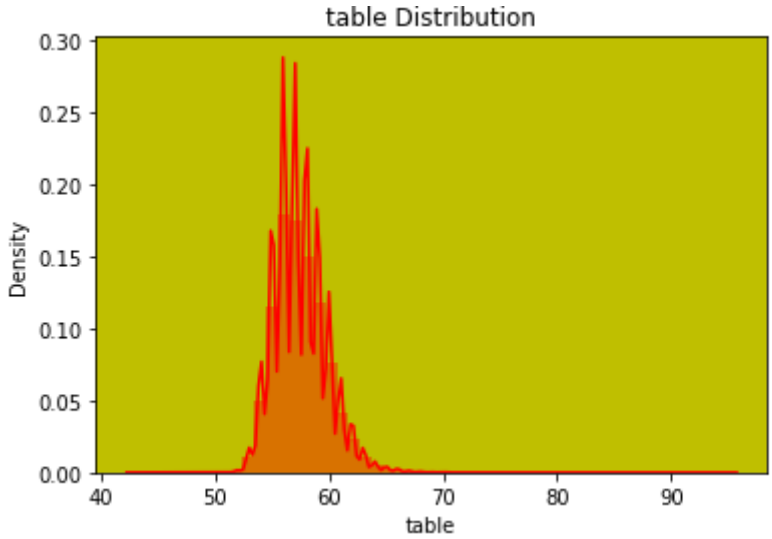
On déduit du tableau que :

- Les diamants avec des valeurs ‘SI1’ de clarity sont les plus fréquents.

- le fréquence de ‘SI1’ est 13065

Table :

C’est un variable quantitatif à valeurs continue, il juge aussi la qualité d’un diamant l’étude de sa distribution donne :



On voit que notre data sur ce variable est concentré dans des valeurs compris entre 54% et 65%.

Sommaire de statistique descriptif de ce variable :

```
count    53940.000000
mean        57.457184
std          2.234491
min         43.000000
25%         56.000000
50%         57.000000
75%         59.000000
max         95.000000
Name: table, dtype: float64
```

count	mean	std	min	25%	50%	75%	max	var
53940	57.457184	2.234491	43	56	57	59	95	4.9929

On déduit de ce tableau que :

-25% des diamants ont une valeur de table inférieure à 56%.

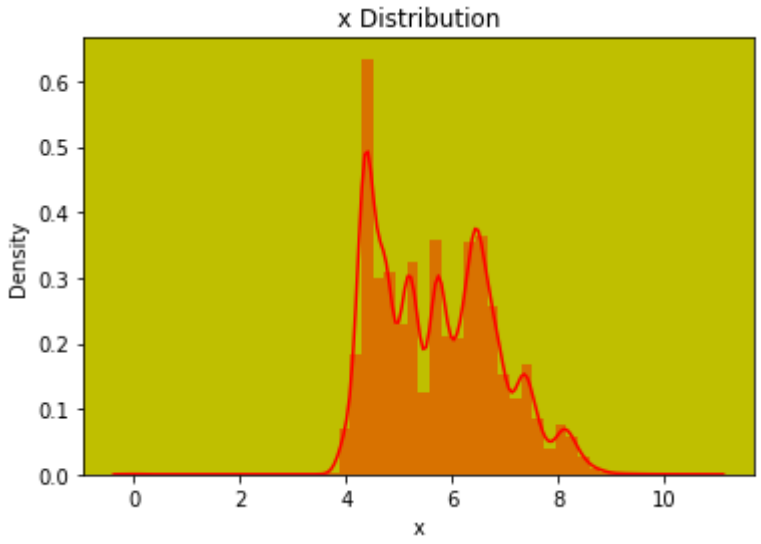
-50% des diamants ont une valeur de table inferieure 57%.

-75% des diamants ont une valeur de table inférieure à 59%.

-le variable table n'est pas très variée (valeur de var n'est pas importante

-x :

C'est un variable quantitatif à valeurs continue, l'étude de sa distribution nous donne :



On voit qu'on a significativement moins de data pour les diamants aves valeur de x supérieur à 7 qu'on a sur celle inférieure à 7. La même chose pour les valeurs inférieures à 4. Cela peut être dû à la difficulté de produire quelque diamant avec ces mesures.

Sommaire de statistique descriptif de ce variable :

```
count    53940.000000
mean         5.731157
std          1.121761
min          0.000000
25%          4.710000
50%          5.700000
75%          6.540000
max         10.740000
Name: x, dtype: float64
```

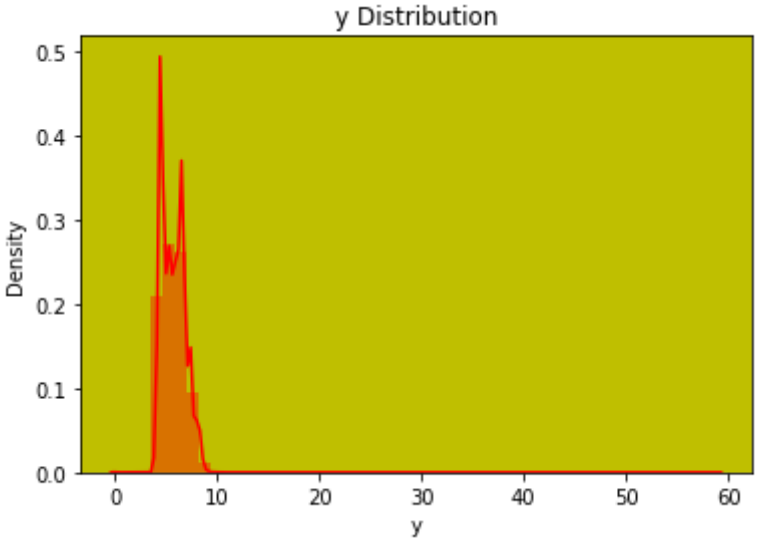
Count	Mean	Std	Min	25%	50%	75%	Max	Var
53940	5.7311	1.1217	0	4.71	5.7	6.54	10.74	1.2583

