

**Rapport de Projet :**

SYSTEME DE RECOMMANDATION D’ORDINATEURS

**Matière : Analyse de données**

**Membres du groupe :**

Mouhamadou Diouf Cissé

Ngoné Ndiaye

Cheikh Ahmadou Bamba Syll

**Introduction :**

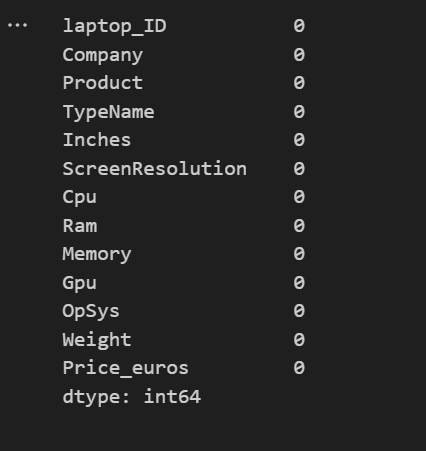
De nos jours, avec la révolution du numérique, les ordinateurs sont devenus omniprésents dans notre vie quotidienne. Que ce soit au travail, à la maison ou en déplacement, ils sont devenus des outils indispensables de notre quotidien, et ce dans de nombreux domaines. Lorsqu'arrive le moment de choisir un nouvel ordinateur, il est donc essentiel de bien réfléchir à ses besoins et à l'utilisation qu'on souhaite en faire avant de se décider pour trouver l'ordinateur qui convient à notre besoin, et qui est plus rentable en rapport qualité-prix. Ainsi, ce projet porte sur un système de recommandation d'ordinateurs à partir de spécifications fournies par l'utilisateur. Nous avons récupéré une base de données d'ordinateurs(laptop.csv) sur Kaggle sur laquelle nous avons appliqué un prétraitement puis une ACM. Et le clustering a permis à la fin de nous retrouver avec trois classes d’ordinateur (bas de gamme, milieu de gamme et haut de gamme).

Nous avons développé une interface pour le système, où l'utilisateur est invité à choisir les spécifications désirées (RAM, ROM, Marque, OS etc.) et le système lui suggère à la suite des ordinateurs répondant à ces critères.

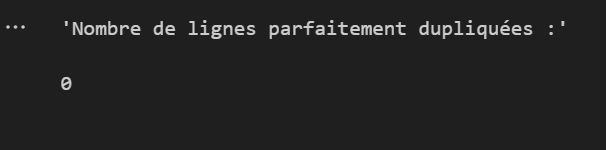
1. **Description et prétraitement des données :**

Les données doivent être décrites et prétraitées avant utilisation. Cela va permettre d’enlever les données qui ne sont pas pertinentes et pourraient fausser le modèle.

Notre base de données contient initialement 1303 ordinateurs portables avec diverses caractéristiques techniques dont le processeur, la mémoire RAM, le stockage ROM etc.

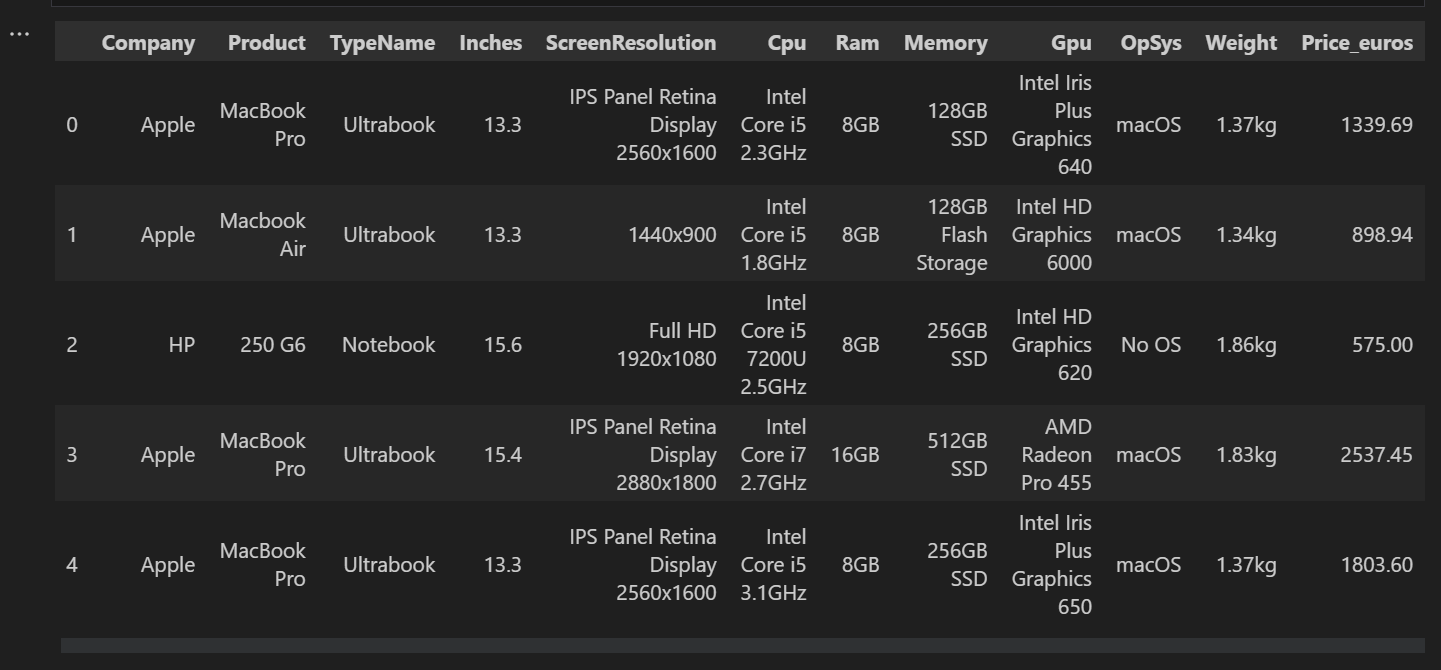


Dataframe des données manquantes :

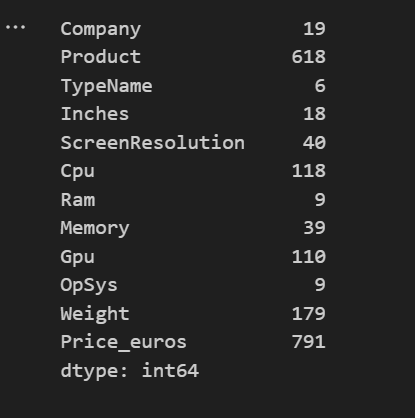


Duplicatas

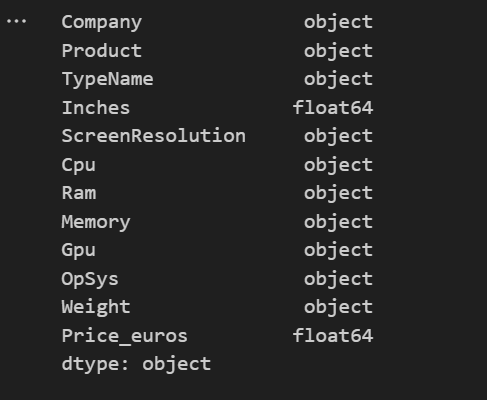
On remarque ainsi qu’il n’y a pas de données dupliquées ni de données manquantes. Et nous allons par la suite supprimer la colonne des ID qui correspondent aux numéros des lignes, elle est donc redondante.



Nouveau DataFrame

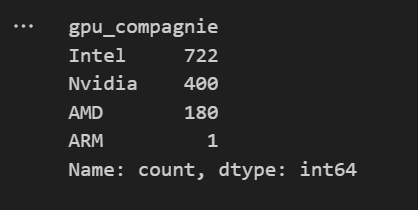


Valeurs uniques

Nous pouvons ainsi visualiser le nombre de valeurs uniques pour chaque colonne.

Types des variables

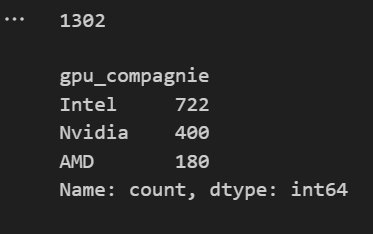
Les variables quantitatives sont celles avec le type float64, les autres sont des qualitatives plus précisément des variables catégorielles.



