Projet MSPL: Diamants

Lilian OURAHA, Doriane RICHARD, Reza ROUDOCI, Ahmed BEN NACIB

20 Mai 2025

Contents

Introduction

Cette étude a été réalisée par des étudiants de L3 MIAGE dans le cadre du module MSPL. Le sujet de cette étude porte sur les diamants.

L'objectif de ce projet est d'identifier le ou les facteurs clés qui déterminent le prix d'un diamant parmi plusieurs variables telles que son carat, sa couleur, sa largeur et d'autres caractéristiques.

Pour cela, nous avons utilisé un jeu de données trouvé sur Kaggle datant de 2017. Il contient plus de 58 000 observations et une dizaine de variables décrivant les caractéristiques physiques et esthétiques des diamants.

Les variables sont les suivantes :

- **price**: Le prix du diamant en dollars ($326–$18,823);
- **carat**: Le poids du diamant en grammes (un carat = 0,2 gramme) (0.2–5.01);

- **cut**: La qualité de la taille du diamant, qui influence fortement la manière dont il reflète la lumière (Fair, Good, Very Good, Premium, Ideal);

- **color**: La couleur du diamant varie du plus coloré au plus incolore ('J' (le moins bon) à 'D' (le meilleur)). Les diamants totalement incolores sont les plus rares. D'autres couleurs naturelles (comme le bleu, rouge ou rose) sont appelées "fancy" et leur évaluation se fait selon des critères différents;
- **clarity**: Un indice mesurant la pureté du diamant (I1 (le pire), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (le meilleur));
- **x**: Longueur en mm (0–10.74);
- **y**: Largeur mm (0–58.9);
- **z**: Profondeur en mm (0–31.8);
- **depth**: Profondeur totale du diamant, exprimée en pourcentage par rapport à son diamètre moyen = z / ((x + y) / 2) = 2 * z / (x + y) (43–79);
- **table**: Largeur de la facette supérieure, exprimée en pourcentage par rapport à la largeur du diamant (43–95).

À partir de ce jeu de données, nous chercherons à répondre à la question suivante :

"**Quel est le facteur le plus déterminant du prix d'un diamant ?**"

Methodologie

Procédure de nettoyage des données

Avant de mener notre analyse, nous avons effectué plusieurs étapes de nettoyage afin de garantir la qualité et la fiabilité des données :

- La première colonne du jeu de données était une variable d'index inutile, que nous avons supprimée, car elle n'apportait aucune information pertinente;
- Nous avons identifié et supprimé 20 lignes contenant des mesures physiques impossibles. Certaines lignes avaient une valeur de 0 pour l'une des dimensions (x, y ou z), ce qui est physiquement impossible pour un diamant réel.

Démarche scientifique

Nous avons d'abord observé que le carat (poids) est de loin le facteur le plus influent sur le prix d'un diamant.

Afin d'évaluer plus précisément l'influence des autres variables, il était nécessaire de contrôler l'effet du carat, c'est-à-dire de le neutraliser pour comparer les autres critères dans des conditions équitables.

Pour cela, nous avons choisi de nous concentrer sur les diamants dont le carat est proche de 1. Ce choix s'explique par leur forte présence dans le jeu de données, mais aussi par le fait qu'un poids d'environ un carat offre une bonne diversité au niveau des autres caractéristiques. Par exemple, certaines coupes ou niveaux de clarté sont plus rares, voire inexistants, pour les très faibles carats. À noter que ce sont justement les diamants de faible carat qui sont les plus présents dans le jeu de données.

En nous limitant à cette plage de carat, nous avons pu analyser l'impact des autres caractéristiques comme la taille (cut), la pureté (clarity) et la couleur (color) sur le prix.

