الجـــمهوريـــة الجزائريـــة الديمـــقراطيـــــة الشـعبيـة
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵜⴰⵣⵣⴰⵢⵔⵉⵜ ⵜⴰⵎⴰⵙⵙⴰⵡⵉⵜ ⵜⴰⵖⵔⴼⴰⵏⵉⵜ
République Algérienne Démocratique et Populaire
وزارة التعليـــم العـــالي والبـــحث العلمـــي
ⵓⵢⵏⴰⵡ ⵏ ⵜⵓⵙⵙⵏⵉⵜ ⵜⵉⵏⵣⵡⴰⵏⵉⵜ ⴰ ⵜⵓⴰⵔⵣ ⵜⵓⵙⵙⵏⵉⵜ
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

ECOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE

المدرسة الوطنية العليا للإعلام الآلي
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵜⴰⵖⵔⴼⴰⵏⵉⵜ ⵏ ⵜⵓⵙⵙⵏⵉⵜ ⵏ ⵜⵓⵙⵙⵏⵉⵜ
École nationale Supérieure d'Informatique

Rapport du TP ANAD

2 ème année Cycle Supérieur (2CS)

Option : Système Informatique (SQ)

Thème :

Analyse Statistique et Multidimensionnelle des Facteurs Déterminants du Prix des Diamants

Réalisé par :

- Benameur Tarek
- Touil Nihel

Introduction:

Dans ce TP, nous analysons un jeu de données contenant des informations sur les prix de près de 54 000 diamants ronds. Ce dataset comporte des variables décrivant les caractéristiques physiques des diamants, telles que leur poids, leurs proportions et leurs dimensions, ainsi que des informations qualitatives liées à leur coupe, couleur et clarté. L'objectif est d'étudier la relation entre ces caractéristiques et le prix des diamants, tout en utilisant l'Analyse en Composantes Principales (PCA) pour réduire la dimensionnalité et mieux comprendre les tendances globales dans les données. voici la description des variables:

- **carat** : Poids du diamant en carats .
- **cut** : Qualité de la coupe du diamant, classée en cinq catégories (Fair, Good, Very Good, Premium, Ideal).
- **color** : Couleur du diamant, évaluée sur une échelle allant de "D" (meilleure qualité) à "J" (qualité inférieure).
- **clarity** : Pureté du diamant, mesurée selon une échelle allant de "I1" (la moins pure) à "IF" (la plus pure).
- **depth** : Proportion de la profondeur totale du diamant, exprimée en pourcentage
- **table** : Largeur relative de la table (surface supérieure plate du diamant) .
- **price** : Prix du diamant en dollars américains.
- **x, y, z** : Dimensions du diamant (longueur, largeur, profondeur).

Tableau de données

carat	cut	color	clarity	depth	table	price	x	y	z
<dbl>	<ord>	<ord>	<ord>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

Figure 1 :Aperçu de la table de données

Distribution des prix des diamants

Nous remarquons que la distribution des prix des diamants suit une tendance décroissante à mesure que le prix augmente. En effet, la plus grande occurrence se situe autour de 1000 USD, ce qui représente la plage de prix la plus fréquente. Cela indique une concentration des prix dans une fourchette modérée. En revanche, au-delà de 12500 USD, les occurrences deviennent beaucoup plus rares, suggérant que les diamants plus chers sont moins fréquents et peuvent représenter des catégories plus exclusives ou des diamants de grande qualité.

