

**Data science programming**

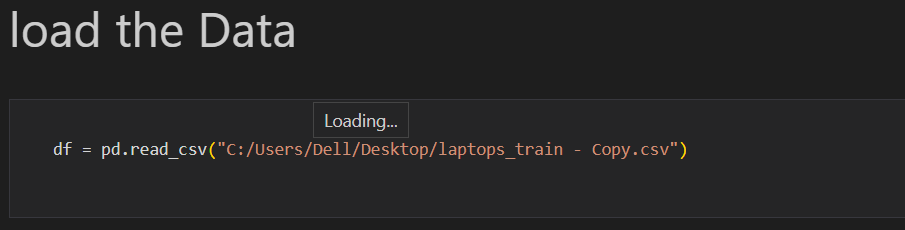
|  |  |
| --- | --- |
| **Name** | **ID** |
| Ahmed khaled | 42010387 |
| Mahmoud Ashraf | 42020014 |
| Zeyad Mohamed | 42010448 |
| Verena Gamal | 42010044 |
| Hagar Galal | 42010084 |
| Hala Khaled | 42020007 |
| Hadeer Alkady | 42010436 |

**Laptops**

**Problem definition :**

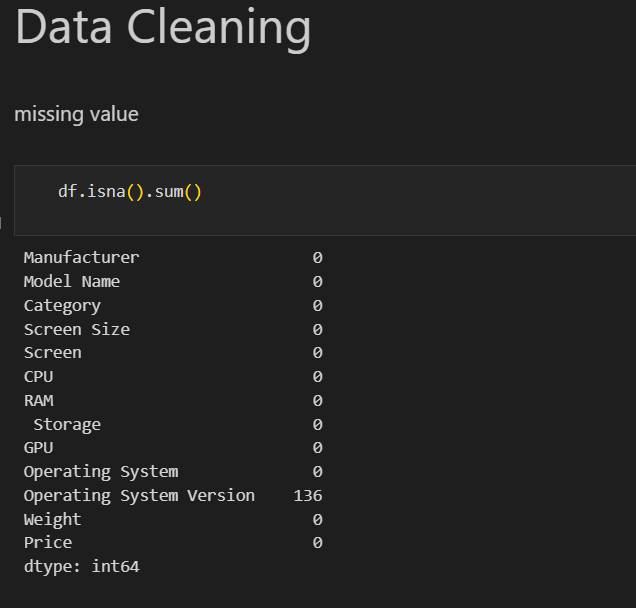
Our data is talking about a laptop store that contains this data (factory name, model name, screen size, screen type, CPU, RAM, Storage, GPU, any operating system, operating system versions, weight and price)

1. **In the beginning , we read the data to work on it.**

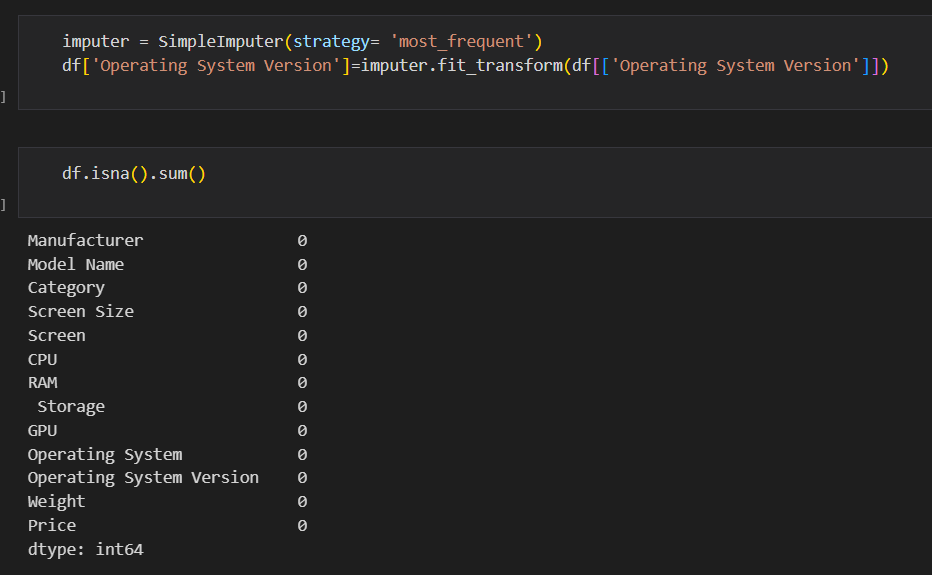


1. **Next, we started doing data cleaning for our data set**
2. We had a missing value in the **operating system version** weFill it with most frequent the value repeated in the Column ,

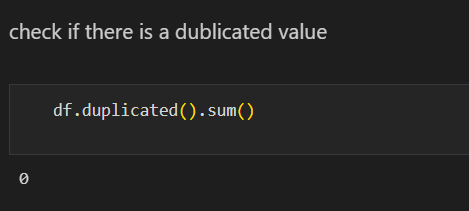
Before the edit :



After editing :