Nous avons également conclu que, parmi les mesures physiques (*x, y, z, depth, table*), c'est le carat qui reste le plus déterminant pour le prix.
Ainsi, nous avons choisi de ne pas approfondir davantage les variables dimensionnelles individuelles (longueur, largeur et profondeur), et nous avons décidé de nous concentrer sur les variables qualitatives (*cut, clarity, color*).

Notre analyse est donc centrée sur ces caractéristiques, après avoir tenu compte du carat.

Choix de représentation graphique

Choix des couleurs : La couleur bleue a été choisie pour les éléments visuels (barres, boxplots, courbes), car c'est celle que l'on associe spontanément à l'image d'un diamant.

Boxplots :
Les *boxplots* ont été utilisés pour comparer le prix selon des caractéristiques esthétiques telles que la coupe, la couleur et la clarté, à carat constant (~1). Ce type de graphique est particulièrement efficace pour visualiser la dispersion des prix et identifier les valeurs extrêmes.

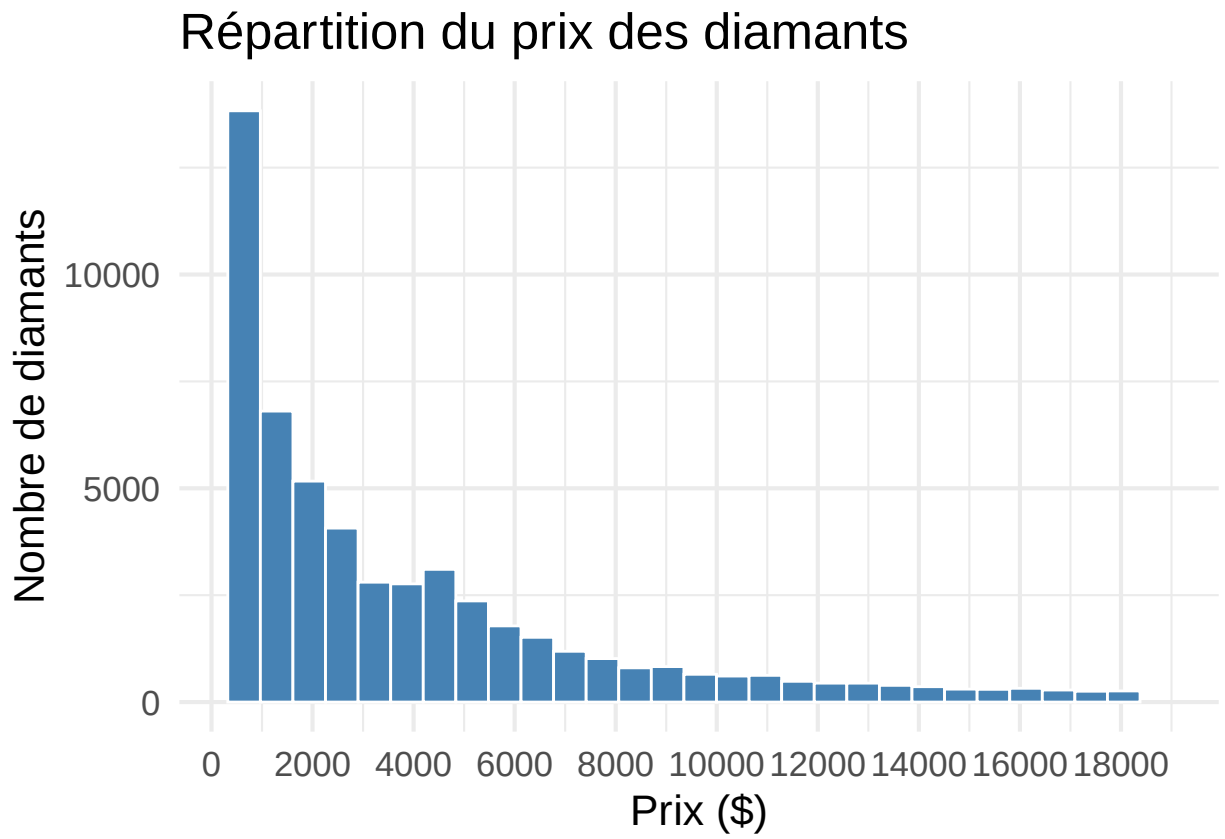
Histogramme : Nous avons utilisé un histogramme pour l'exploration initiale des prix. Il nous a permis d'identifier rapidement une forte concentration de diamants dans les tranches de prix les plus basses.