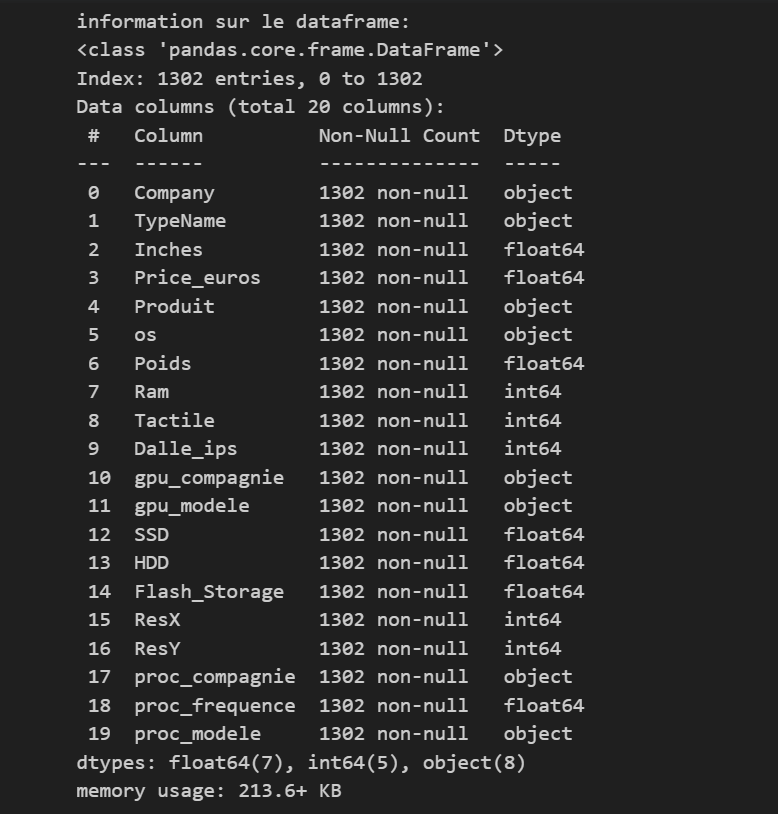
Types de GPU

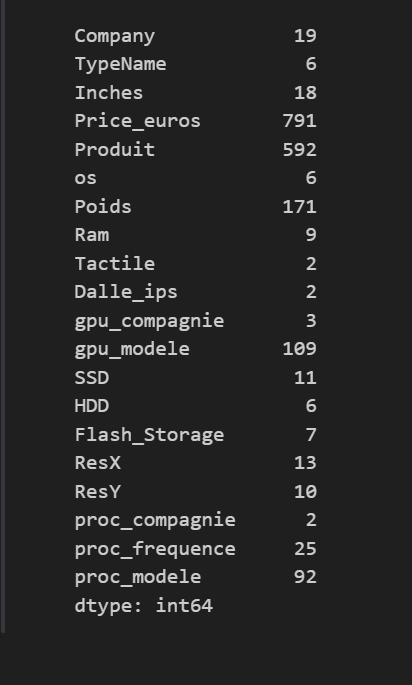
Nous remarquons ici que le type de GPU ARM n’est représenté qu’une seule fois parmi 1303 ordinateurs donc ce n’est pas pertinent de le prendre en compte. Il est donc considéré comme un outlier et sera supprimé pour la suite.

On obtient ainsi :



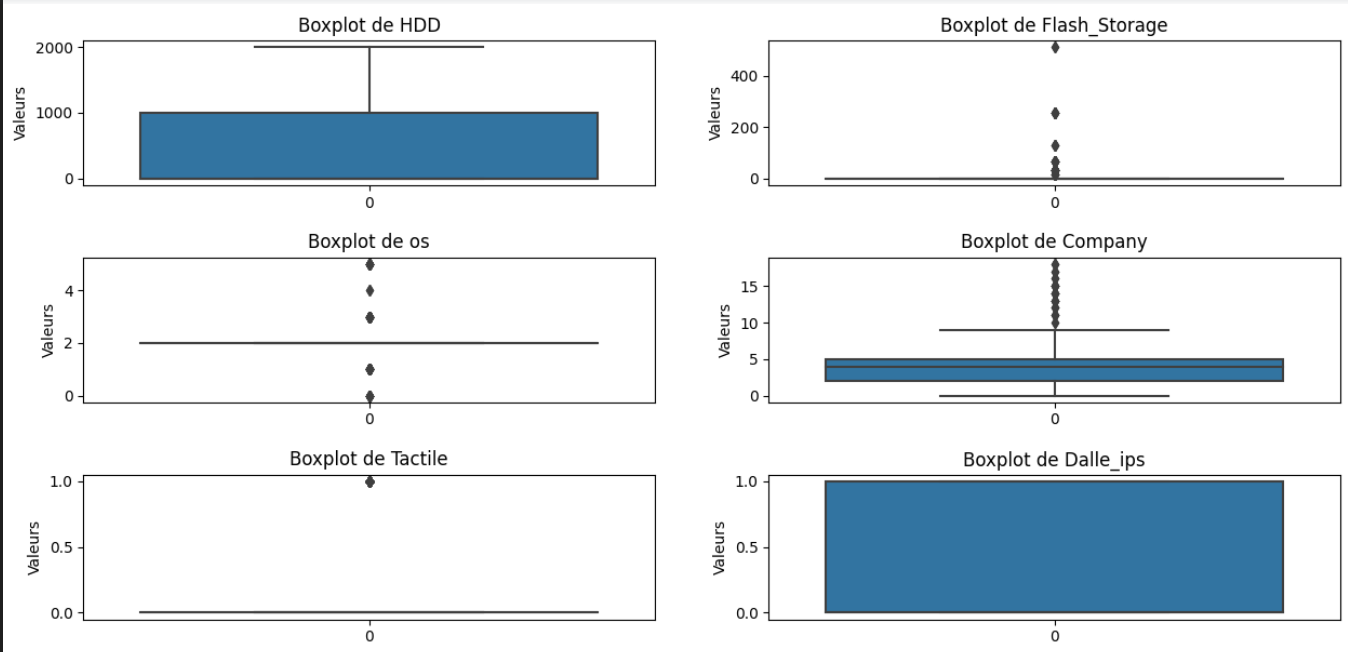
Ce nouveau tableau contient des colonnes dissociées. Ces nouvelles colonnes donnent des informations pouvant constituer des critères pour le programme.