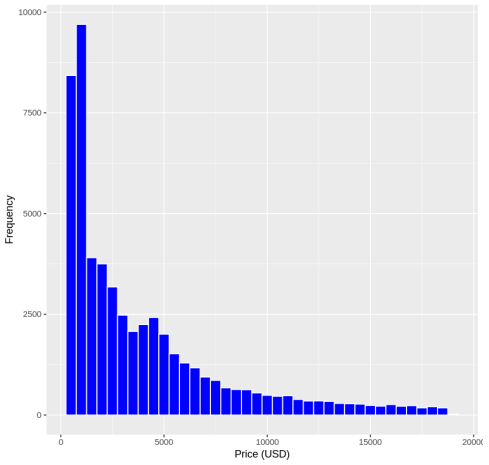

Figure 2 : Distribution des prix

Corrélations et relations entre caractéristiques physiques et prix des diamants :

Cette figure met en évidence une forte corrélation positive entre **le poids en carats** des diamants et leurs dimensions physiques **(x, y, z)**. Par ailleurs, une forte corrélation est également observée entre **le poids en carats et le prix des diamants**. Ces observations suggèrent que les dimensions physiques et le poids sont des facteurs majeurs influençant le prix. Afin de mieux comprendre et visualiser les relations complexes entre ces variables numériques fortement corrélées, une analyse en composantes principales (PCA) peut être utilisée pour réduire la dimensionnalité.

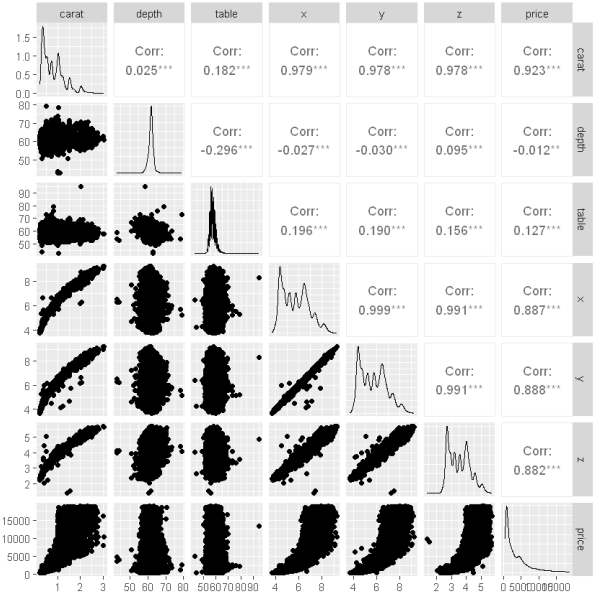

Figure 3 : Matrice de corrélation enrichie avec distributions et relations croisées

Relation entre les variables catégoriques et le prix :

À première vue, nous constatons la présence de nombreuses valeurs aberrantes. Cela s'explique par le fait que certains diamants affichent des prix nettement supérieurs à la majorité des autres, probablement en raison de caractéristiques particulières. Cependant, certaines valeurs aberrantes sont clairement dues à des erreurs, comme des dimensions nulles (par exemple, (x,y,z)=0). Dans ces cas, nous choisissons de supprimer les entrées correspondantes pour garantir la qualité des données.

Les boxplots montrent un chevauchement significatif entre certaines catégories des variables catégoriques (par exemple, *clarity*, *color* et *cut*). Ce phénomène peut s'expliquer par une distribution similaire des prix dans ces catégories, ce qui suggère que les différences de prix ne sont pas significativement marquées entre certaines d'entre elles. Cette superposition des boîtes met en évidence une faible variabilité du prix en fonction de ces catégories spécifiques.