**3- Then we made sure that it was not duplicate, and we did not have in our data.**

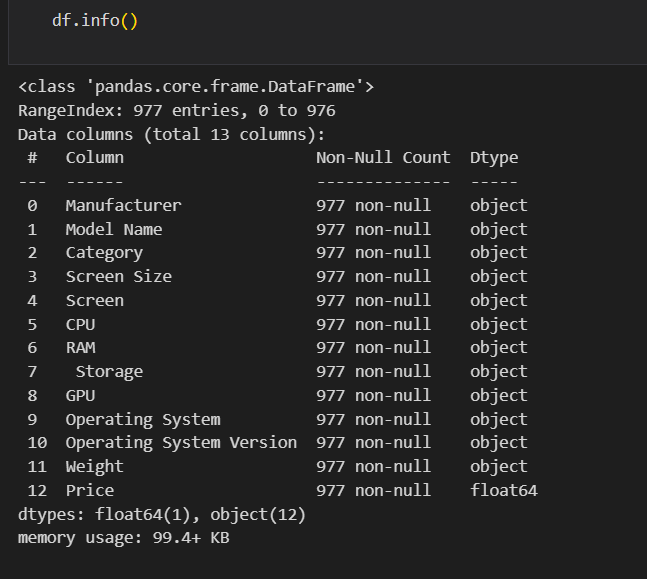


**4-** **we made unique for each column ,**

It was in weight kg shilling

And there was a screen size that had a double quotation for it

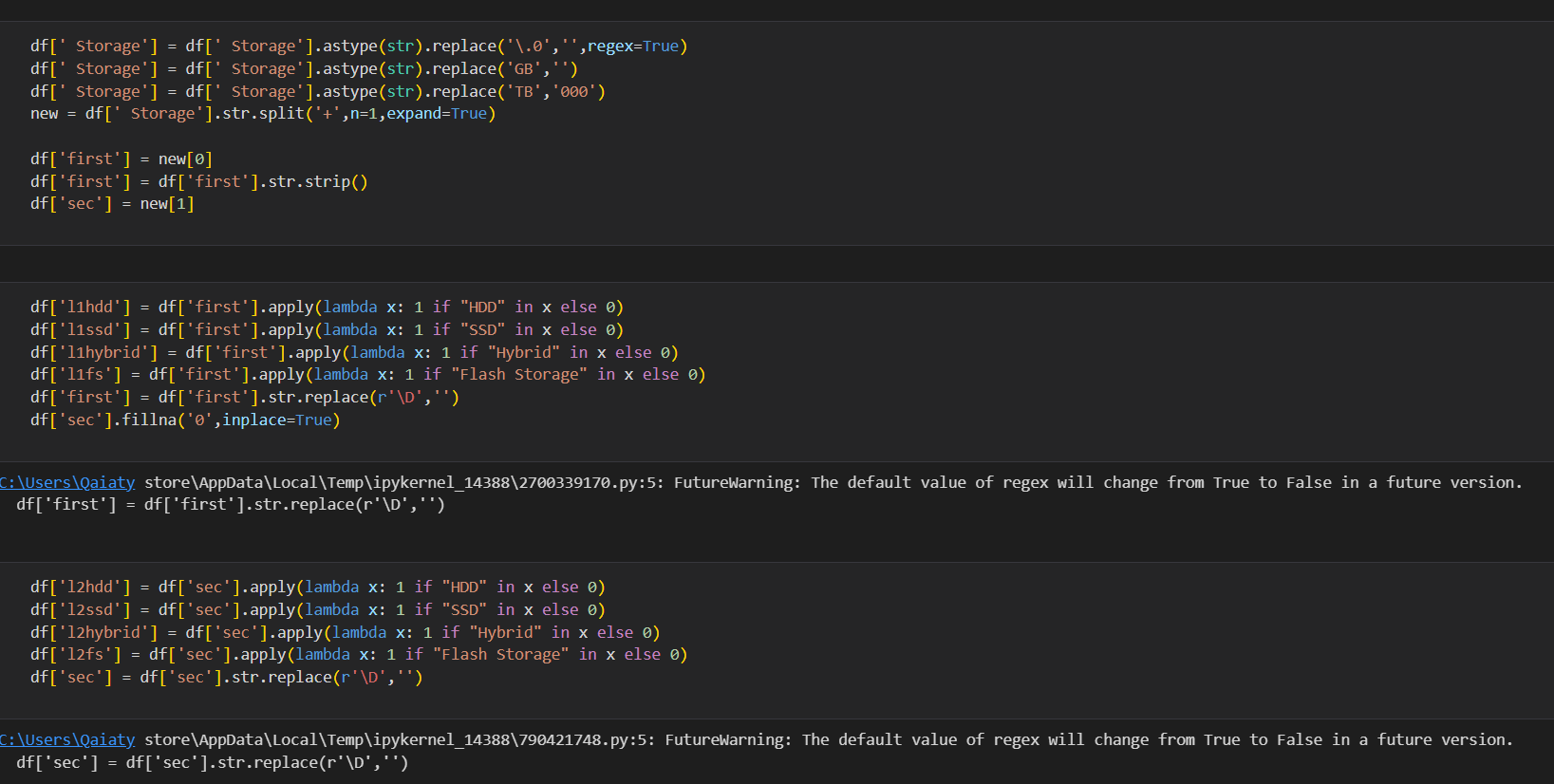
And the ram had GB shillings, and all of this was converted to float



