



Projet de Python Avancé

Classification des prix des bijoux féminins du site Histoire d'Or

Fancello Marie Clara, Germini Eva, Gutfreund Eloise – Décembre 2021

Présentation de l'analyse



CADRE D'ANALYSE

Cette analyse porte sur les articles féminins du site Histoire d'Or.

Les modèles développés utilisent PySpark ainsi que la librairie Scikit-learn de Python.



OBJECTIF

Notre objectif est de prédire la catégorie de prix des articles.

Par extension, nous développons deux modèles de classification supervisée et comparons les résultats.

Plan

- 1.** Présentation du site web d'Histoire d'Or
- 2.** Web Scraping : récupération des données
- 3.** Premiers traitements et première description de la base de données
- 4.** Analyses statistiques et transformations de variables
- 5.** Modélisation et résultats

Présentation du site Histoire d'Or

Histoire d'Or, vendeur de bijoux (bagues, bracelets...). est installé en France. Il est possible d'acheter les produits en magasin ou sur son site: (histoiredor.com)

Collier Clothilde Or Blanc Topaze Et Oxyde De Zirconium

Collier Or Blanc 375/1000 Maille Forçat Coeur Topaze Bleue 1ct Et Oxydes De Zirconium 42cm

Référence: B3CFBTB505C



DÉTAIL PRODUIT

Genre	Femme
Poids total (gr)	1.02
Matière principale	Or
Couleur matière	Blanc
Titrage matière ?	375/1000
Longueur (cm)	42
Type de chaîne	Forcat
Type de motif	Cœur

PIERRE PRINCIPALE

Type de pierre	Topaze
Couleur	Bleu sky
Nombre de pierres	1
Forme	Coeur
Caratage (ct) ?	1.0000
Type de sertis ?	Grains
Traitement	Irradiation

Chaque article est accompagné d'une photo, de son prix ainsi que de sa description.

Web Scraping

Nous effectuons le Web Scraping dans un Jupyter Notebook. Les articles présents sur le site d'Histoire d'Or possèdent trois indicateurs intéressants pour notre étude:

- le titre de l'article
- le prix de l'article
- la description de l'article.

Nous récupérons les informations sur les articles grâce aux balises du code HTML du site.

		article	prix	description
2779		\nBague Anja Or Blanc Rubis Et Diamant\n	360.00	{Ge...
2780		\nBague Solitaire Victoria Or Blanc Diamant\n	590.00	{Ge...
2781		\nCréoles Telya Lisses Fil Carré Or Jaune\n	99.00	{Ge...
2782		\nBague Leni Or Blanc Diamant\n	319.00	{Ge...
2783		\nAlliance Valentine Or Blanc Diamant Synthet...	790.00	{Ge...

Premiers traitements et première description de la base de données

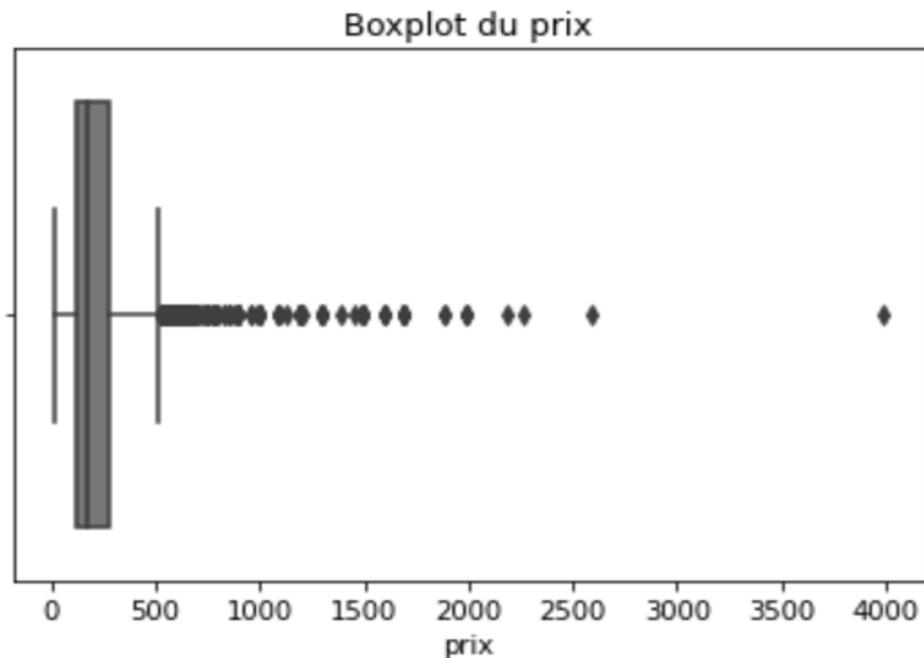
Les premiers traitements de la base de données se font avec Pyspark sur Databricks grâce à l'utilisation de regex.