On déduit du tableau que :

- Le minimum de x est 0 et le maximum est 10.74.
- La valeur moyenne de x est 5.73 et son std est 1.12
- 25% des diamants ont des valeurs de x inférieure à 4.71.
- 50% des diamants ont des valeurs de x inférieure à 5.7.
- 75% des diamants ont des valeurs de x inférieure à 6.54.
- X est peu variée.

y :

C’est un variable quantitatif à valeurs continue, l’étude de sa distribution nous donne :



On voit que notre data est concentré entre 3 et 9 pour les valeurs de y. probablement pour des raisons similaires à celle de x.

Sommaire de statistique descriptif de ce variable :

```
count    53940.000000
mean         5.734526
std          1.142135
min          0.000000
25%          4.720000
50%          5.710000
75%          6.540000
max         58.900000
Name: y, dtype: float64
```

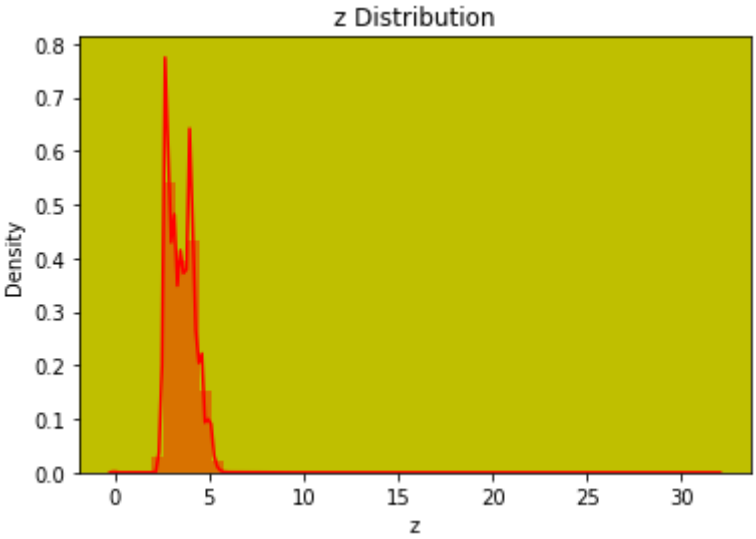
Count	Mean	Std	Min	25%	50%	74%	Max	var
53940	5.73	1.14	0	4.72	5.71	6.54	58.9	1.3044

On déduit du tableau que :

- Le minimum est 0 et le maximum 58.9.
- La valeur moyenne est 5.73 et le std est 1.14.
- 25% des diamants ont une valeur de y inférieure à 4.72.
- 50% des diamants ont une valeur de y inférieure à 5.71.
- 75% des diamants ont une valeur de y inférieure à 6.54.
- Y est peu variée.

z :

C’est un variable quantitatif à valeurs continue, l’étude de sa distribution nous donne :



On voit que pour les valeurs de la variable z, notre data est importante seulement dans un intervalle entre 2 est 4. Cela est probablement dû aux mêmes raisons que celle des variables x et y.

Sommaire de statistique descriptif de ce variable :

```
count    53940.000000
mean         3.538734
std          0.705699
min          0.000000
25%          2.910000
50%          3.530000
75%          4.040000
max         31.800000
Name: z, dtype: float64
```

Count	Mean	Std	Min	25%	50%	75%	Max	Var
53940	3.538	0.705	0	2.91	3.53	4.04	31.8	0.49

On déduit du tableau que :

- 25% des diamants ont une valeur de z inférieure à 2.91.
- 50% des diamants ont une valeur de z inférieure à 3.53.

- 75% des diamants ont une valeur de z inférieure à 4.04 .
- z est peu varié.

Sommaire de statistique descriptif de tous les variables:

	ID	carat	depth	table	price	x	y	z
count	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000
mean	26970.500000	0.797940	61.749405	57.457184	3932.799722	5.731157	5.734526	3.538734
std	15571.281097	0.474011	1.432621	2.234491	3989.439738	1.121761	1.142135	0.705699
min	1.000000	0.200000	43.000000	43.000000	326.000000	0.000000	0.000000	0.000000
25%	13485.750000	0.400000	61.000000	56.000000	950.000000	4.710000	4.720000	2.910000
50%	26970.500000	0.700000	61.800000	57.000000	2401.000000	5.700000	5.710000	3.530000
75%	40455.250000	1.040000	62.500000	59.000000	5324.250000	6.540000	6.540000	4.040000
max	53940.000000	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