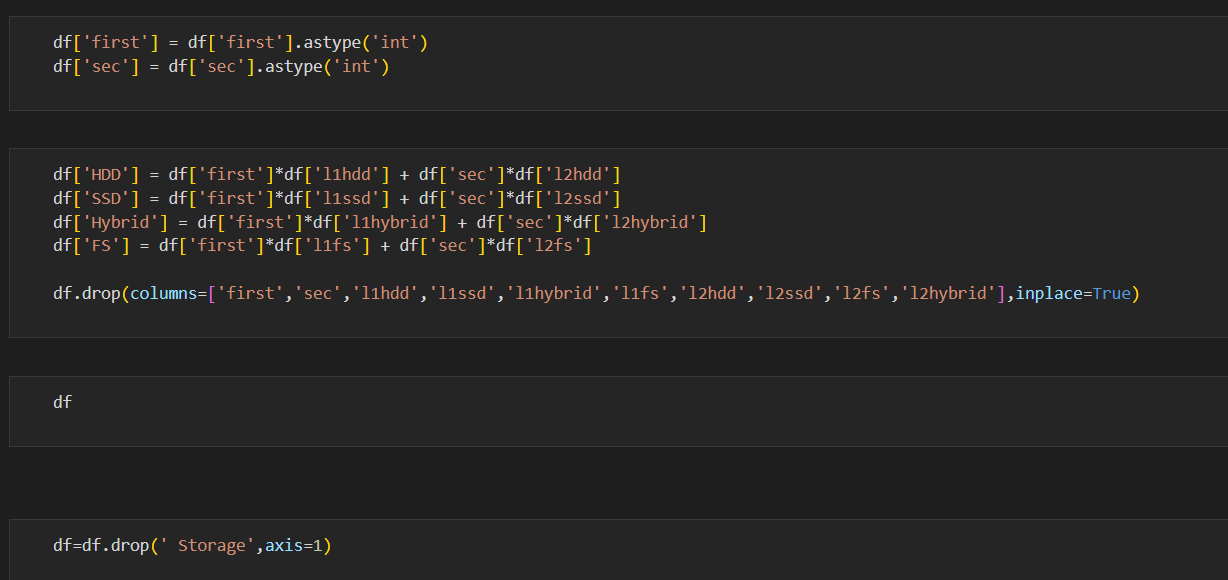
And we start to fix this

**5-And the storage we split for four coulombs hdd, ssd, flash storage, hybrid**

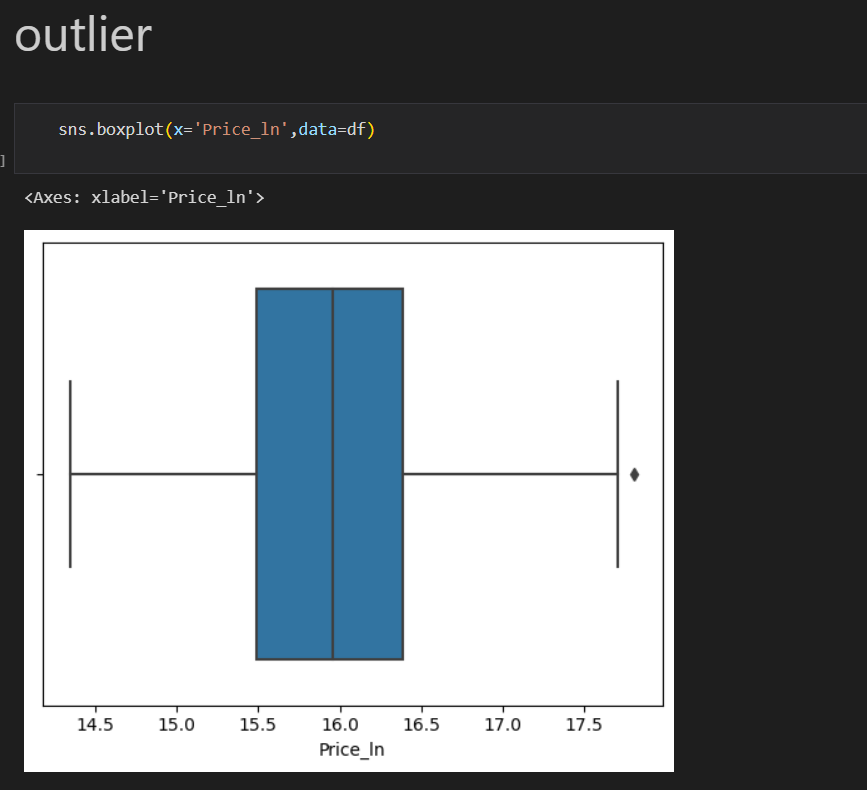
**For each lab the size storage store in the type of storage hdd, ssd, flash storage, hybrid**