- Création de la variable "catégorie" à partir de la variable article donnant la catégorie de l'article.
- Création de nouvelles variables comme : poids de l'article, matière de l'article, couleur de la matière, type de pierres sur le bijou, couleur des pierres, forme des pierres, nombre de pierres, type de motifs sur le bijou, la longueur de l'article et largeur de l'article

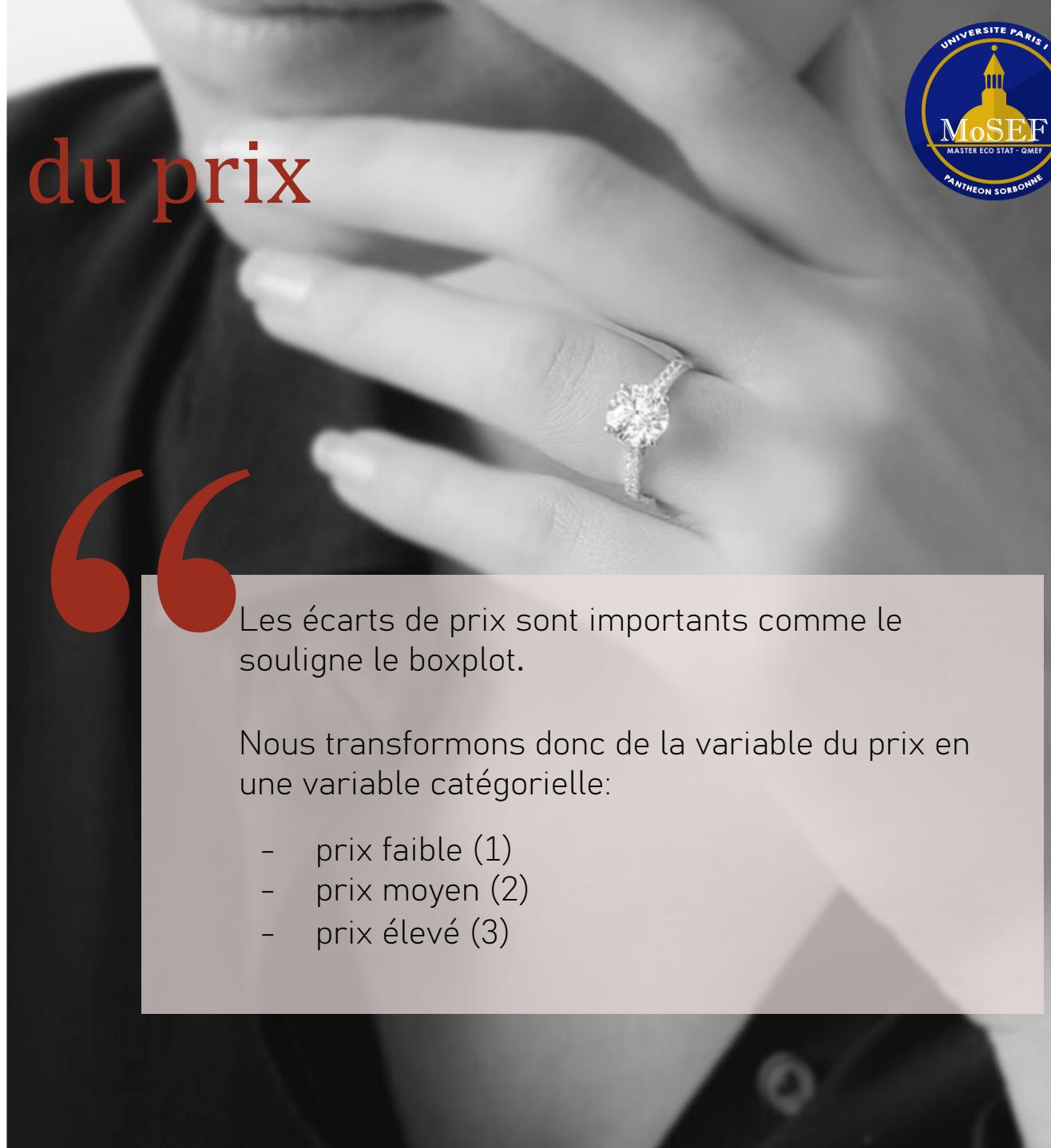
prix	categorie	poids	matiere	couleur_mat	type_pierre	couleur_pierre	forme	type_motif	nb_pierre	longueur	largueur
159.00	Bague	0.9	Or	Blanc	Oxyde	Blanc	Ronde	Cur	2		
169.00	Bague	0.94	Or	Blanc	Oxyde	Blanc	Ronde		12		
89.00	Boucles	0.70	Or	Jaune	Strass	Blanc	globaleRonde				