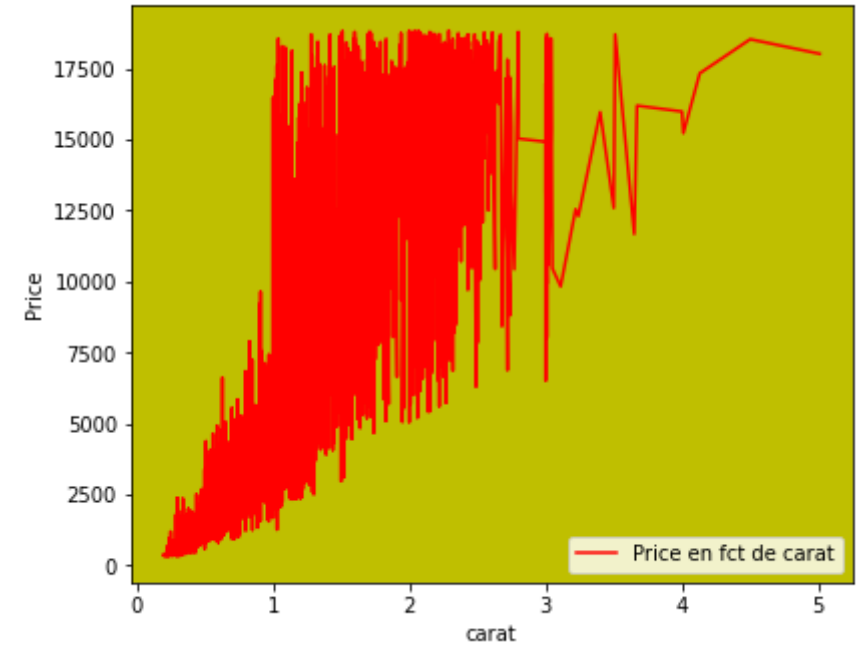
Maintenant après cet étude uni varié ont connu notre data. Ce tableau résume toute l'étude de statistique descriptif de la dataset (variables numériques).

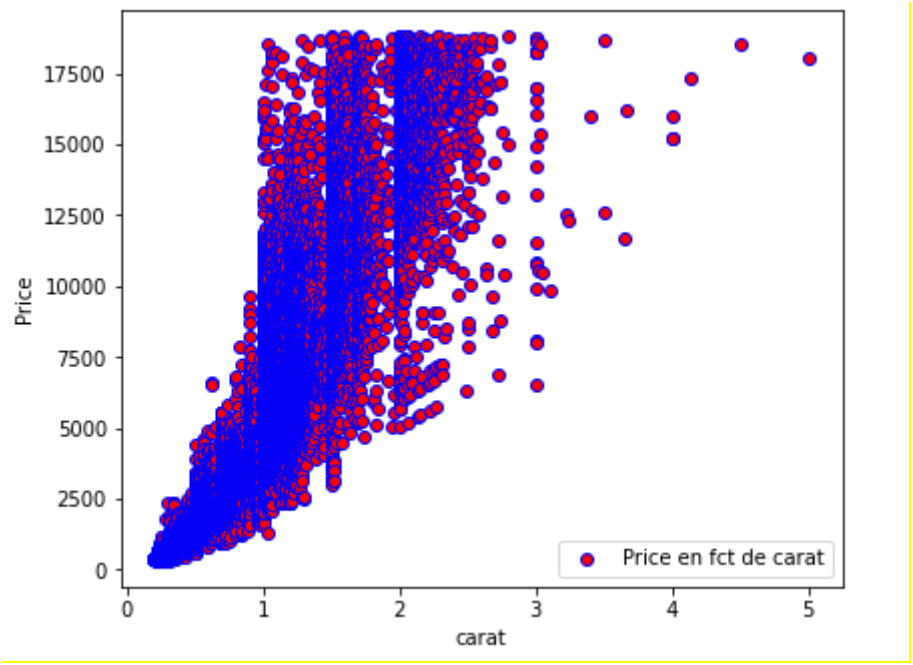
II. ETUDE DU RELATIONDU PRIX AVEC LES AUTRE ATRIBUE :

1. Price avec les variables numériques :

On commence avec l'étude bi varié

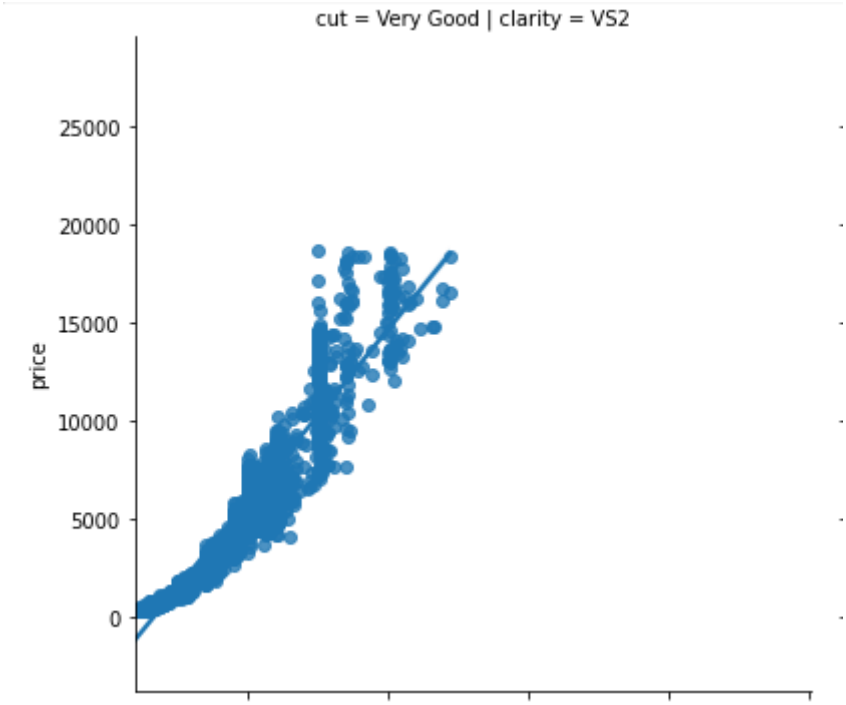
-Carat :