Courbe (*line plot*) : Nous avons utilisé un *line plot* pour étudier la relation entrele le carat et le prix moyen. Il nous a permit de visualiser l'évolution du prix moyen en fonction du poids du diamant, et de mettre en lumière la forte corrélation entre le prix et le carat.

Diagramme en barres (*geom_bar*) : Nous avons utilisé un *geom_bar* pour analyser la fréquence des carats dans notre jeu de données. Cela nous a permis de détecter un déséquilibre dans la distribution des observations en fonction du carat, ce qui nous a aidés à interpréter le *line plot* précédent.

Analyse en programmation lettrée

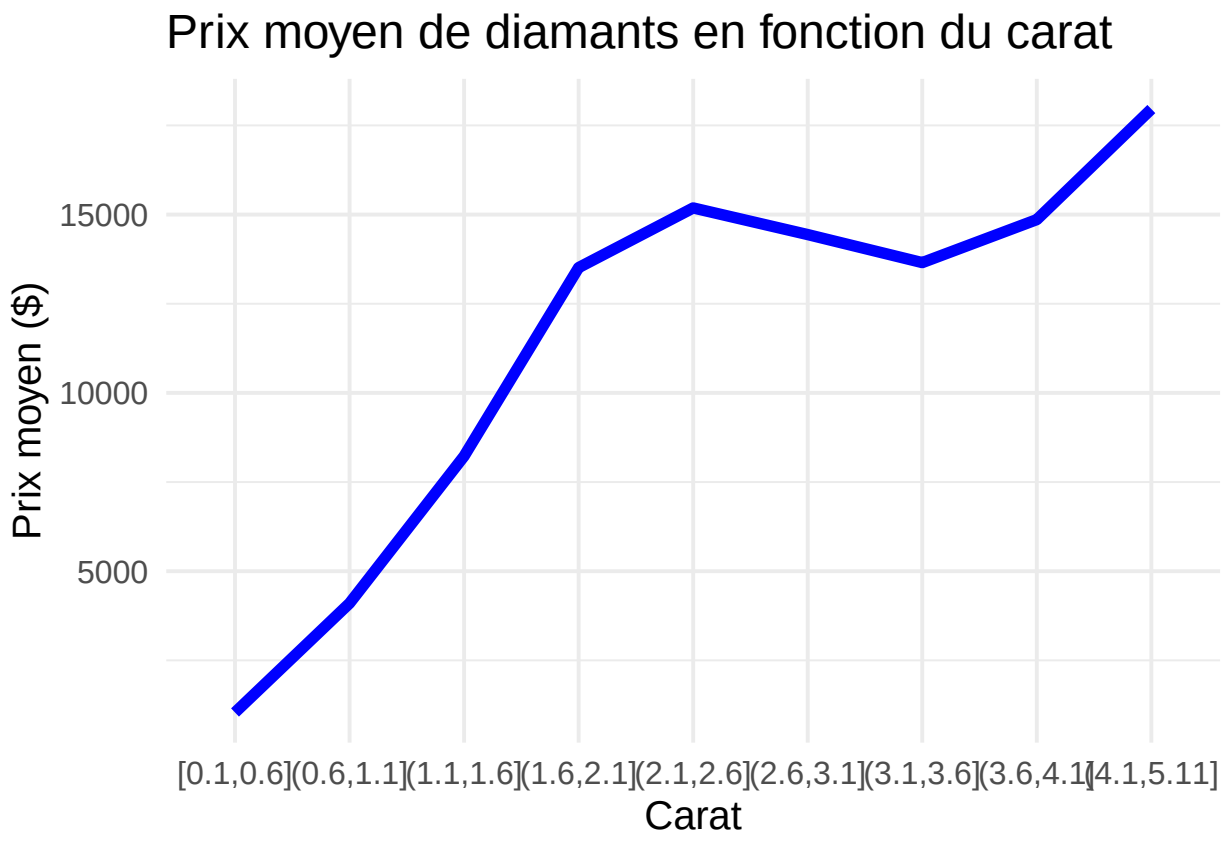
Tout d'abord, afin de nous familiariser avec notre sujet, nous avons analysé la répartition des diamants en fonction de leur prix.