Taille de la base de données: 2367 articles et 12 variables

Analyse descriptive du prix

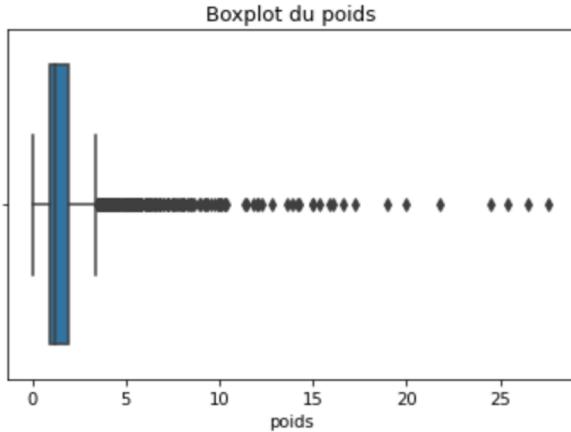


- En moyenne le prix des articles est de 259 euros.
- Le prix de l'article le moins chère est de 15€
- L'article le plus chère coûte 3990€

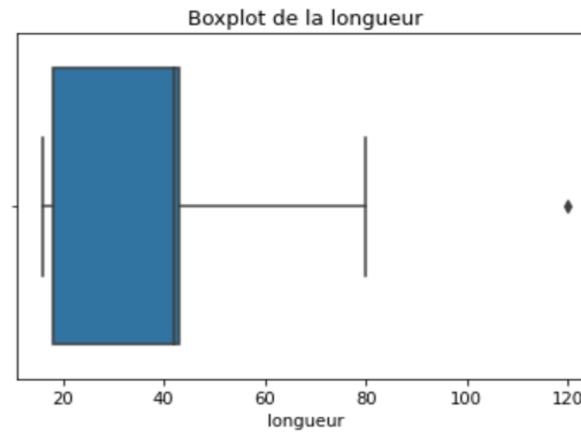


Après observation des distributions, nous transformons les variables numériques suivantes :

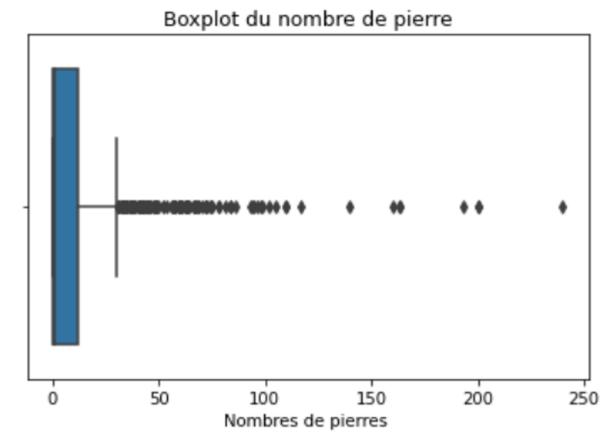
Le poids :