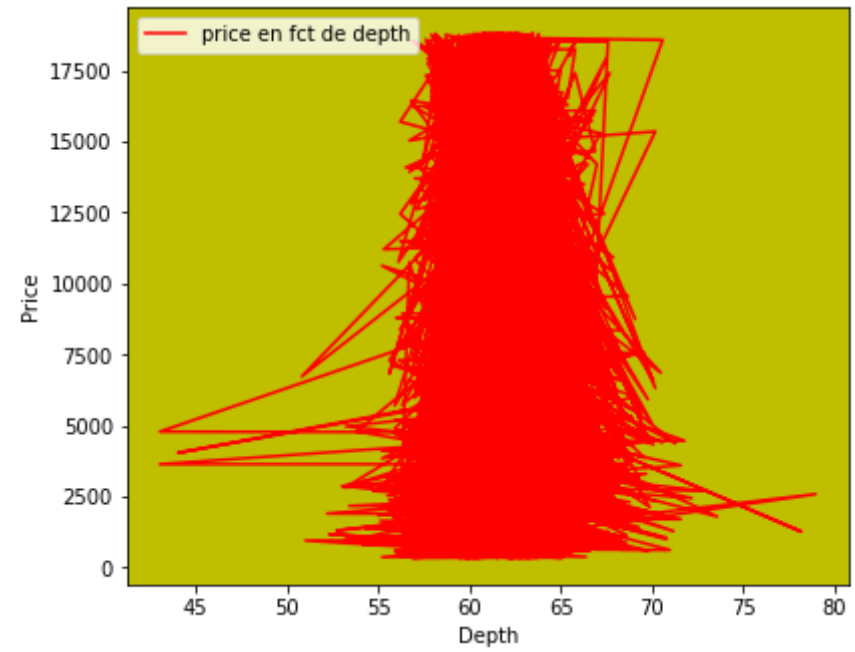
D’abord Le chaos dans cette visualisation est dû aux autres variables qui influencent le prix, mais comme on voit dans les deux visualisations (line plot, scatter plot), généralement la valeur de carat influence le prix positivement (si le carat augment alors le prix augment aussi) et son influence et très important. Ce résultat est évident car le carat est le plus important aspect d’un diamant, il exprime sa rareté.

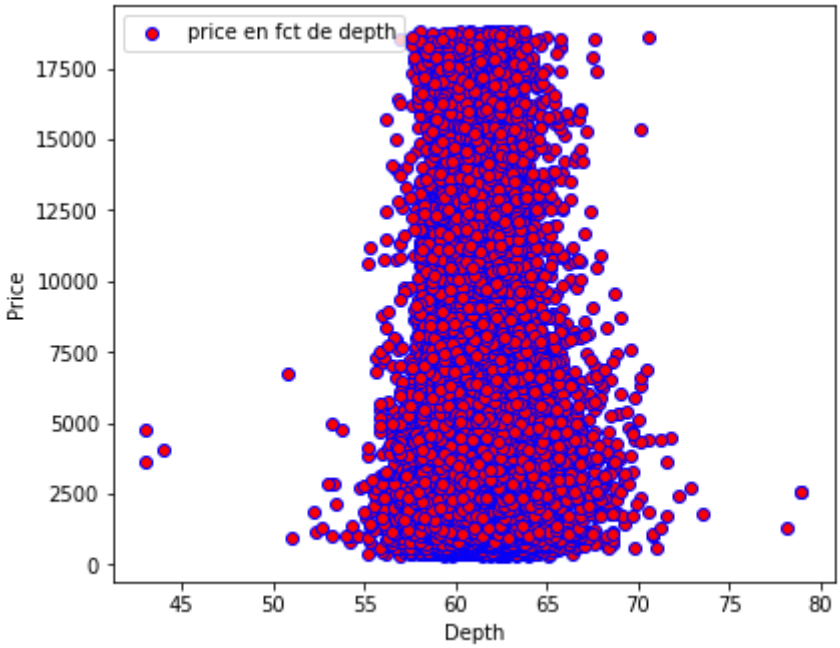
Pour plus justifier l'influence positive du carat sur Price, on réalise le plot suivant de l'évolution du prix en fonction de carat en fixant une valeur de cut et de clarity par exemple:



Comme on voit il y'en a moins de chaos car on fixé deux autres variables qui influence le prix à savoir cut et clarity (on verra ca encore). L'influence du carat est donc assurée.

-Depth :