Nous avons constaté que la majorité des diamants se situaient dans le premier tiers de notre échelle de prix, ce qui implique que disposerons de moins d'observations pour les tranches de prix plus élevées.

À la suite de ce constat, nous avons eu une réflexion : il existe deux grandes catégories de catactéristiques. D'une part, les **caractéristiques physiques** comme le poids, la longueur ou la largeur; d'autre part, les **caractéristiques esthétiques** comme la couleur, la coupe et la clarté du diamant.

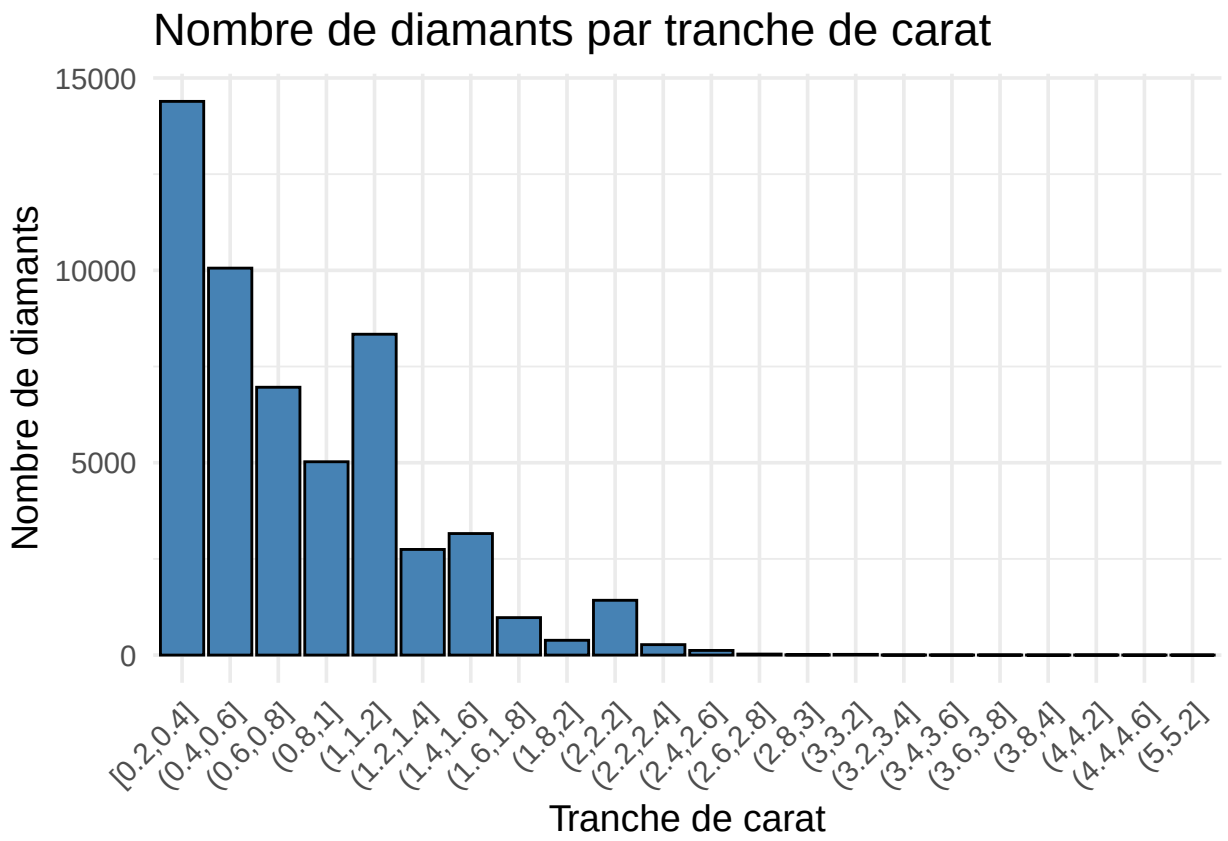
Nous avons donc commencé par analyser une **caractéristique physique** qui nous semblait pertinante, à savoir **le carat**, en lien avec le prix.