La longueur :



Le nombre de pierres :



- Les poids faibles (1)
- Les poids moyens (2)
- Les poids élevés (3)

- Les longueurs courtes (1)
- Les longueurs longue (2)
- Pas de longueur (3)

- Pas de pierre (1)
- Nombre moyen de pierres (2)
- Nombre élevé de pierres (3)

Après observation des distributions, nous transformons les variables catégorielles suivantes :

Matière:

- Or (1)
- Autre (2) représentant les autres matières
- Non connues (3) représentant les valeurs manquantes (33%)

Couleur:

- Jaune (1) étant le plus représenté
- Blanc (2)
- Autres (3) dans laquelle nous plaçons les NaN car ils sont au nombre de 3

Forme:

- Ronde (1)
- Autre (2) représentant les autres formes
- Non connues (3) représentant les valeurs manquantes

Couleur de la pierre:

- Blanc (1)
- Autre (2) représentant les autres couleurs
- Non connues (3) représentant les valeurs manquantes

	matiere_new	count
3		639
1		1114
2		150
	couleur	count
3		264
1		846
2		793
	categorie	count
Boucles		146
Collier		421
Autres		206
Bracelet		299
Bague		831
	forme_new	count
3		639
1		1058
2		206
	coul_pierre_new	count
3		665
1		989
2		249

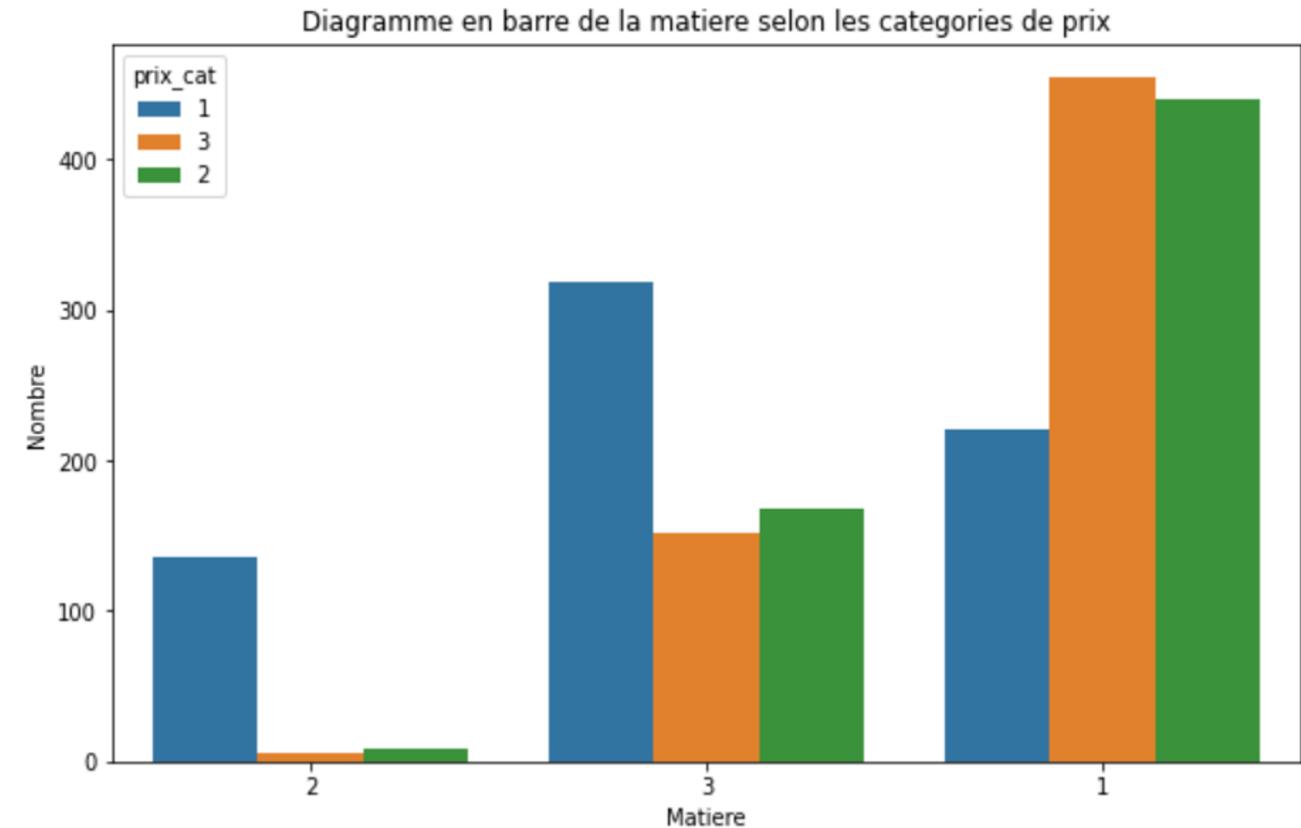
Nous observons également la distribution des matières par rapport au prix :