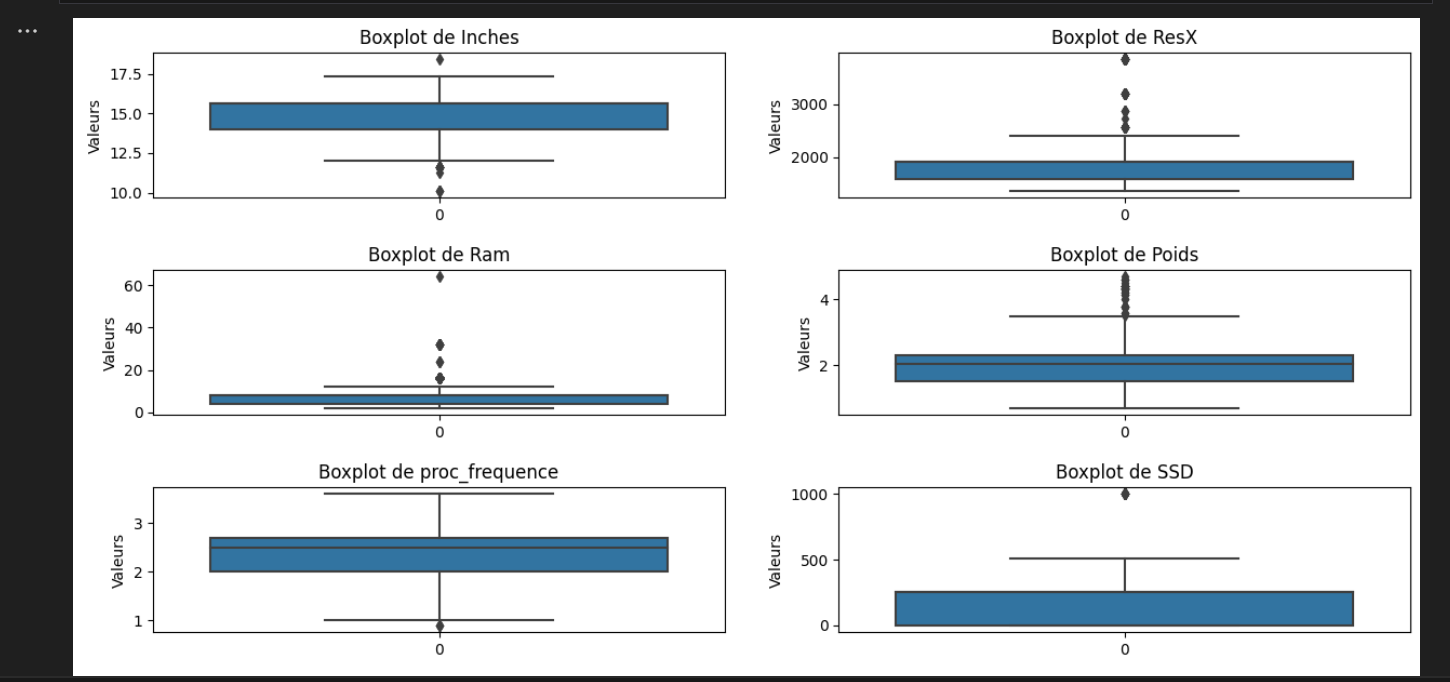
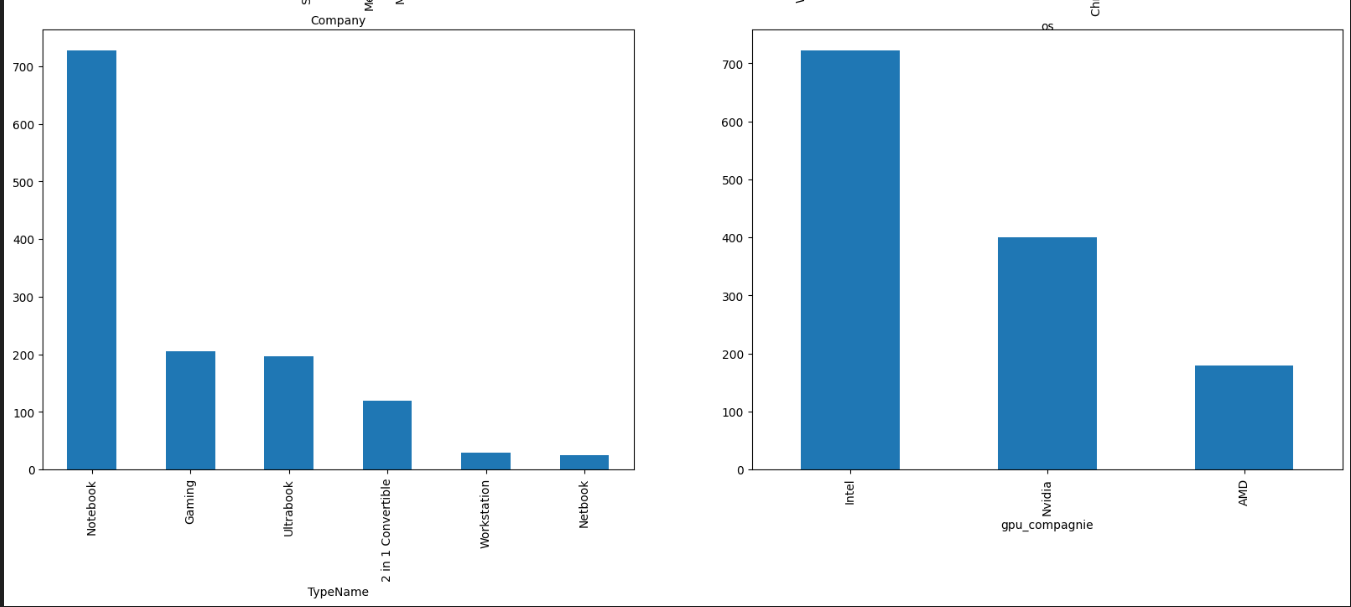
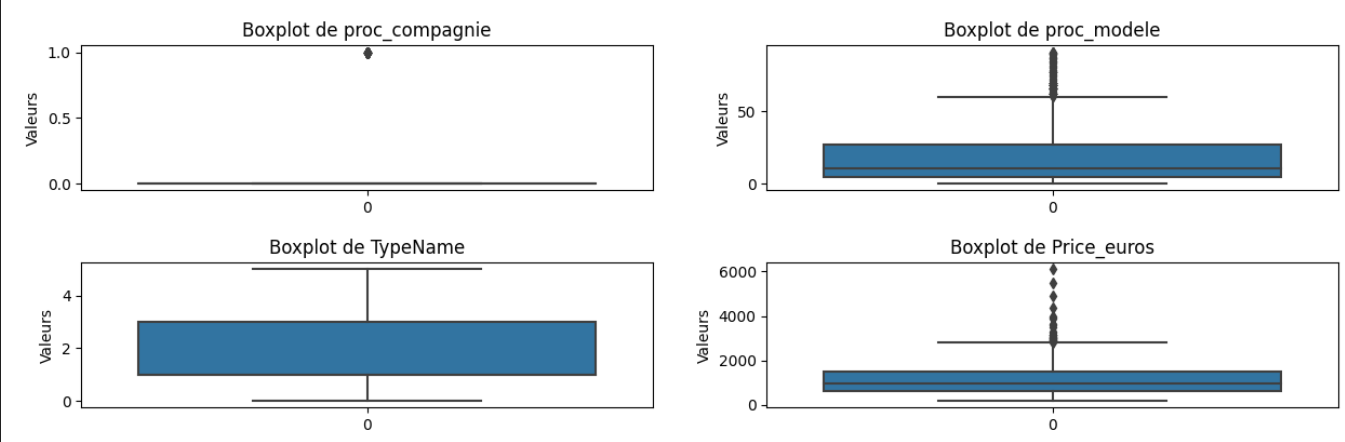