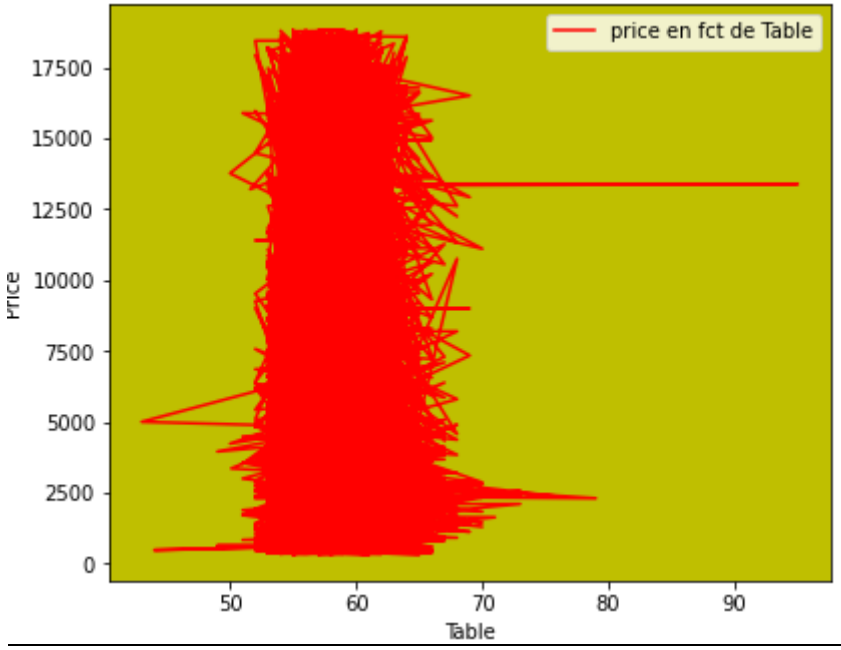


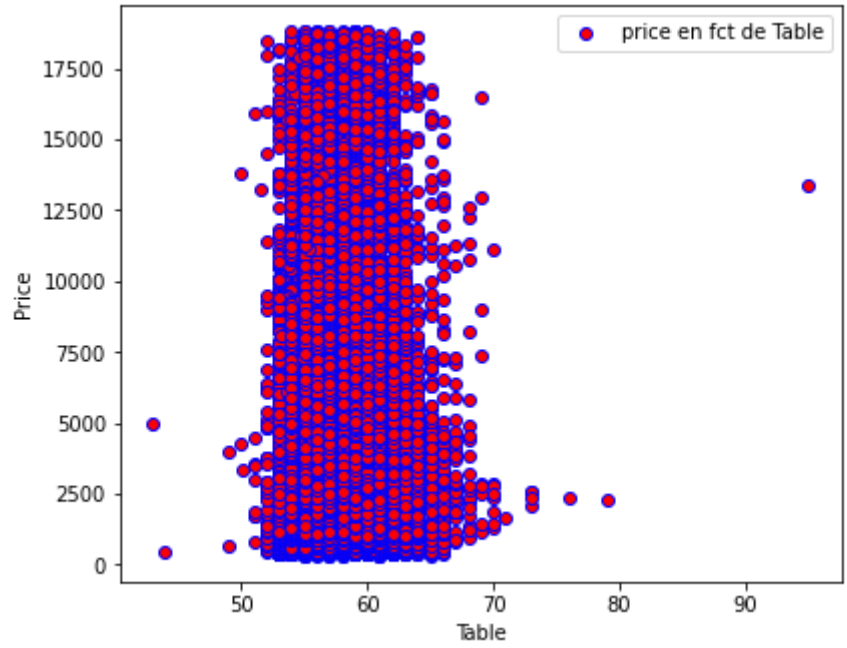


Alors on voit dans ces deux visualisations que le prix est maximal pour des valeurs compris entre 58 et 68, un diamant avec un depth médiocre ou très grands affect son prix négativement. Ce résultat est aussi évident car le depth (profondeur) affect l'apparence d'un diamant spécialement lorsqu'il est exposé à la lumière ; Un diamant court avec une faible profondeur permettra à la plupart de la lumière de passer, réduisant ainsi l'étincelle et le feu visibles lorsqu'il est

exposé à la lumière. Cela peut également se produire lorsqu'un diamant est trop profond.

-table :

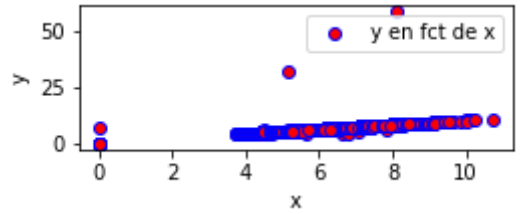


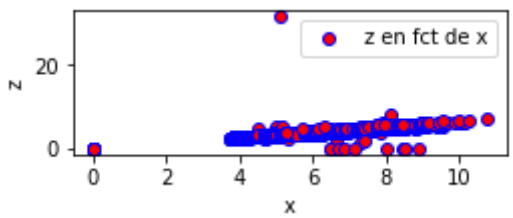


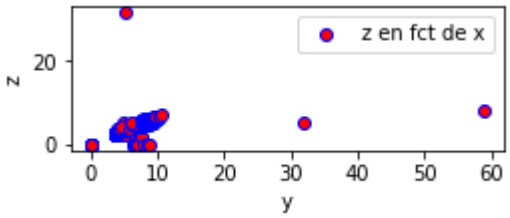
Alors on voit dans les deux visualisations que le prix est maximal pour des valeurs de Table compris entre 54 et 64. La taille de la table est l'un des facteurs les plus importants dans l'apparence de tout diamant.

x, y and z :

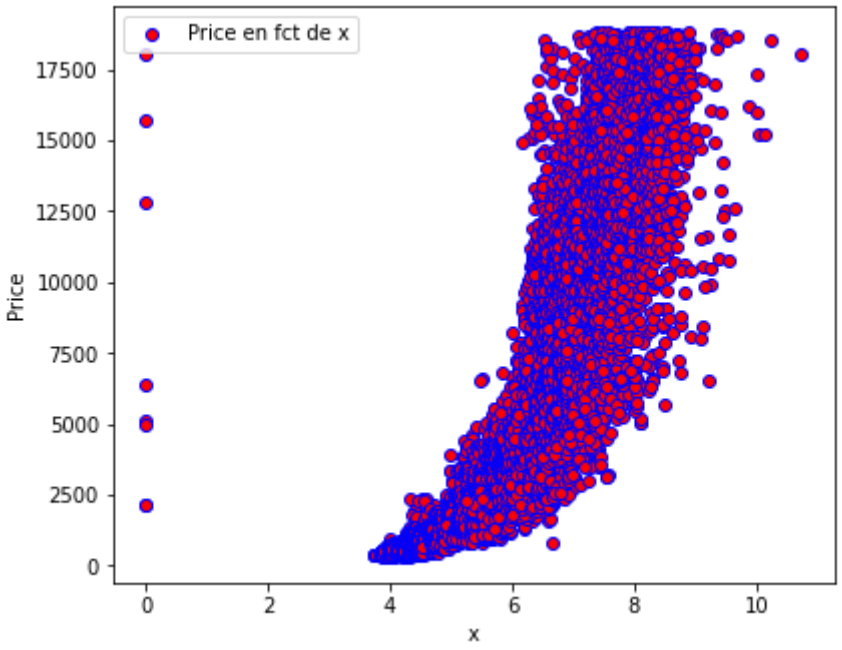
D'abord on commence par une étude de proportionnalité entre les trois variables x et y et z pour faciliter le travail :