“

Lorsque le bijou est en or (1) son prix fait majoritairement parti des catégories avec un prix élevé (orange) ou un prix moyen (vert)

Pour les produits composés d'autres métaux (2), le prix est faible (bleu).

Lorsque la matière est non renseignée, le résultat est davantage contrasté.



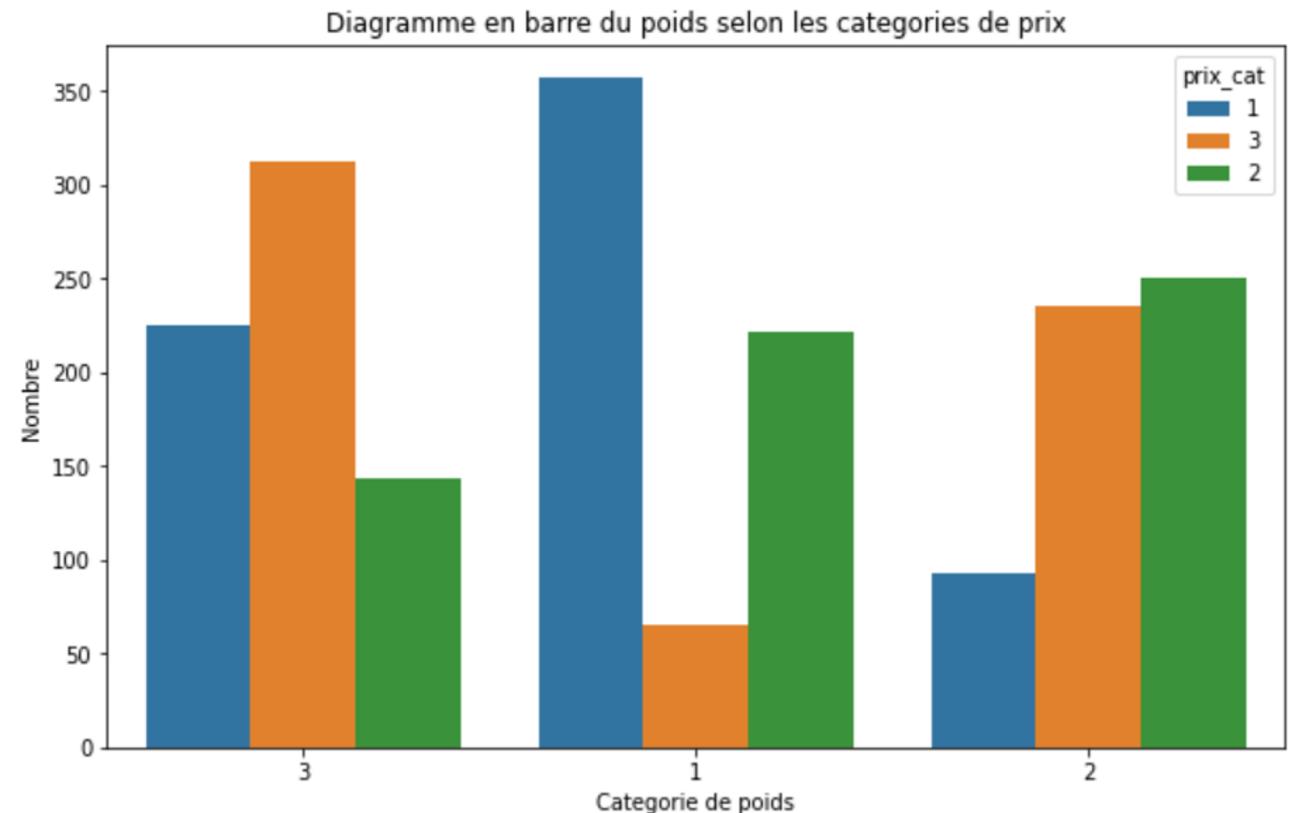
Nous observons également la distribution du poids par rapport au prix :