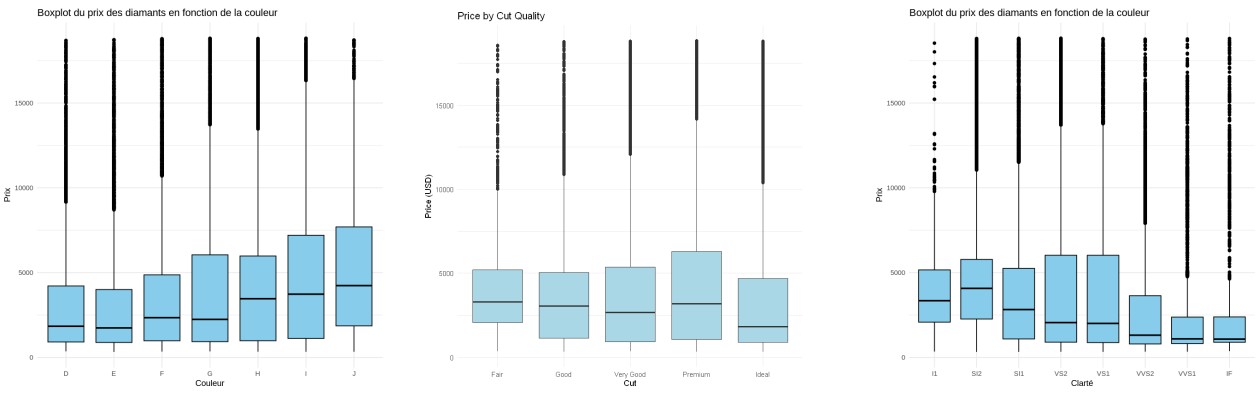

Figure 4 : boxplots des prix de diamants en fonction de couleur, qualité de coupe et clarté

Nous avons réalisé trois graphiques de dispersion en croisant le **carat** avec les variables qualitatives *cut*, *clarity* et *color*. Lorsqu'elles étaient analysées de manière isolée, ces variables qualitatives ne permettaient pas d'obtenir des résultats cohérents. Cependant, en intégrant le **carat**, les relations entre ces variables et le **prix** des diamants sont devenues plus intelligibles. Cela démontre que le **prix** d'un diamant est influencé par plusieurs caractéristiques, et que l'ajout du **carat** améliore la compréhension de ces interactions.

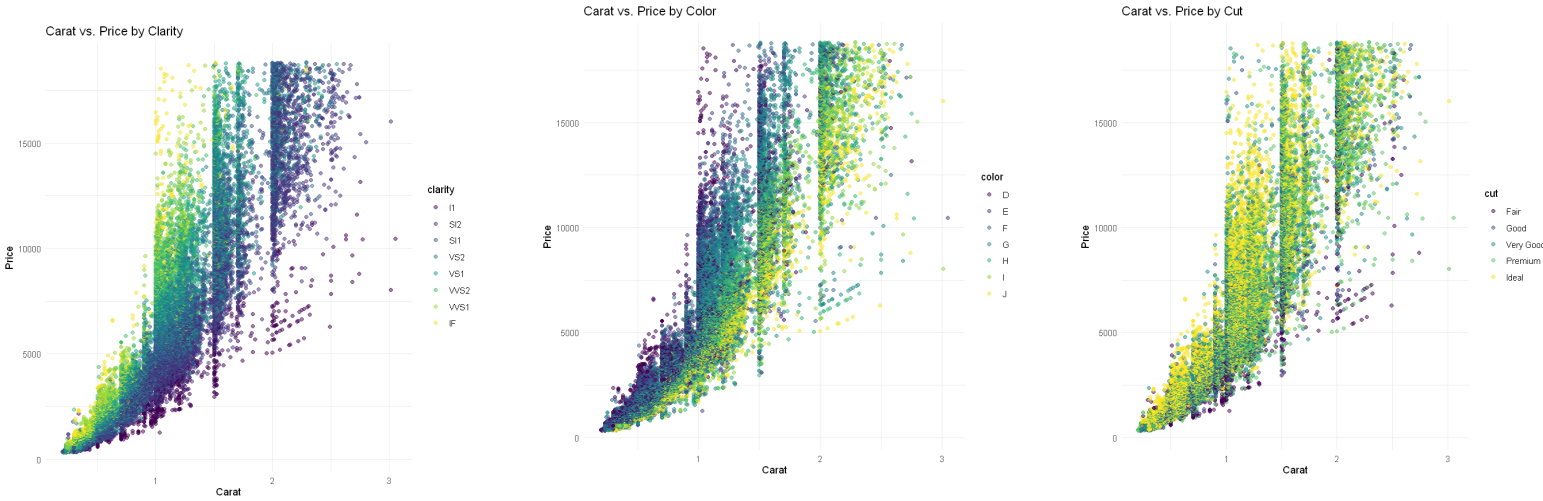

Figure 5 : Graphiques de dispersion des prix selon la clarté, couleur et qualité de coupe