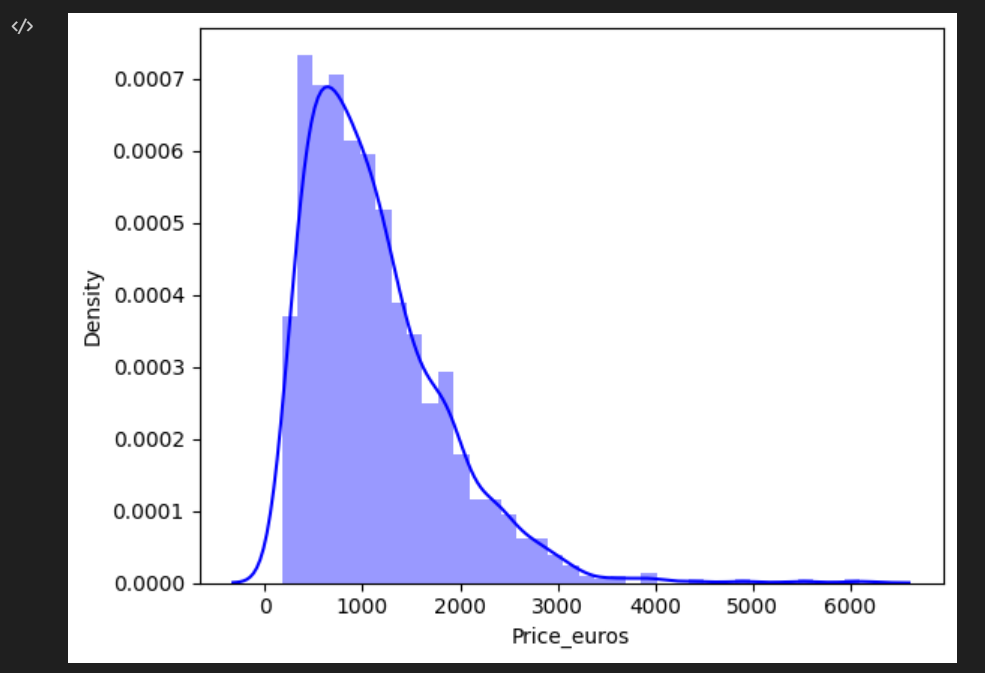
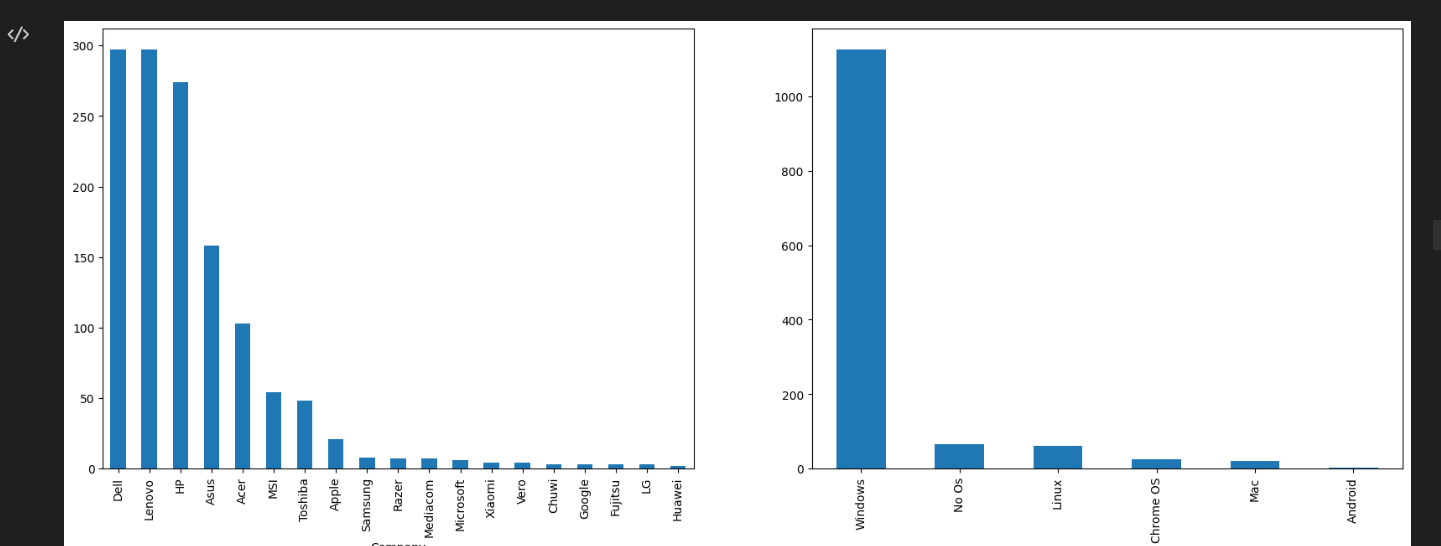


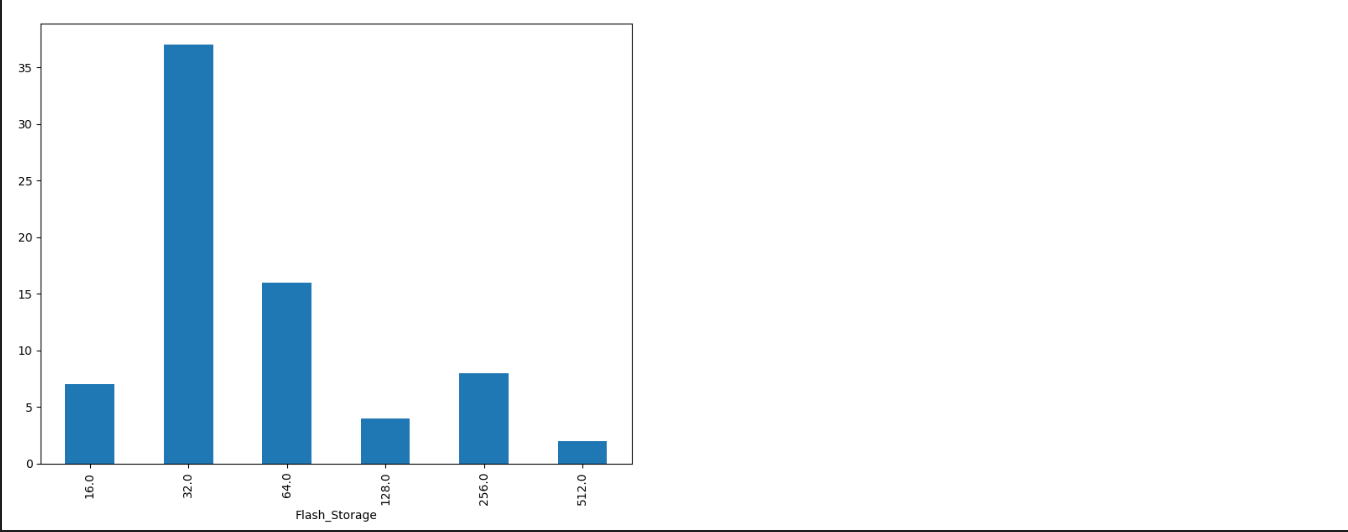
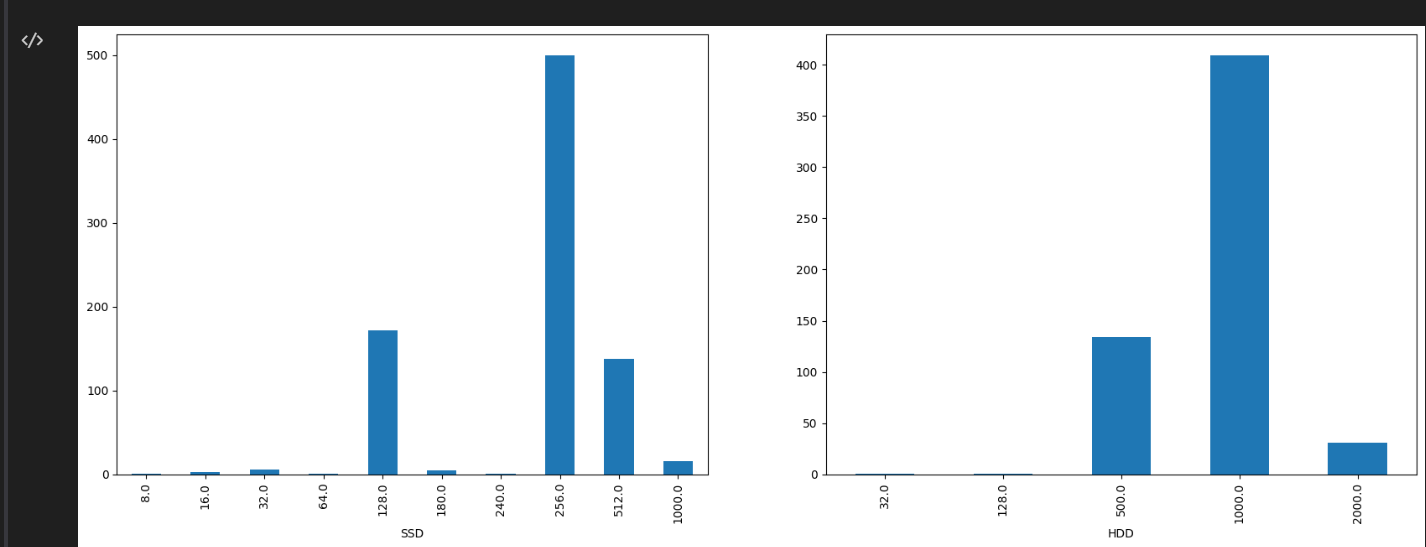
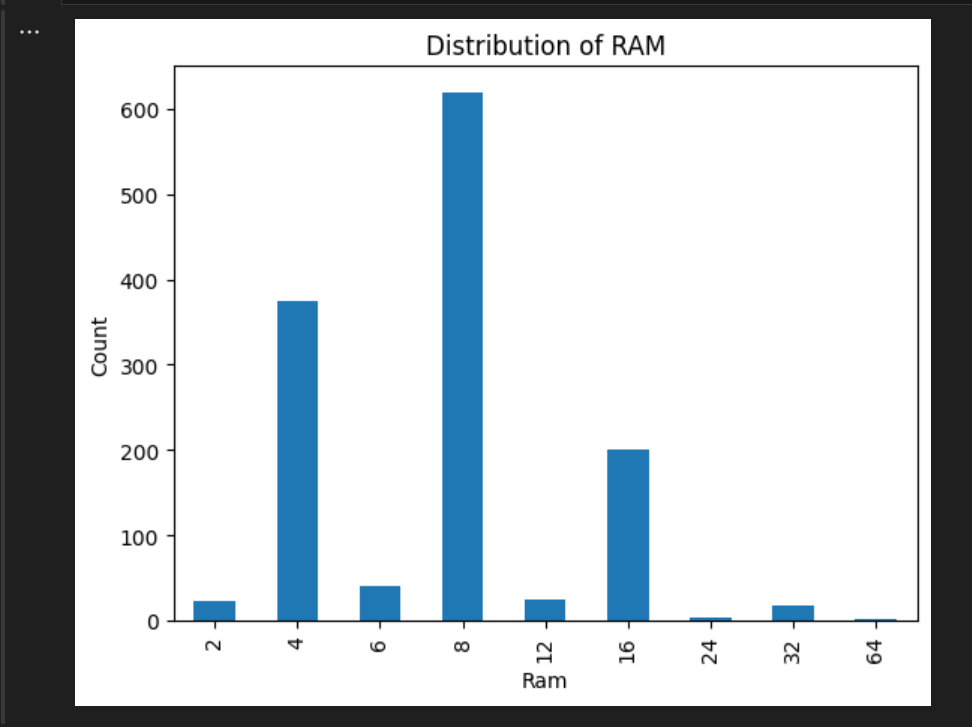
Des informations sur la nouvelle Dataframe