“

Pour la catégorie avec un poids élevé (3), son prix fait majoritairement parti de la catégorie avec un prix élevé (orange).

A l'inverse, lorsque le bijou est léger (1), son prix fait majoritairement parti de la catégorie avec un prix faible (bleu).

Pour la catégorie avec un poids moyen (2), le bilan est davantage contrasté.



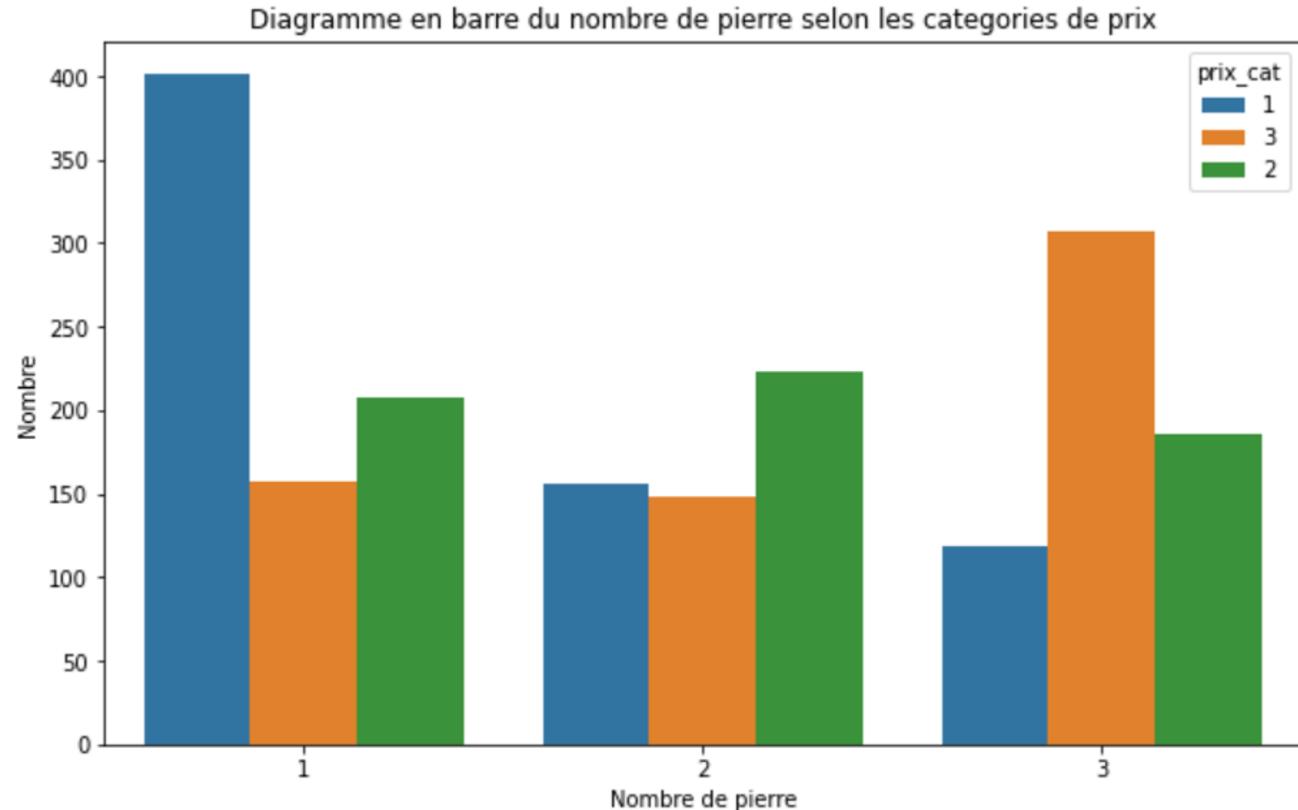
Nous observons également la distribution du nombre de pierres par rapport au prix :