Analyse des valeurs numériques :

Passons maintenant à l'examen des variables numériques. Nous avons choisi d'effectuer une Analyse en Composantes Principales (ACP) afin de mieux comprendre les relations entre les différentes variables et de réduire la dimensionnalité des données. Comme mentionné précédemment, il existe une forte corrélation entre certaines variables, ce qui rend l'ACP particulièrement adaptée pour résumer l'information sans perdre de pertinence.

Il est également essentiel de normaliser les données avant de procéder à l'ACP, car les variables sont exprimées sur des échelles différentes. Cette normalisation garantit que toutes les variables contribuent de manière équitable à l'analyse, évitant qu'une variable ayant une amplitude élevée n'influence de manière disproportionnée les résultats.

Les résultats de l'analyse en Composantes Principales (ACP) montrent qu'en utilisant uniquement les deux premières composantes principales, nous expliquons déjà 86 % de la variance totale des données. Cela indique que la majorité de l'information peut être capturée avec seulement deux axes, permettant une réduction de la dimensionnalité tout en conservant une proportion significative de la variabilité des données. Cette réduction simplifie l'analyse sans perte d'informations importantes.

Eigenvalues							
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
Variance	4.839	1.285	0.691	0.162	0.020	0.001	0.001
% of var.	69.127	18.360	9.874	2.320	0.282	0.019	0.018
Cumulative % of var.	69.127	87.488	97.362	99.682	99.963	99.982	100.000

Figure 6 : table des valeurs propres

Projection des variables :

Le cercle des corrélations présente la projection des variables sur les deux premières

composantes principales. Ce graphique met en évidence comment chaque variable est positionnée selon les axes principaux de la variance. Par exemple, on observe que les variables *carat*, *price*, *x*, *y* et *z* sont fortement corrélées et proches l'une de l'autre sur l'axe de la première composante principale (69.1%), indiquant qu'elles expliquent une proportion importante de la variance des données. En revanche, *table* et *depth*, bien que moins corrélées entre elles et éloignées de cet axe, contribuent davantage à la deuxième composante principale (18,4 %). De plus, ces deux variables présentent une corrélation négative, signalant une relation inverse dans leur influence sur les composantes principales.

Projection des individus :

Le graphique de dispersion illustre la projection des individus (ici, chaque diamant) selon ces deux premières composantes principales. Chaque point représente un diamant, coloré en fonction de la qualité de la coupe (*cut*). Cette projection nous permet de visualiser comment les différentes catégories de coupe se répartissent dans l'espace des deux premières composantes principales. On observe une certaine séparation entre les catégories de coupe, avec une concentration plus dense des diamants *Ideal* et *Premium* dans la zone centrale de la projection, tandis que les diamants avec des coupes de qualité inférieure (*Fair*, *Good*) semblent plus dispersés.
Par exemple, les diamants de qualité ***Ideal*** se situent au centre par rapport à l'axe 2, influencé par les variables ***depth*** et ***table***, qui déterminent la qualité de la ***coupe***. En revanche, les variables ***carat*** et ***price*** montrent une plus grande variabilité, traduisant une large gamme de prix et de poids, même pour cette qualité.

Ces visualisations nous aident à mieux comprendre la structure des données et à identifier des patterns intéressants, notamment en ce qui concerne la relation entre la qualité de la coupe et d'autres variables comme le prix et le carat.