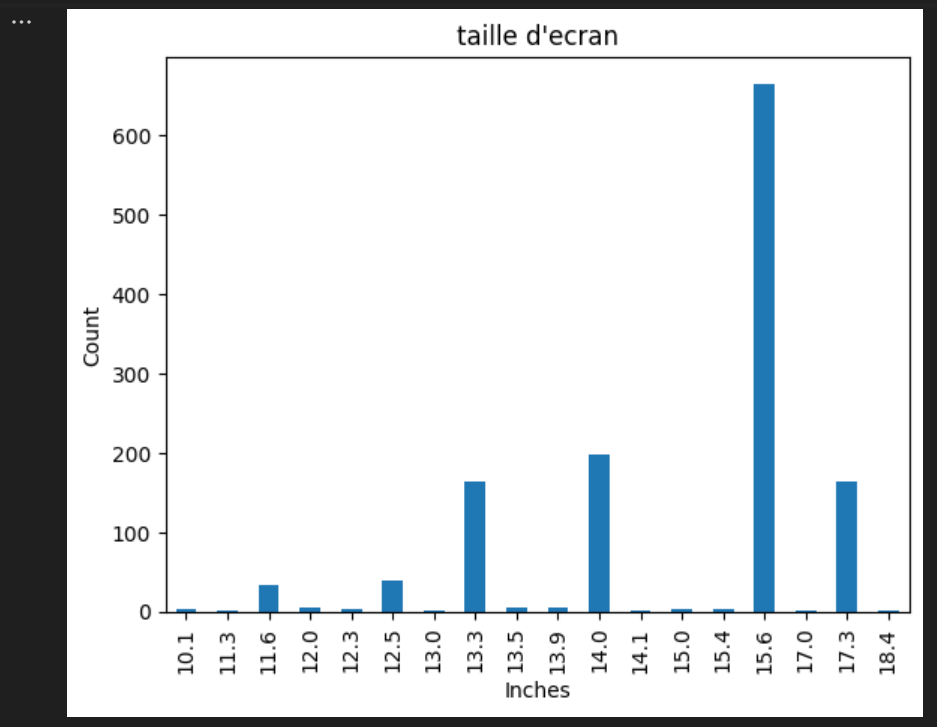
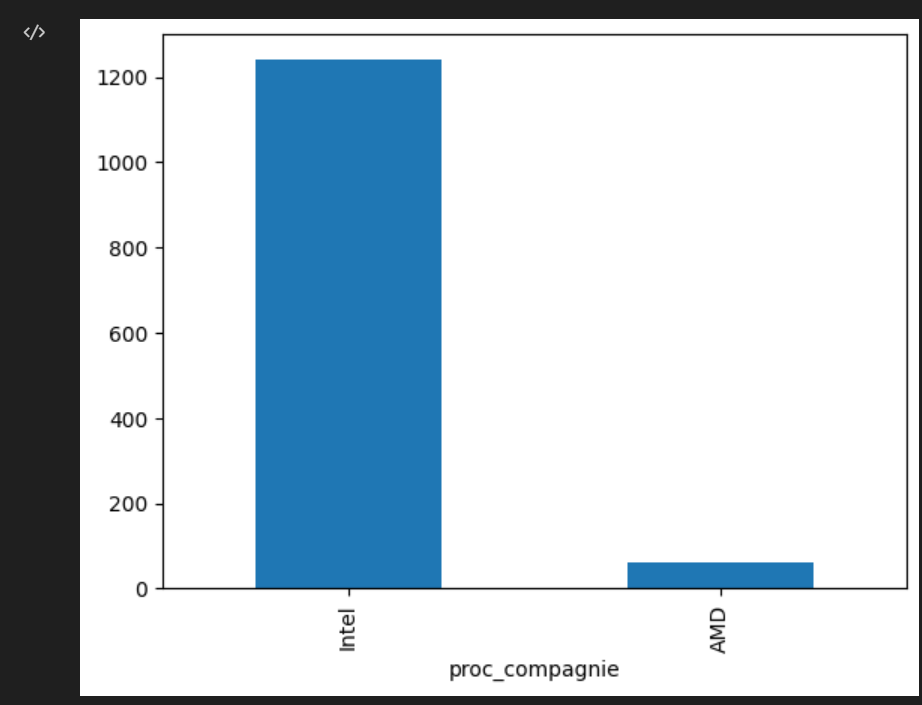
On voit les nouvelles variables, leurs types et leurs nombres de valeurs uniques. On voit également que chaque colonne est représentée pour les 1302 ordinateurs.