“

Pour la catégorie avec un nombre de pierres élevé (3), son prix fait majoritairement parti de la catégorie avec un prix élevé (orange).

A l'inverse, lorsque le bijou n'a pas de pierres (1) son prix fait majoritairement parti de la catégorie avec un prix faible (bleu).

Lorsque le bijou a un nombre moyen de pierre (2) son prix fait majoritairement parti de la catégorie avec un prix faible (bleu).



Présentation de la méthodologie

Séparation de la base de données :

- 1903 observations pour entraîner le modèle
- 464 observations pour tester le modèle

One-Hot-Encoding: Transformation des variables catégorielles en variables binaires

Cross validation: Séparation de notre jeu de données train en 5 groupes

A la première itération, le jeu de données est entraîné sur 4 groupes et validé sur un groupe. A la deuxième itération, le jeu de données est entraîné sur 4 autres groupes et validé sur un autre groupe ...

GridSearchCV: Tester plusieurs modèles avec des hyperparamètres différents afin d'obtenir le minimum d'erreurs de prédictions par la pénalisation

Modèles utilisés:

- Modèle de classification Logistique
- Modèle de classification Random Forest

Accuracy: Métrique mesurant la probabilité que le modèle prédit la bonne classe

- Une accuracy de 100% signifie que le modèle prédit parfaitement les classes
- Une accuracy de 50% signifie que le modèle n'est pas meilleur que le hasard

Modélisation – Résultats

“

- Pyspark:

- L'accuracy du modèle logit de Pyspark est de 63%. Cela signifie que le modèle ne prédit pas la bonne classe de prix dans 37% des cas.
- L'accuracy du modèle random de Pyspark est de 70%. Cela signifie que le modèle ne prédit pas la bonne classe de prix dans 30% des cas.

- Sklearn:

- L'accuracy du modèle logit de Pyspark est de 63%. Cela signifie que le modèle ne prédit pas la bonne classe de prix dans 37% des cas.
- L'accuracy du modèle random de Pyspark est de 70%. Cela signifie que le modèle ne prédit pas la bonne classe de prix dans 30% des cas.

- **Comparaison:** Le modèle de classification logit a une meilleure performance avec la librairie sklearn. Cependant, la random forest a une capacité prédictive inférieure à celle effectuée sur pyspark.

	Pyspark	Sklearn de Python
Logit	63%	66%
Random Forest	70%	67%

Conclusion

“

Nous avons web-scape le site Histoire d'Or afin d'obtenir les caractéristiques de chaque article féminin.

Ensuite, nous avons créé puis nettoyé la base de données et procédé à des analyses descriptives. Par extension, nous avons proposé deux modèles de Machine Learning afin de prédire les catégories de prix des produits. Les modèles Logit et Random Forest avaient respectivement une accuracy de 63% et 70%.

Nous avons également exécuté ces mêmes modèles avec la librairie Sklearn de Python. Nous constatons que le modèle de classification logit a une meilleure performance avec la librairie Sklearn (accuracy de 66%). Cependant, la Random Forest a une capacité prédictive inférieure à celle effectuée sur pyspark (67% VS 70%).

En conclusion, le modèle qui permet de prédire la catégorie de prix des articles avec une erreur minimale est celui de la Random Forest de Pyspark.