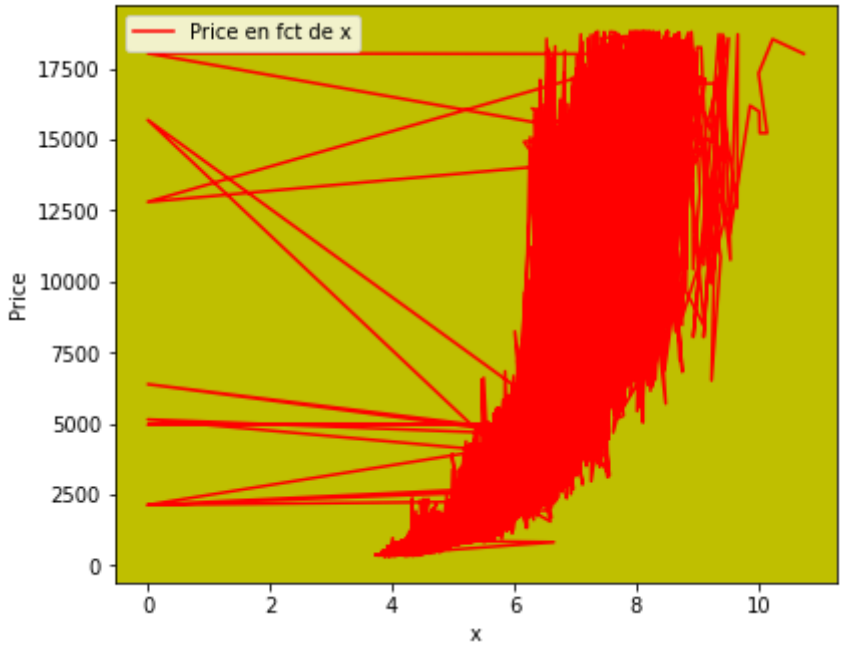






Ces trois scatter plots montrent que x et y et z sont proportionnels 2 en 2, alors il suffit de choisir l'un entre eux pour étudier l'influence sur le prix (dans ce cas x):



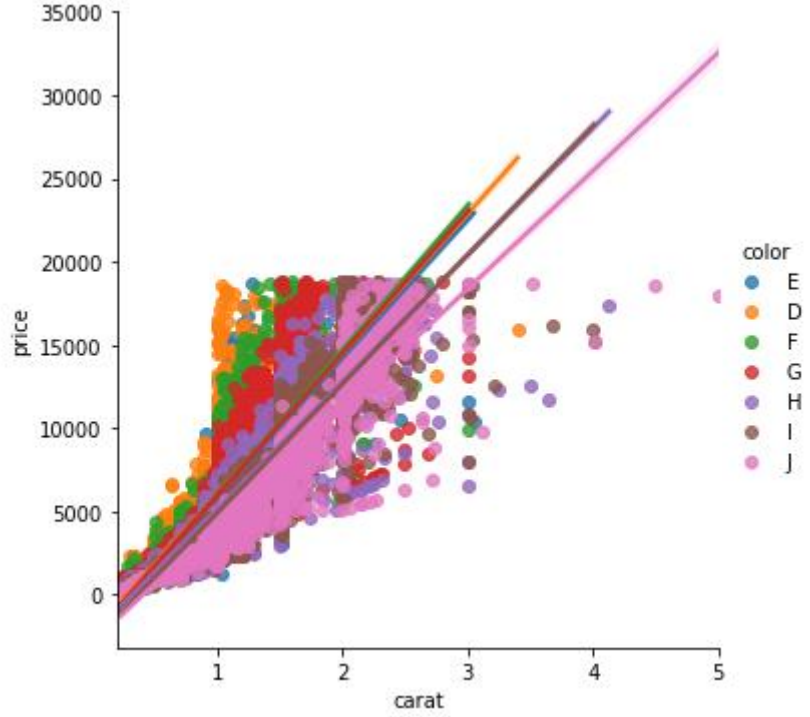


On voit donc dans ces deux visualisations que généralement x influence Price positivement (lorsque x augment le prix augment aussi). On verra après la relation entre ces variables et le depth.

2. Price avec les variables catégoriques :

-Color :

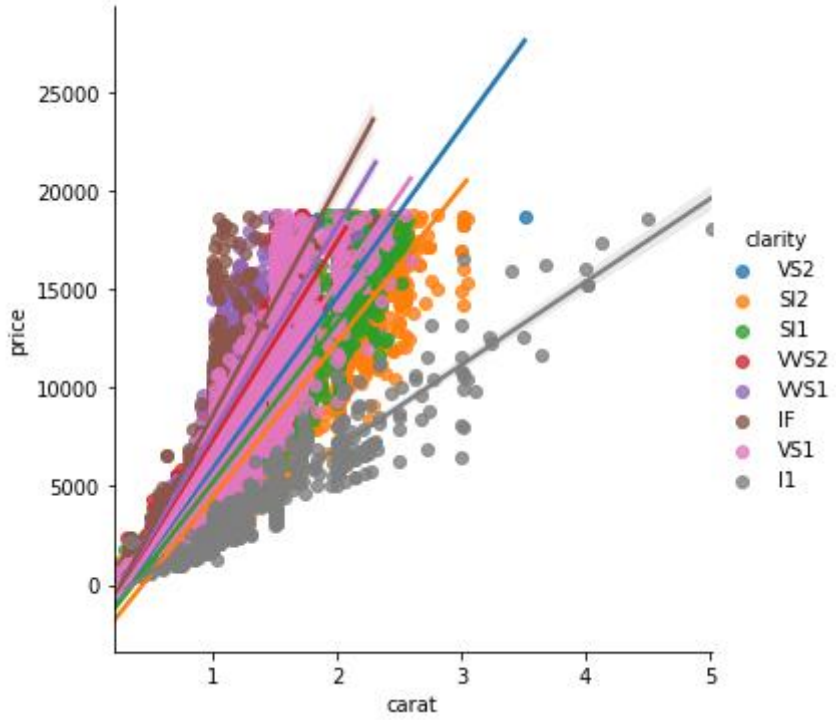
Pour étudier la relation entre Color qui est une variable qualitatif et Price qui est quantitatif, on va utiliser le scatterplot suivant :