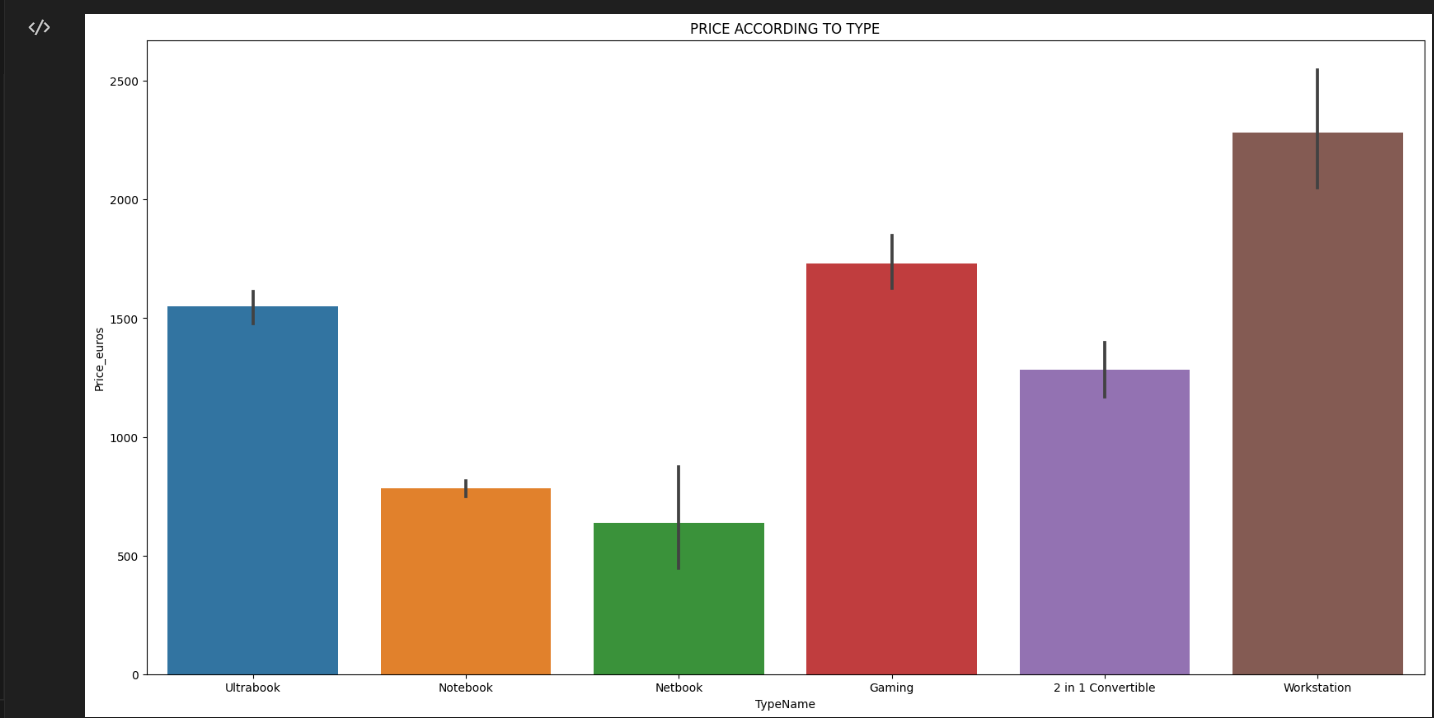
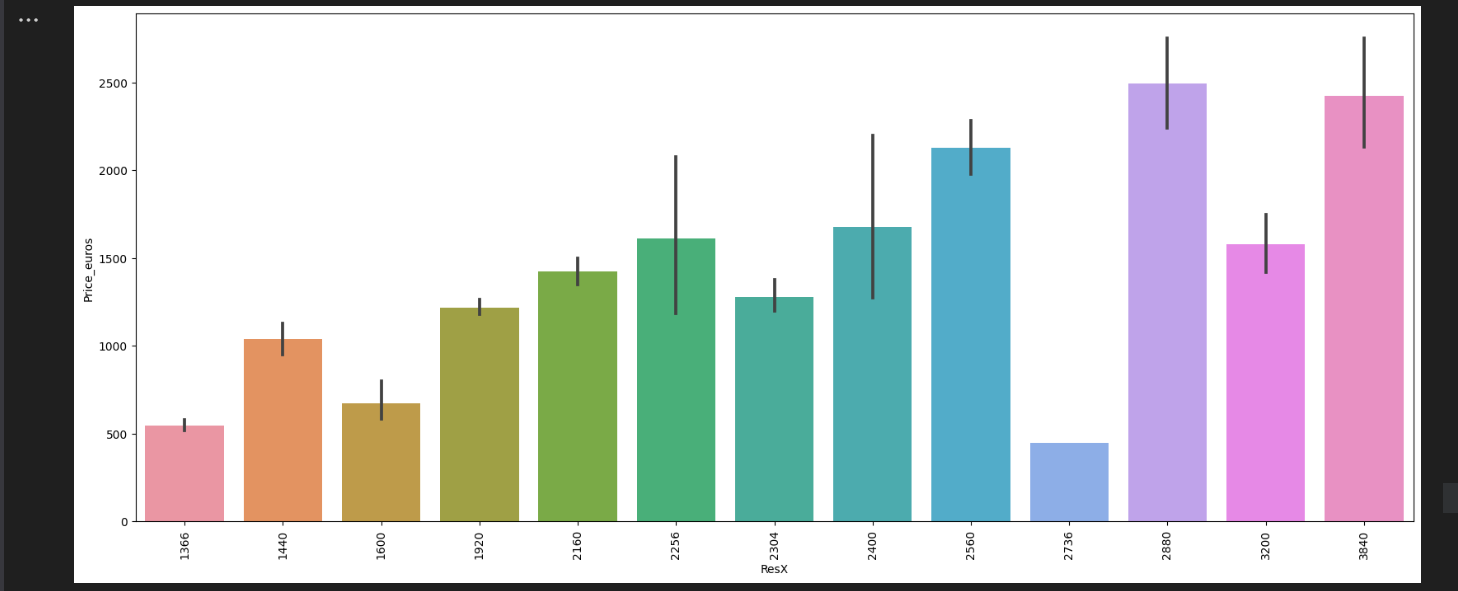
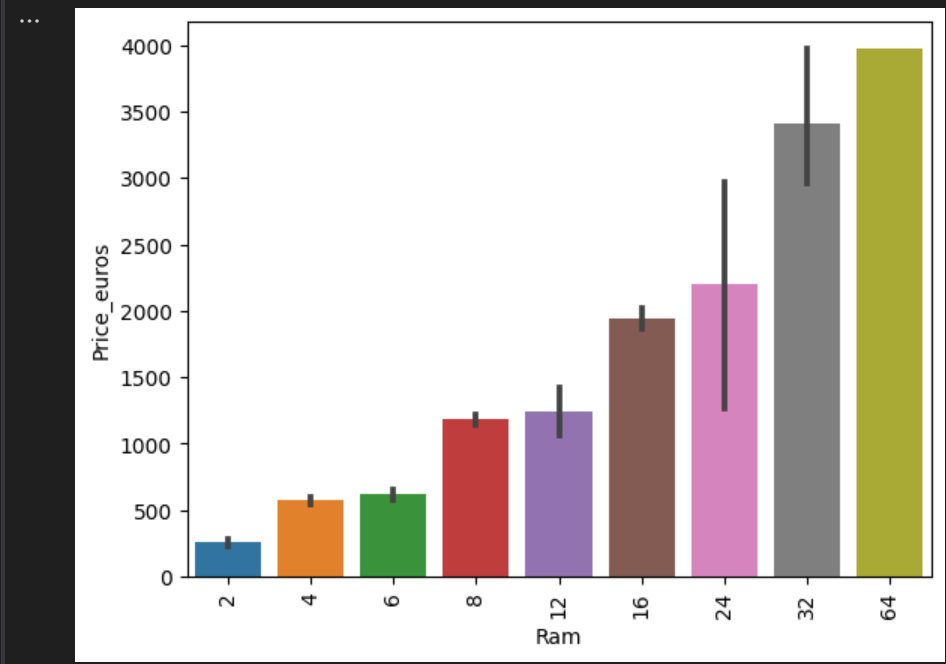
Voici les graphiques traduisant les tendances pour les différentes variables