****

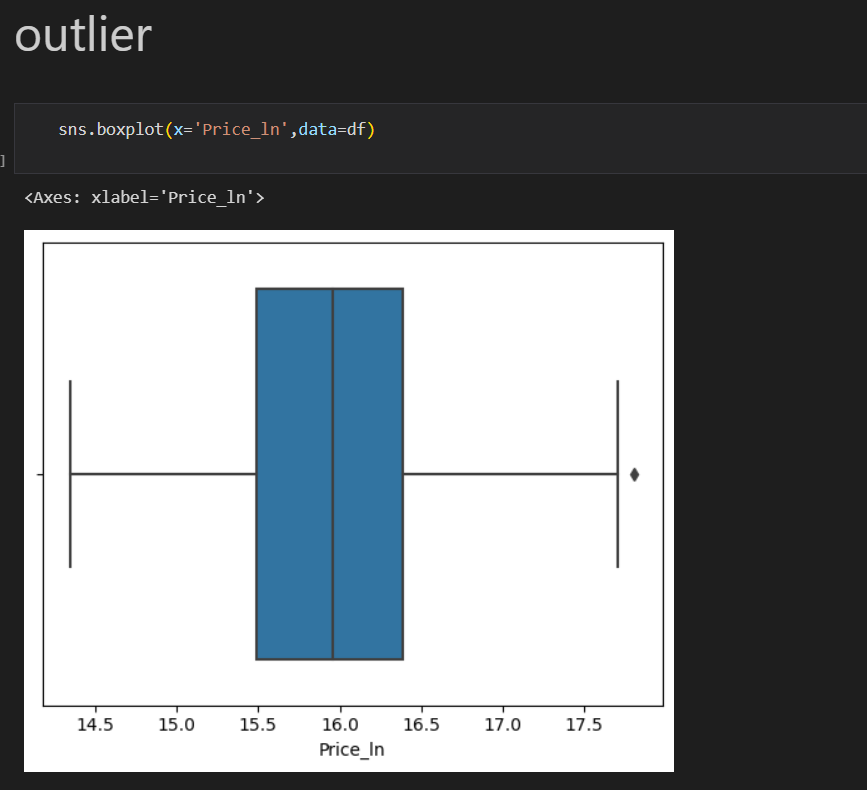
**Then we drop the column of storage**

****

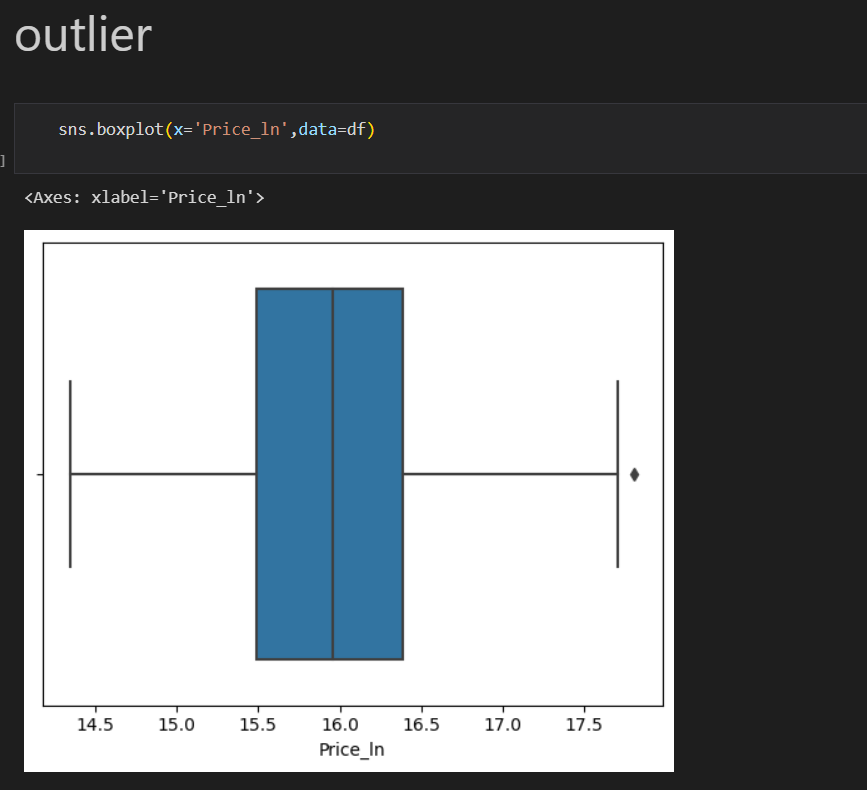
1. **Then we calculated the outlier price in ln and make Boxplot**

****

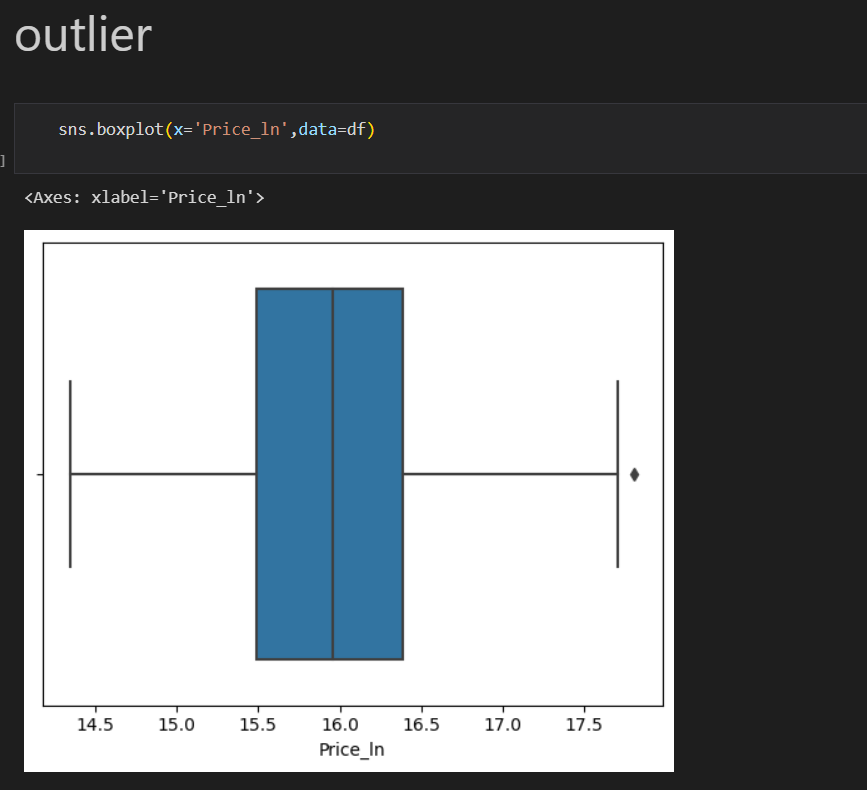
1. **RAM was a normal outlier and we made Boxplot to it**