La visualisation est une Price-carat scatter, et les couleurs sur les points représentent les différents grades de couleur de diamants. D'un part la figure montre la relation qu'on a montrée entre le prix et le carat, d'autre part il montre qu'une couleur plus transparente conduit généralement à des prix plus élevés. Alors la relation entre le couleur et le prix est assuré.

-Clarity :

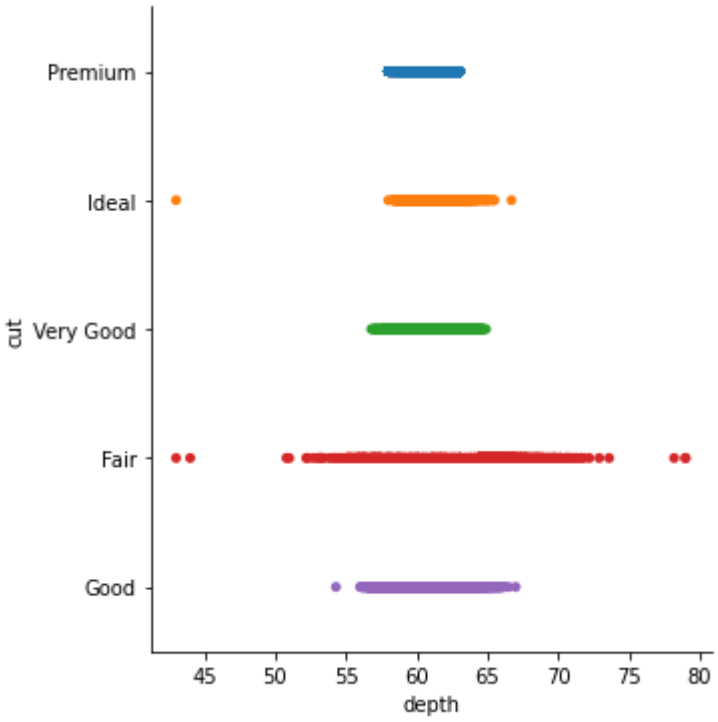
Par le même démarche, clarity est une variable qualitatif alors on utilise le scatterplot suivant pour étudier la relation entre clarity et Price :



La visualisation est une Price-Carat scatter, les points sont colorés par rapport à la clarté des diamants. Une clarté plus pure conduit généralement à des prix plus élevé. Alors la relation entre la clarté et le prix est assuré.

-Cut :

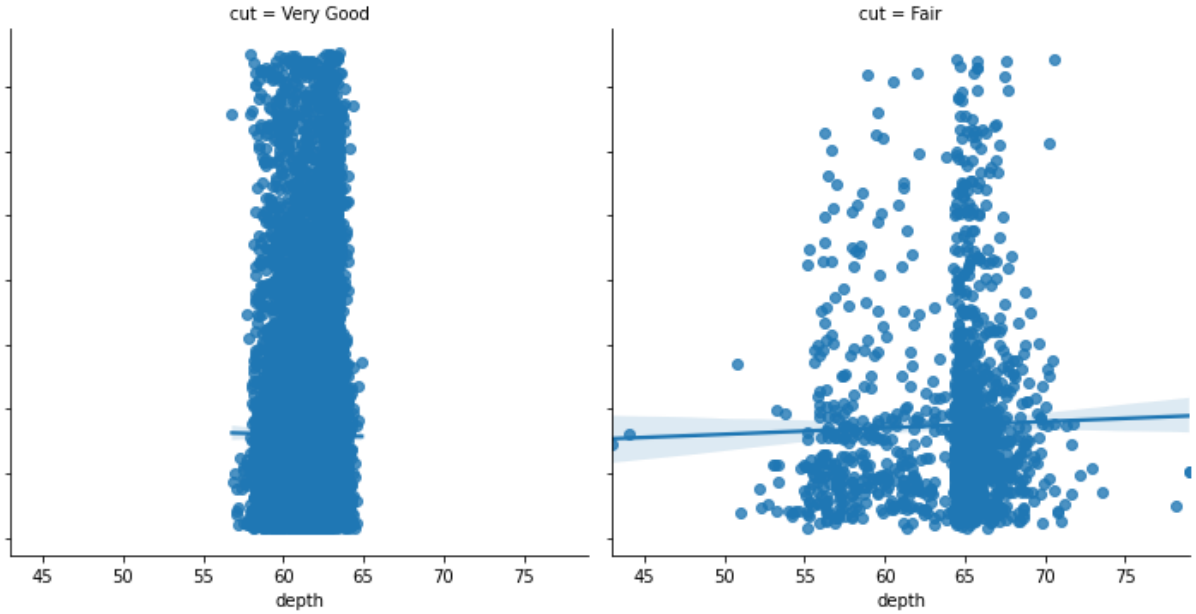
Pour trouver une relation entre cut est Price il suffit de trouvé une relation entre Cut et Depth,



On a montré d'abord que le nombre de diamants avec cut='Ideal' est plus important que le nombre de diamants avec cut='fair', on a aussi montré que le prix augment dans un intervalle de depth compris entre 55 et 65.