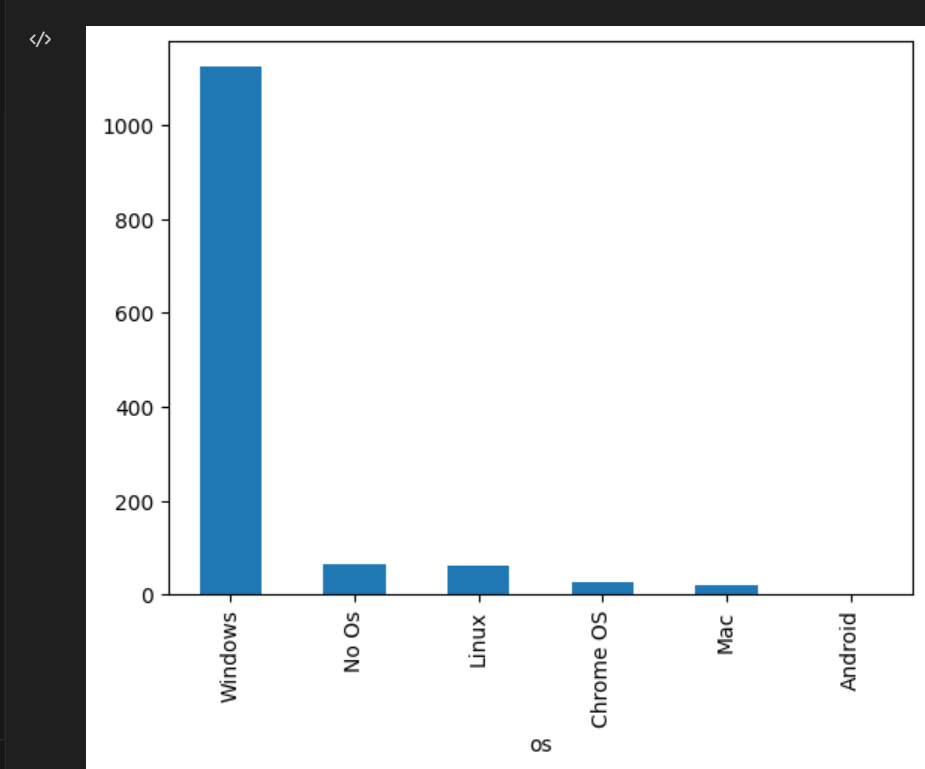
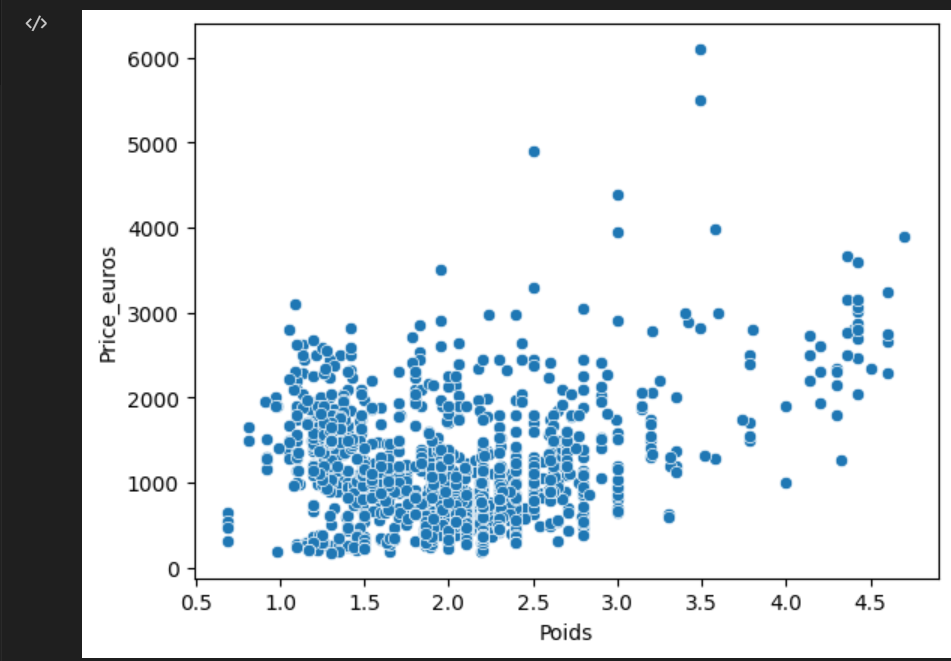


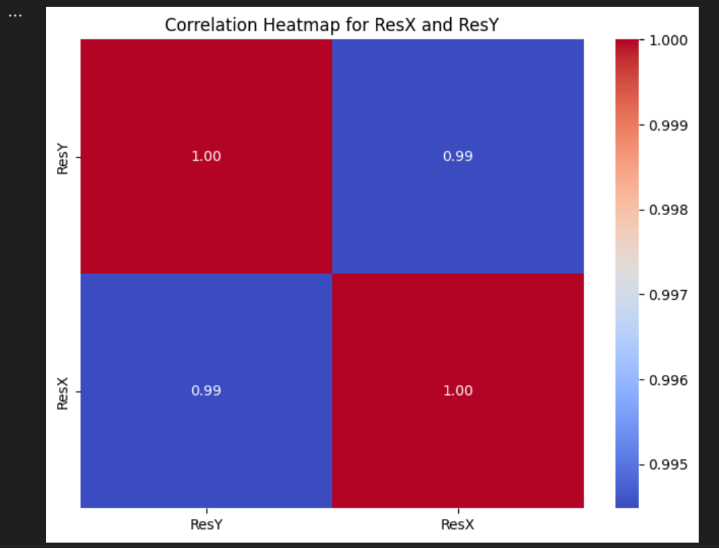
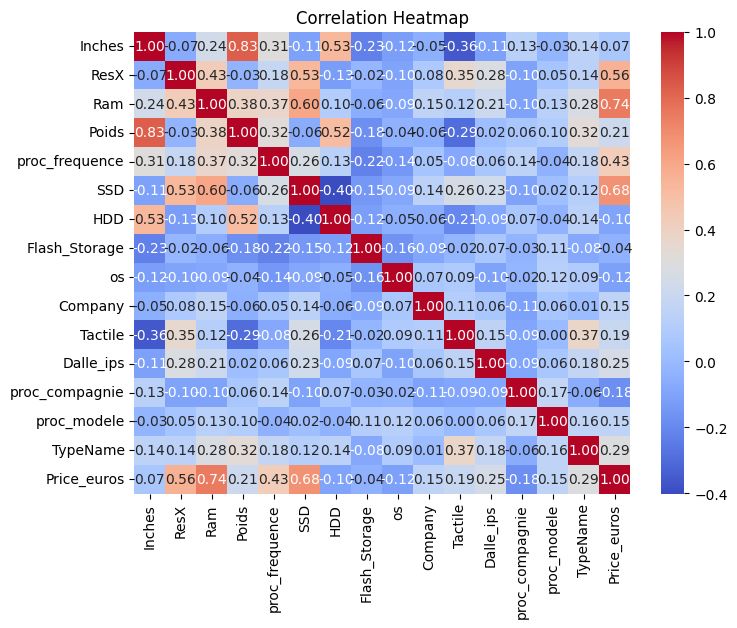