****

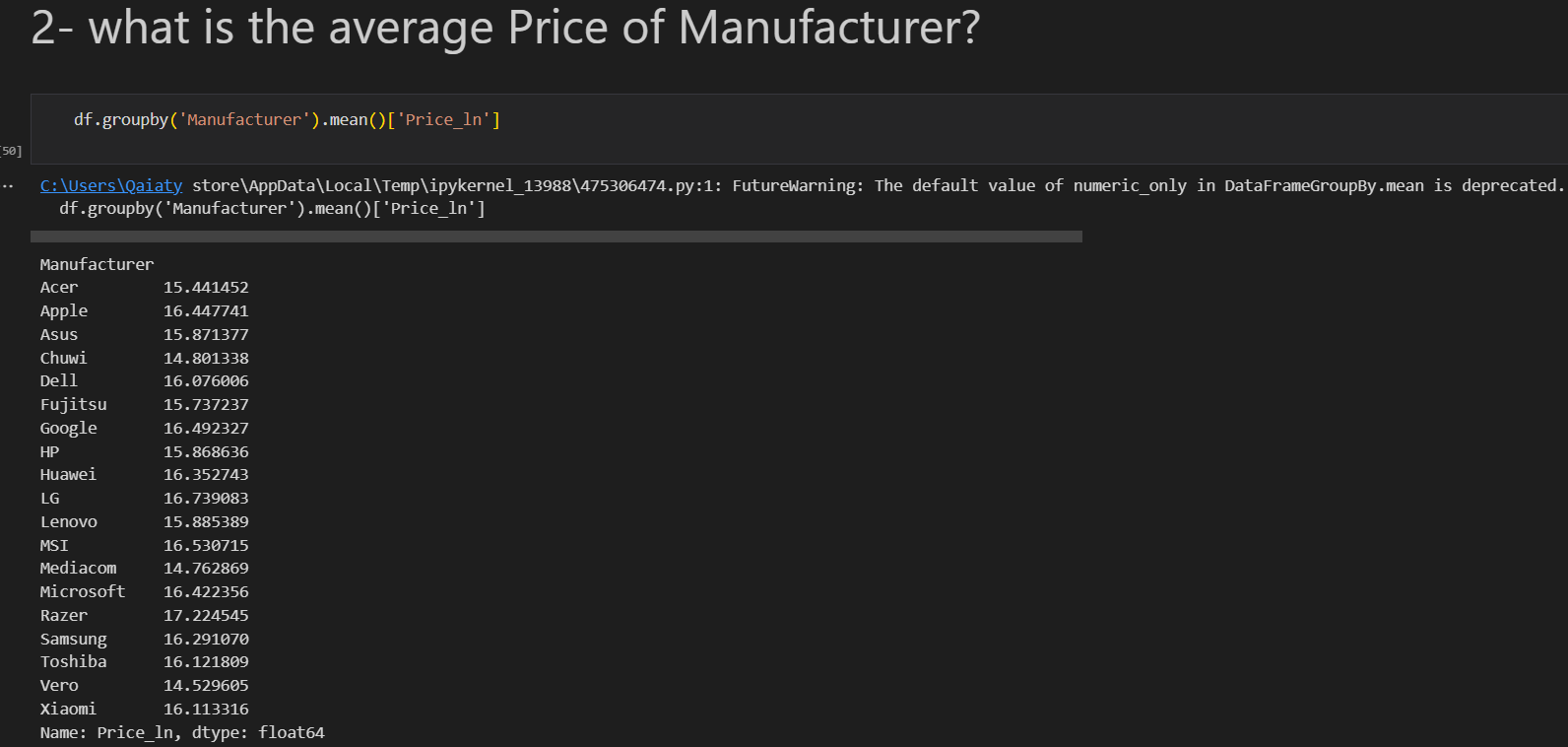
1. **wight was normal outlier , and we made Boxplot**