La visualisation **depth-cut scatter** montre que plus depth est proche de cet intervalle de [55,65] plus probable que la cut='ideal' et alors plus la coupe est bon. Et donc ici se voit la relation entre le depth et cut ; cut influence depth.

On peut visualiser cette dépendance avec une autre méthode. Dessous on voit deux Price-depth scatters pour deux différents types de cut:



Alors, de tous ca on peut conclure que Cut influence Price positivement (plus cut est bon plus depth reproche de l'intervalle idéal plus le prix augment).donc la relation entre cut et le prix et assuré.

III. Simplification des résultats :

Dans cette partie on va essayer de trouver les attribue principales qui influence le prix. Donc il s'agit principalement d'éliminé quelque variables en justifiant ca

- x, y, z, Depth et Cut :

On a trouvé d'abord que x, y, z sont proportionnels 2 en 2.

Puis, Puisqu'on définit depth par depth = z / mean(x, y) alors il n'est pas nécessaire de visualiser la relation entre depth et x ou depth et y ou depth et z.

Enfin, puisqu'on a trouvé que Cut influence Depth et vice versa. On dit que x, y, z et depth sont des attribues de Cut et ***on prend Cut comme attribue influenceur (positivement) du prix (Price).***

- Clarity

On a prouvé que clarity influence (positivement) le prix (Price). Plus la clarté des diamants est pure plus le prix augment.

- Carat :

On a prouvé que Carat influence (positivement) le prix (Price). Plus la valeur du carat est important plus le prix augment. La visualisation de cette relation montre que ce variable est le plus important dans l'évolution du prix.

- Color :

On a prouvé que Color influence (positivement) le prix (Price). Plus le Couleur soit de grade haute plus le prix augmente (plus le diamant sont transparent plus le prix augment)

Sommaire :

Pour conclure le prix (Price) est influencé par quatre attribues (4C's) à savoir ***le couleur*** (Color), ***le carat*** (Carat), ***la clarté*** (Clarity) et ***la coupe*** (Cut)

IV. Analyse :

Le diamant est la plus belle et la plus rare des gemmes, il n'est pas difficile de comprendre pourquoi il représente un sujet fascinant pour beaucoup des gens, le prix des diamants est aussi un sujet très intéressant et le but de cette étude était d'utiliser la dataset « bijouterie » afin de visualiser la relation s'il existe entre le prix des diamants et ces divers attribues à savoir la coupe (cut), la clarté (clarity), Le couleur (color), le carat (carat), la profondeur (depth), la table (Table), x, y et z. en effet le but est de prouvé que le prix est affecté par les 4C's (cut, clarity, color et carat).

La dataset « bijouterie » utilisé dans ce projet contient les informations de 53 940 diamants, et heureusement il y'en a pas des donnés manquant. D'abord on à commencer avec **une étude uni variée** pour mieux comprendre notre data et sa distribution. Dans cette étape on a trouvé que le prix des diamants dans notre dataset est distribué d'une façon que la plupart de ses valeurs sont inférieurs à 5000 §, ensuite pour le couleur on a trouvé que les meilleurs grades de couleur G, H, J et I constitue plus de la moitié de notre dataset (51%), pour la coupe on a trouvé que les meilleures états de coupe (Idéal, premium et très bon) constitue 88%, pour le carat on a trouvé que la majorité de notre data concerne des diamants de poids de carat inférieur à 1, pour la clarté on a trouvé qu'il y'en plus de diamants avec une clarté de type SI2 que de type VS1 (qui est le meilleur parmi les types), enfin les valeurs de table, x, y et z des diamants de notre dataset sont concentré sur des intervalles respectifs [54%-65%], [4-7], [3-9], [2-4], on a compris beaucoup de chose à propos du data mais ce qu'on peut conclure de cette étape est que le bijoutier a choisi sa marchandise en prenant en considération la qualité mais aussi le prix de production, il essaye de maximisé son profit en vendant le meilleur qualité possible avec un cout de production raisonnable. Ensuite dans une deuxième étape on commence **une étude bi**

variée entre le prix et ses attribues pour visualiser les relations s'ils existent, on a commencé avec les variables numérique d'abord dont on a pu visualiser une relation entre le prix et le carat (le carat influence le prix positivement) en utilisant des scatters et line plots et cette relation se voit lieux lorsqu'on un ou deux des autres 4C's,ensuite on a trouvé que le prix se maximise dans une intervalle du profondeur (depth) [57,64],et la même chose pour la table dans une intervalle [55,65], ensuite pour les variables x, y et z d'abord on a trouvé qu'ils sont fortement proportionnels entre eux à l'aide de quelque scatterplots puis qu'il influence le prix positivement. De même manière pour les variables quantitatives on a trouvé que la qualité de clarté et le grade du couleur influence aussi le prix positivement, ensuite on a visualisé une relation entre la coupe (cut) et la profondeur (depth) (on a trouvé que le plus la profondeur s'approche de l'intervalle qui maximise le prix plus la coupe serait bonne) et comme ça on a prouvé que la coupe influence le prix positivement après ça on a utilisé la relation mathématique entre la profondeur et les trois variables x, y et z pour choisir enfin la coupe comme représentante de ses relations. En finit cette étude bi varié en concluant qu'il y'en a une fonction qui relie le prix avec la coupe du diamant, sa qualité de clarté, le grade de sa couleur et son carat.

Les graphiques de visualisation de données peuvent nous donner un aperçu des relations des données et ainsi que le rang d'importance de chaque variable. Dans notre étude on a réussi à conclure l'existence d'une telle relation pour le prix du diamant et ses attribues et on aussi conclue que le carat et l'influenceur le plus important du prix.