Nous constatons que le carat est fortement corellé au prix. La courbe obtenue montre que l'augmentation du prix est presque proportionnelle à celle du carat, à l'exception de la plage comprise entre 2.1 et 4.1 carats, où la progression semble stagner.

Le carat apparaît donc comme un facteur extrémement déterminant du prix, du moins parmis les caractéristiques physiques.

Cependant, pour mieux comprendre cette stagnation du prix entre 2.1 et 4.1 carats, nous avons réalisé un autre graphique, sous forme de *geom_bar*, afin d'observer la répartition des observations dans notre jeu de données en fonction du carat.



Ce graphique montre que notre jeu de données contient une majorité d'observations pour de faibles carats, et que l'intervalle entre 2.1 et 4.1 carats ne comporte que très peu d'observations.

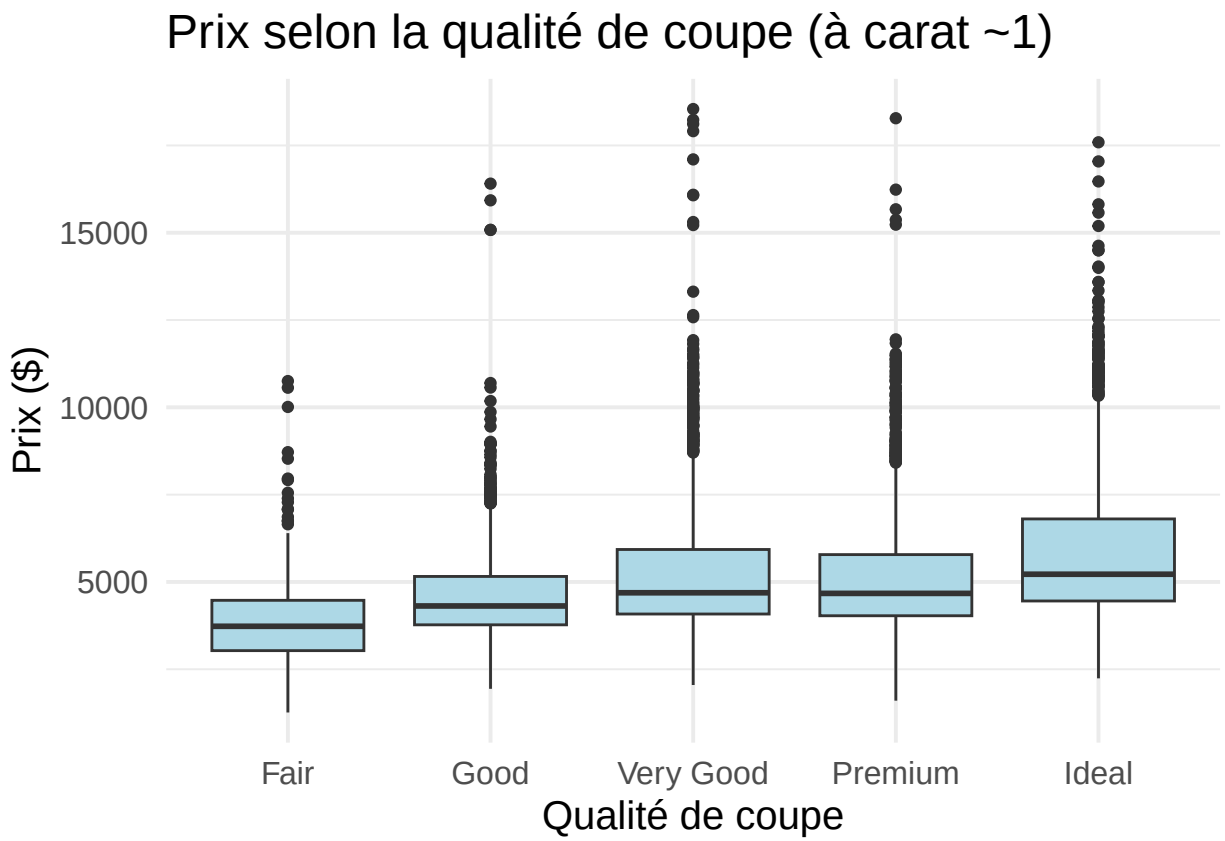
Ce manque de données justifie probablement le résultat inatendu observé sur le graphique précédent.