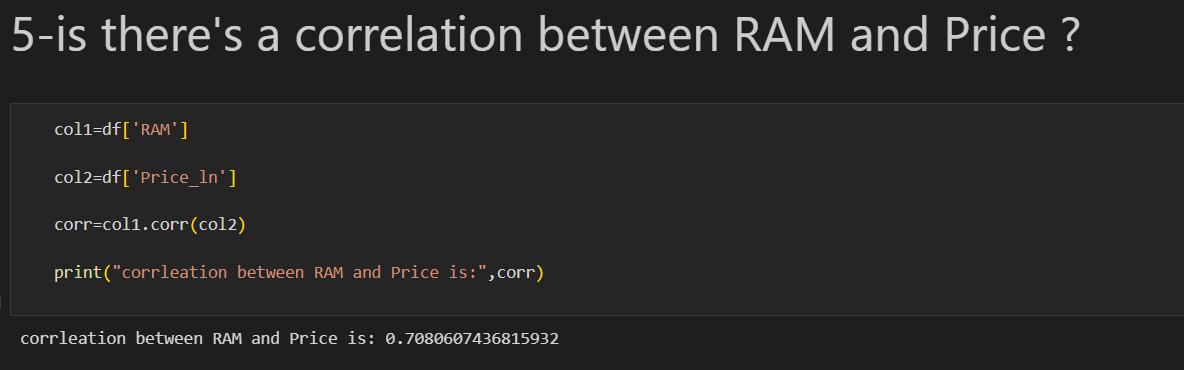
****

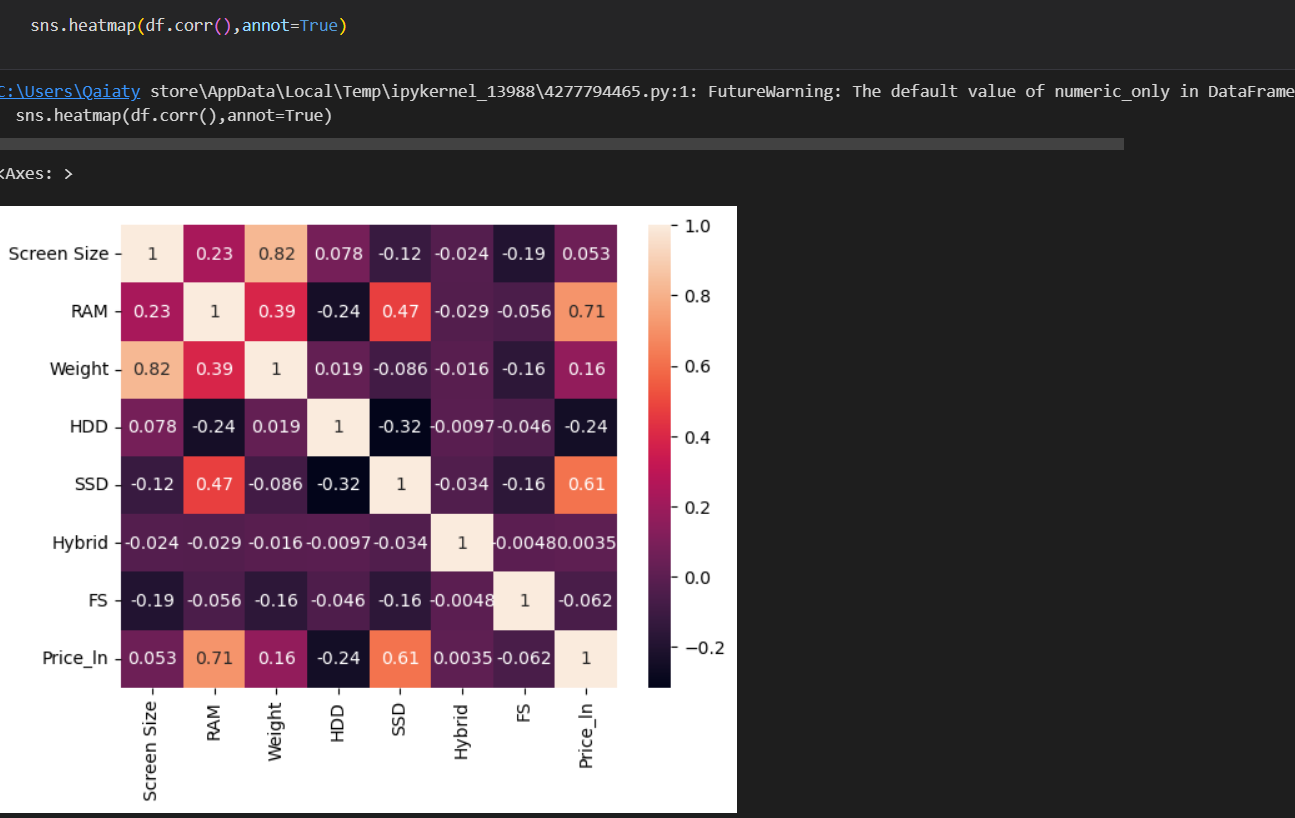
1. **Made Boxplot to Screen size**

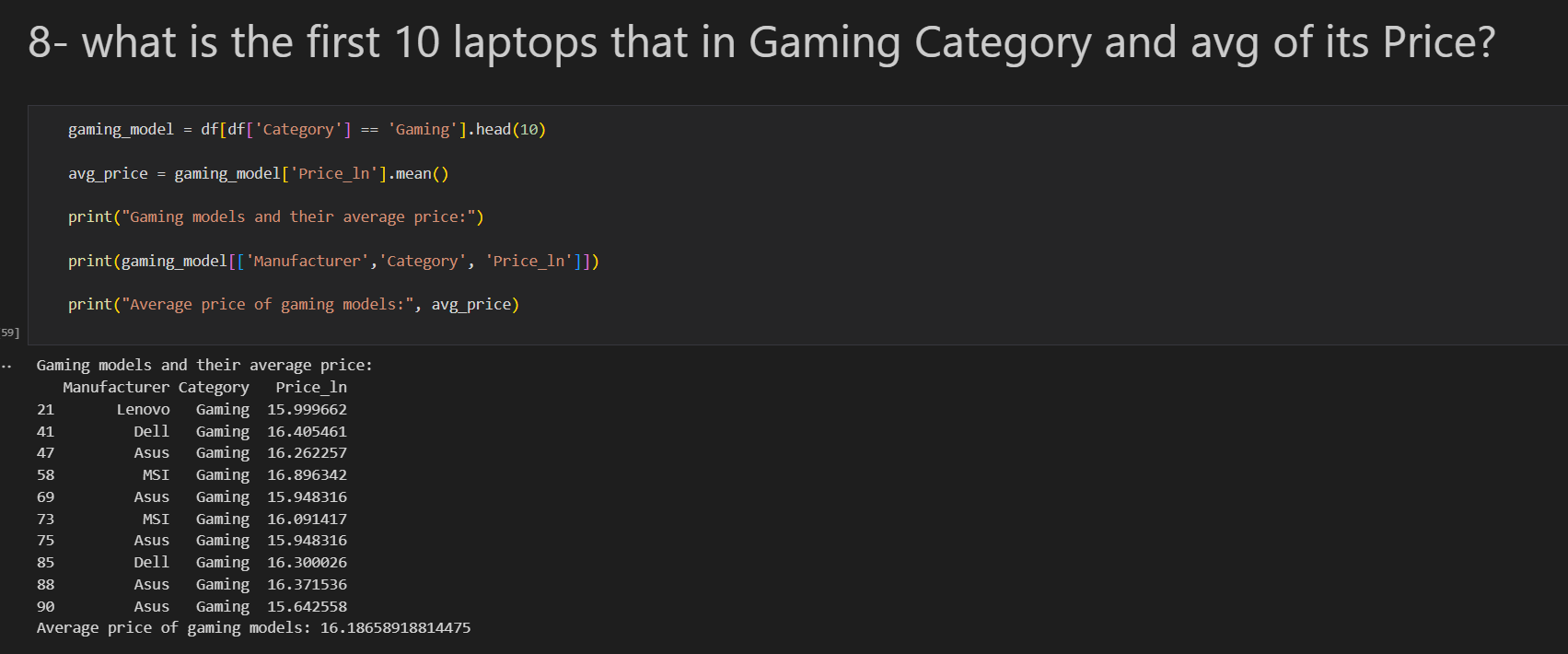