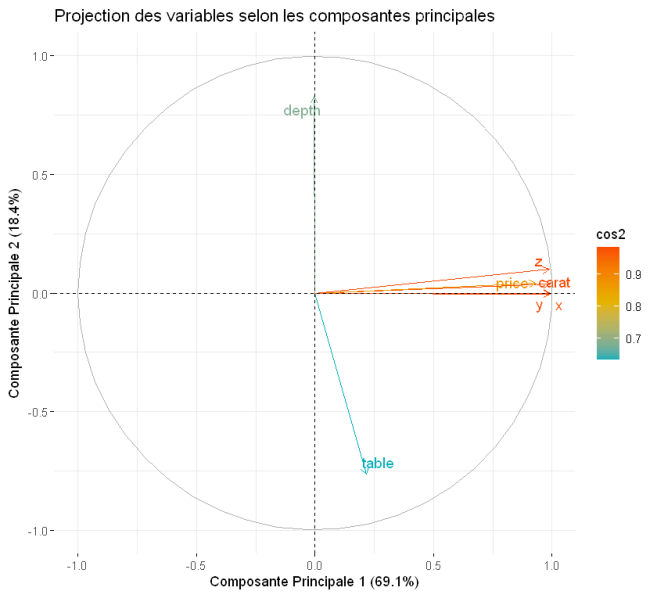

Figure 6 : Cercle de corrélations de l'ACP

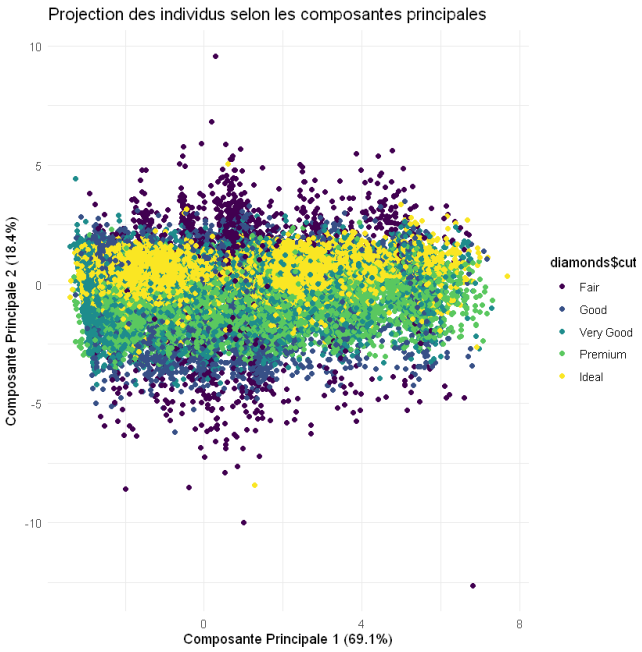

Figure 7 : Représentation des individus selon la qualité de coupe

Conclusion :

Grâce à une combinaison d'analyses statistiques, de visualisations graphiques et de réduction de dimensionnalité via l'Analyse en Composantes Principales (ACP), nous avons mis en évidence *les relations clés entre les caractéristiques des diamants et leur prix*.

L'ACP a permis de **réduire la dimensionnalité des données** tout en conservant l'essentiel de la variance, *révélant une forte corrélation entre certaines variables*.

Cependant, l'analyse révèle également une influence importante des dynamiques du marché. Les diamants de haute qualité (comme "**Ideal**") et de poids élevé en carats sont rares, car leur production naturelle est limitée. Cette rareté crée une demande forte et contribue à la dispersion des prix observée, particulièrement pour les diamants avec des caractéristiques exceptionnelles. Ainsi, les qualités inférieures ("**Fair**" ou "**Good**") sont plus dispersées car elles couvrent une gamme plus large en termes de disponibilité et de variabilité des caractéristiques.

En conclusion, cette analyse a permis d'**identifier les principales caractéristiques influençant le prix des diamants, essentile poids en carat, la qualité de la coupe, la couleur, et la clarté ainsi que l'impact de la rareté et de la demande sur le marché.** Ces facteurs combinés expliquent la variabilité des prix et des positions des diamants sur les différentes composantes principales.