Nous en concluons que le carat est la caractéristique physique la plus déterminante du prix.

Pour la suite, nous allons chercher à identifier la caractéristique esthétique la plus influente sur le prix d'un diamant.

En raison du faible nombre d'observations pour les prix et carates, nous avons décidé de ne conserver que les diamants ayant un carat proche de 1, car cette tranche est très bien représentée dans notre jeu de données.

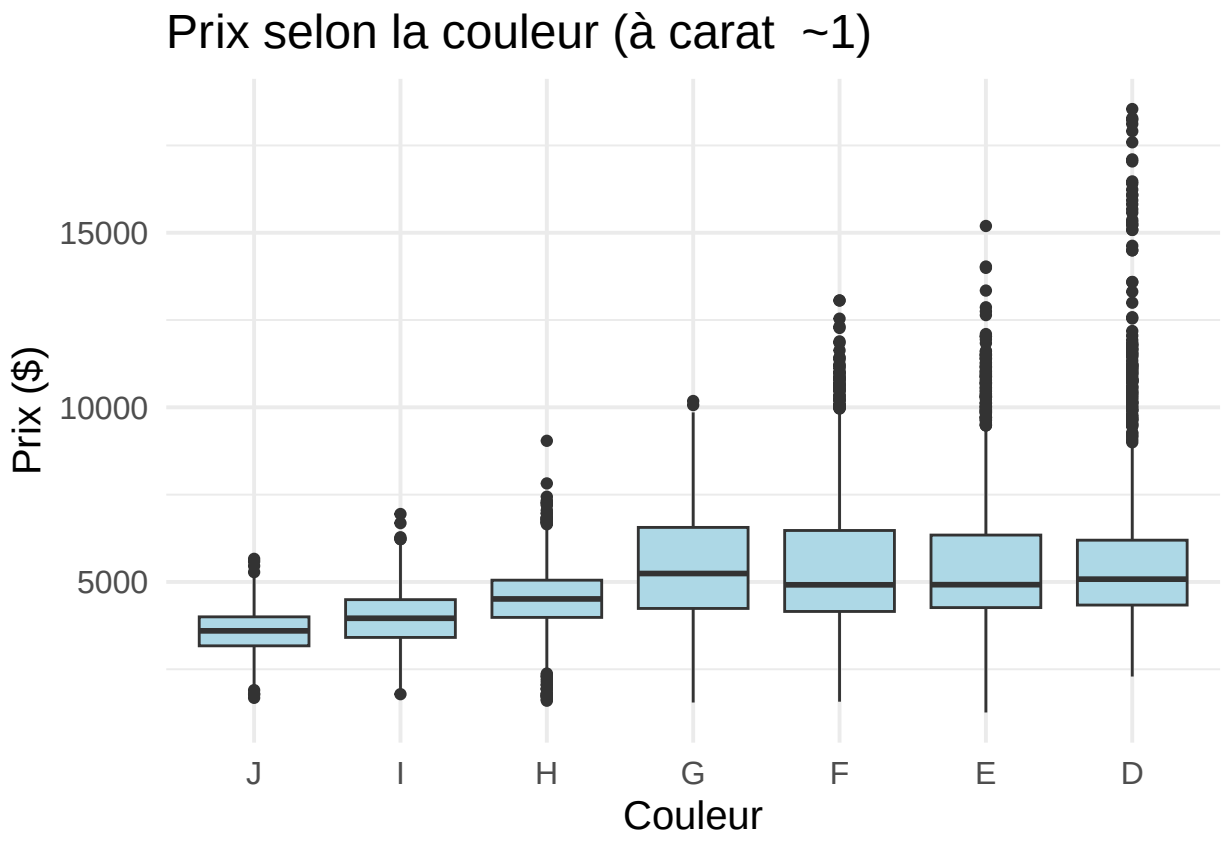
Nous allons donc analyser trois caractéristiques esthétiques en fonction du prix (à carat ~1), en commençant par la qualité de la coupe.