****

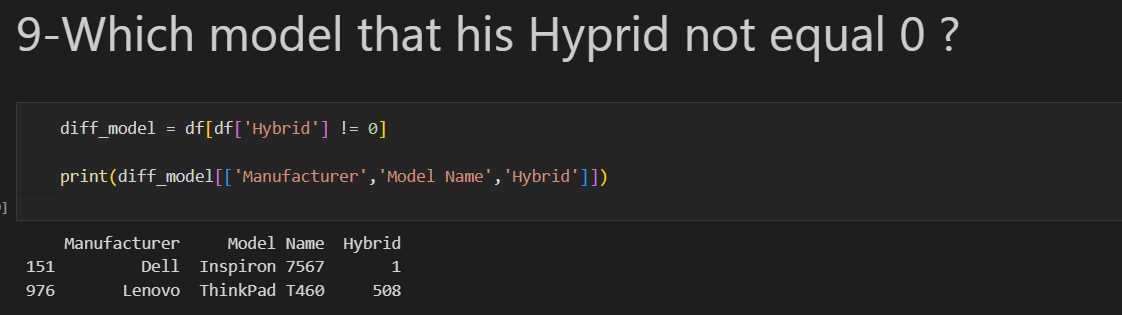
**2- Visualization by python**

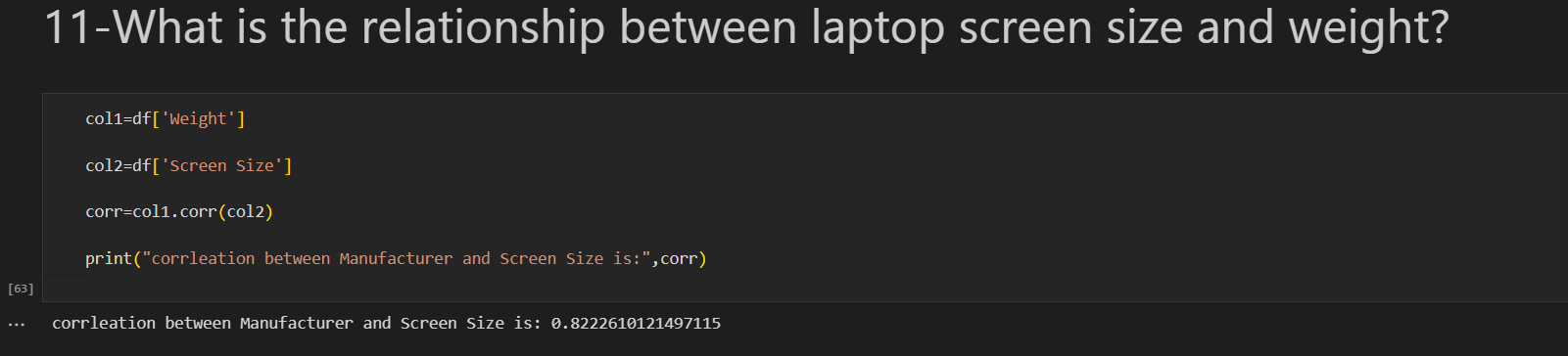
****

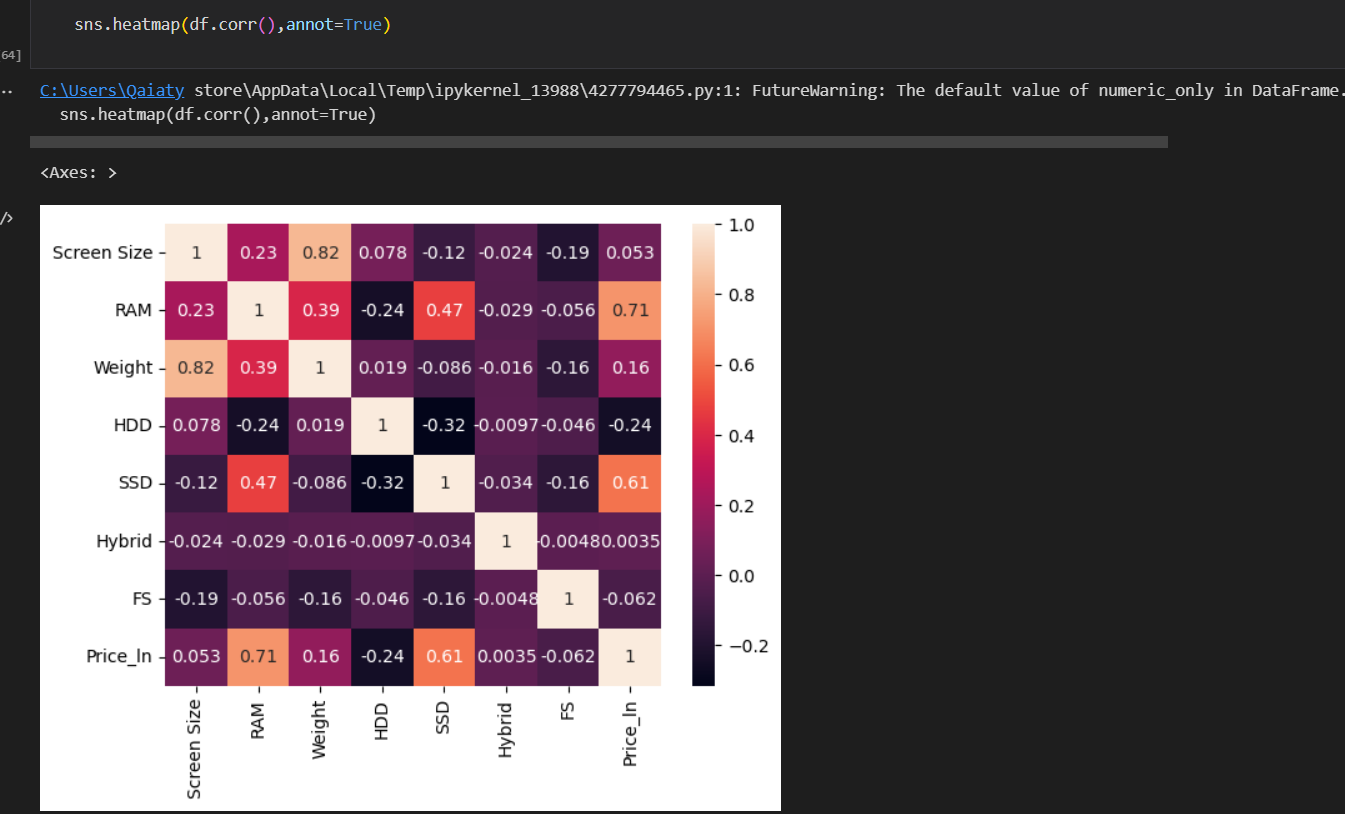