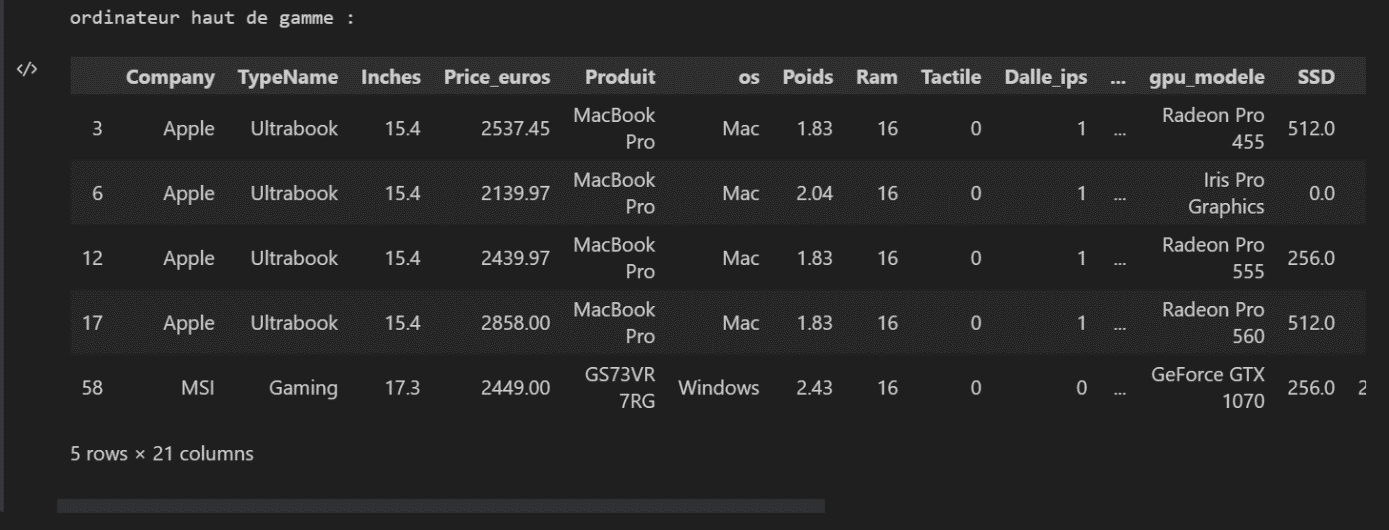
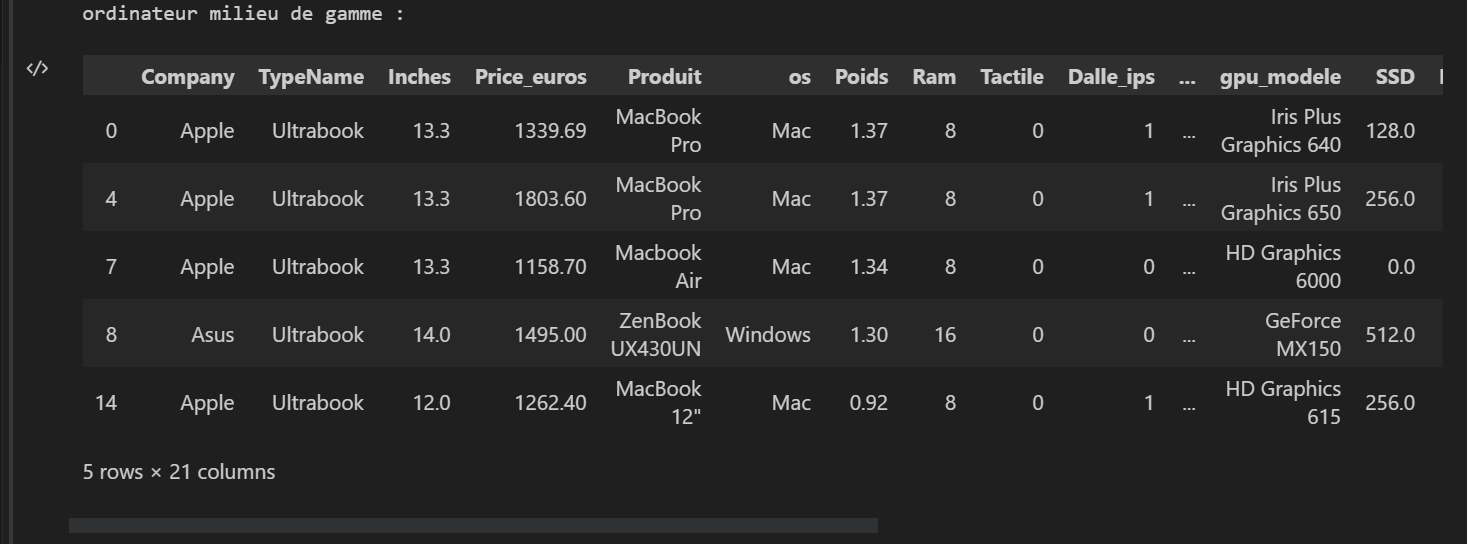
Ces boxplots et autres graphiques montrent les répartitions des valeurs de chaque variable ou spécification. Elles permettent de voir les valeurs les plus récurrentes. Et sur cette base, les différentes classes d’ordinateur seront créées lors du Clustering.

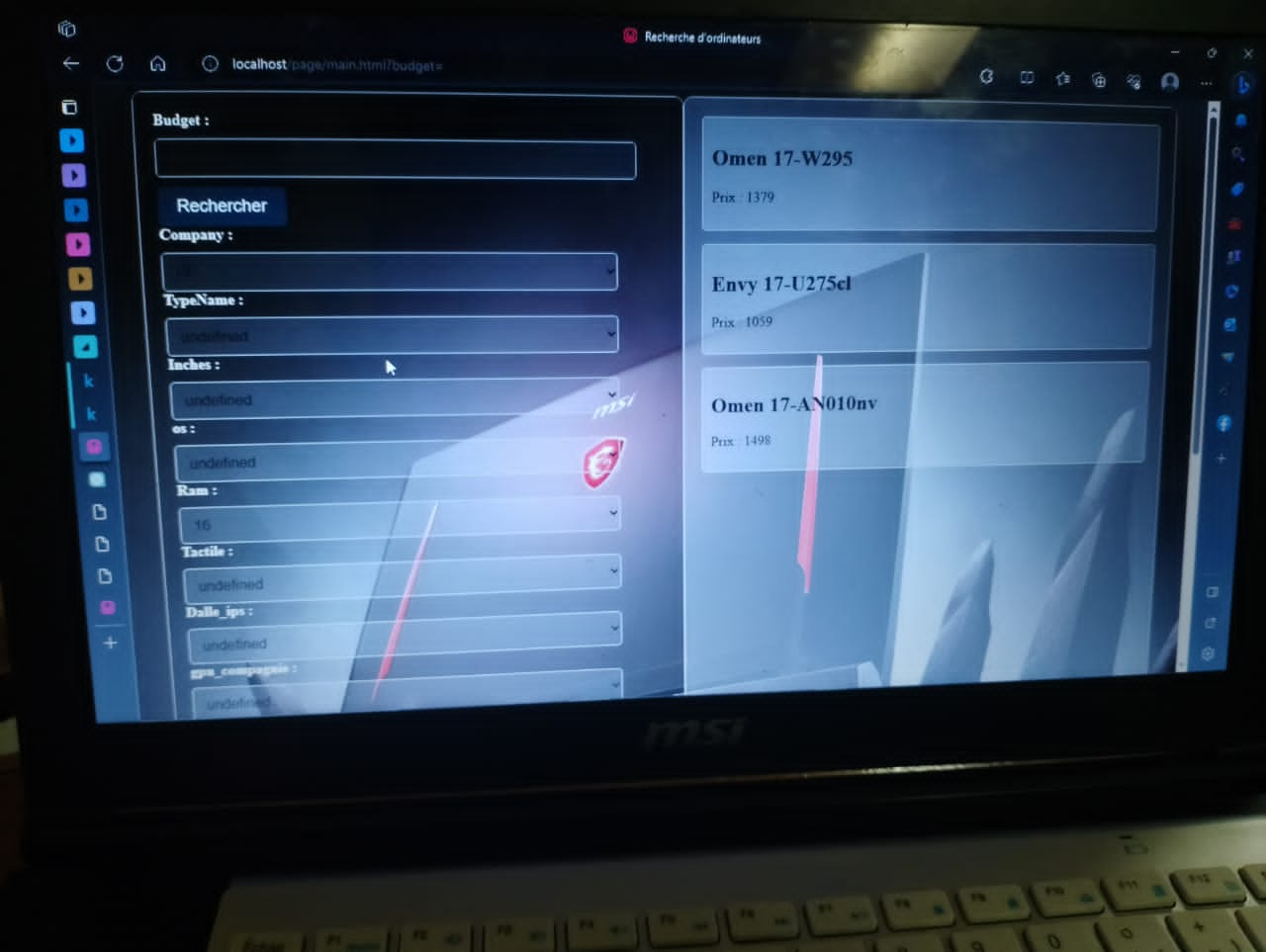
Ces tableaux de corrélation montrent comment les variables sont corrélées entre elles et permettent ainsi de regrouper certaines variables fortement corrélées qui rassemblent donc des observations similaires. Pour la suite, nous décidons de regrouper les variables dont le coefficient de corrélation est supérieur à 0.90.

1. **Le Clustering**

Après avoir traité les données pour obtenir les nouvelles colonnes ou variables correspondant aux différentes spécifications, nous allons procéder au Clustering par la méthode des k-means.

Nous obtenons ainsi trois classes d’ordinateur décrites dans l’introduction.

L’ACM étant terminée, les données sont récupérées dans un fichier de format JSON qui sera utilisée pour la conception de l’interface.



Le programme est ainsi fonctionnel et permet de faire des suggestions personnalisées et adaptées aux besoins de l’utilisateur.

1. **Perspectives**

* Enrichissement de la base de données avec davantage de modèles et de caractéristiques techniques.
* Intégration d’informations sur les usages (Gaming, mobilité, graphisme…)
* Utilisation du machine Learning (Deep Learning) pour augmenter les performances et le temps de calcul.
* Intégration des feedbacks des utilisateurs (User Expérience) pour enrichir le système et renforcer la pertinence
* Combinaison avec d’autres algorithmes de filtrage collaboratif ou hybride
* Extension à d’autres types de produits nécessitant un conseil personnalisé (exemple des voitures, ou d’appartements de location)
* Déploiement en ligne du système pour une utilisation facile et partout

L’intégration des feedbacks des utilisateurs et le passage à l’échelle nous semblent les axes prioritaires de recherche futurs.