Ce boxplot montre que la qualité de la coupe n'a pas d'impact significatif sur le prix du diamant, bien que les diamants de qualité *"Fair"* soient globalement vendus à un prix légèrement inférieur aux autres.

La qualité de la coupe ne semble donc pas être un facteur déterminant du prix.

Nous avons ensuite analysé la couleur du diamant à l'aide d'un autre *boxplot*, construit de la même manière.

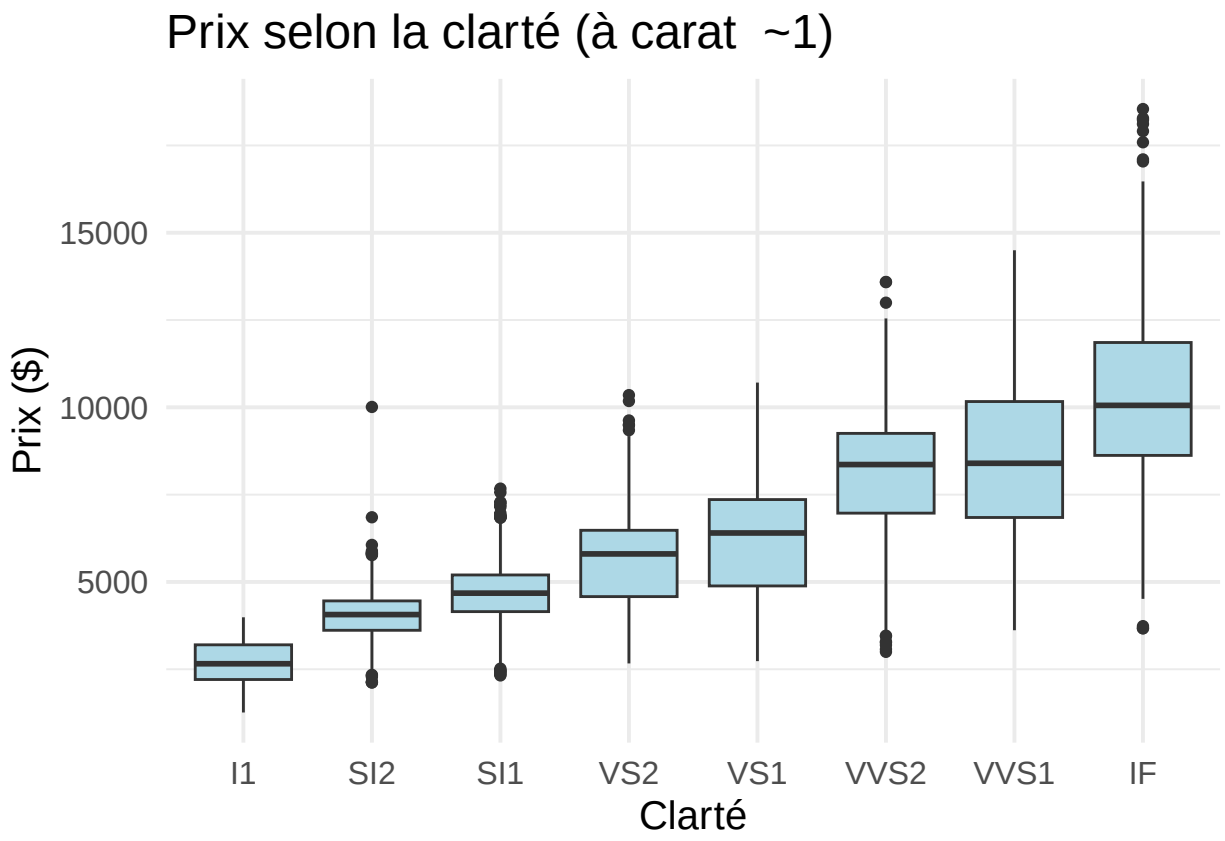


Ce boxplot montre que le prix moyen est légèrement influencé par la couleur.

Cependant, on observe très clairement au niveau des valeurs extrêmes que la couleur a un impact plus marqué sur le prix : plus la couleur est bonne, plus la valeur extrême du prix est élevée.

On peut ainsi en conclure que, dans l'optique d'obtenir les diamants **les plus cher**, la couleur D est la plus favorable.

Enfin, nous avons analysé la dernière caractéristique esthétique, à savoir la clarté, de la même manière que les autres.



Ici, on observe que la clarté influence très clairement le prix du diamant de manière croissante : plus la clarté est élevée, plus le prix augmente.

En ce qui concerne les valeurs extrêmes, on constate qu'il est indispensable d'avoir une clarté optimale pour atteindre les prix les plus élevés.

Conclusion

En conclusion, le diamant est un matériau précieux dont la valeur peut atteindre des sommets impressionnants, comme le montre le jeu de données, où le prix d'un diamant commence à plus de 300 dollars.

Ce prix est déterminer par une combinaison de caractéristiques physiques et esthétiques.

Parmi les différents facteurs influençant la valeur d'un diamant, le carat, c'est-à-dire le poids du diamant, s'est révélé être le plus déterminant.

Nos analyses ont mis en évidence une corrélation claire entre le carat et le prix : plus un diamant est lourd, plus il tend à coûter cher.

Cependant, cette relation ne suffit à elle seule à expliquer les écarts de prix, notamment pour les diamants les plus cher.

En effet, un diamant de grande taille n'est pas nécessairement le plus onéreux. Par exemple, dans notre jeu de données, le diamant le plus cher pèse 2.29 carats, tandis que le plus lourd atteint 5.1.

Aucun des **dix diamants les plus lourds** ne figure parmi les **dix diamants les plus cher**, ce qui met en lumière l'importance d'autres critères de qualité, comme la clarté et la couleur.

Ces deux caractéristiques, lorsqu'elles atteignent leur niveau de qualité maximal, semblent indispensables pour qu'un diamant soit considéré comme "ultra haut de gamme".

Ainsi, bien que le carat soit le facteur principal dans la détermination du prix, il ne sufit pas à lui seul : il doit être accompagné d'une excellente clarté et d'une couleur optimale pour atteindre les prix les plus élevés.

Références

https://www.kaggle.com/datasets/shivam2503/diamonds