****

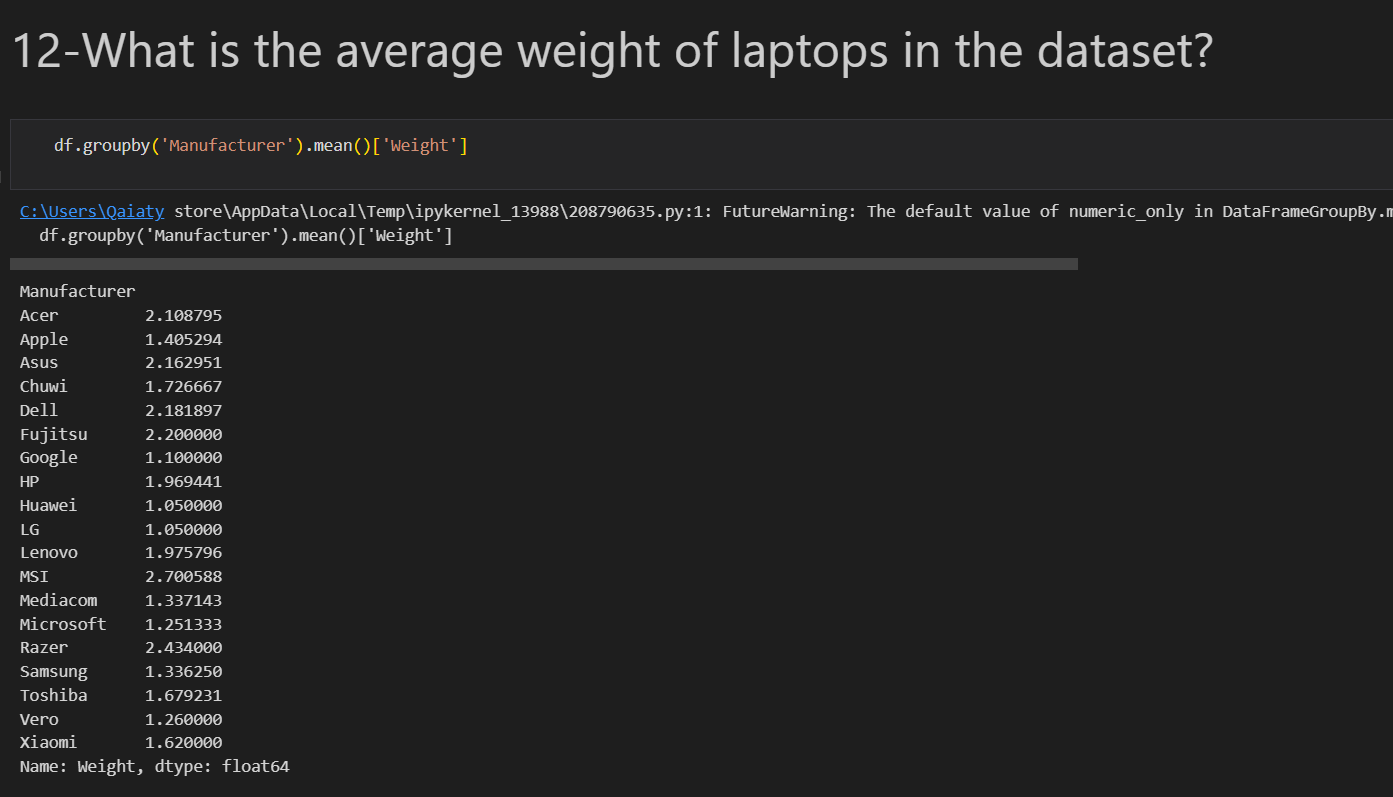
****

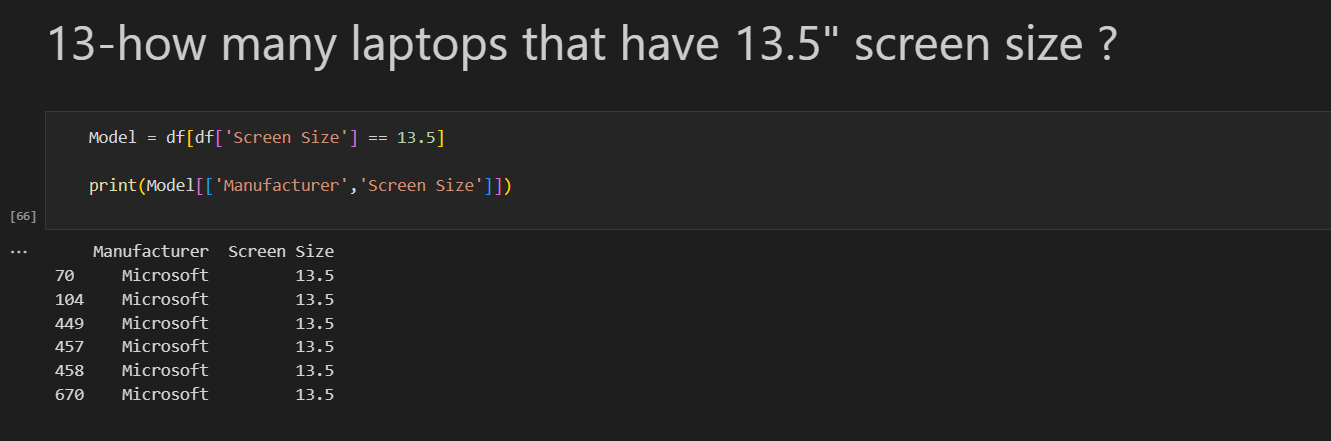
****

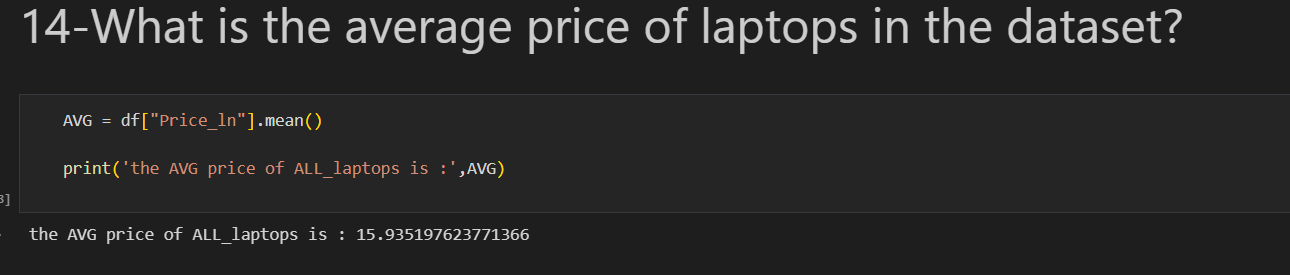
****

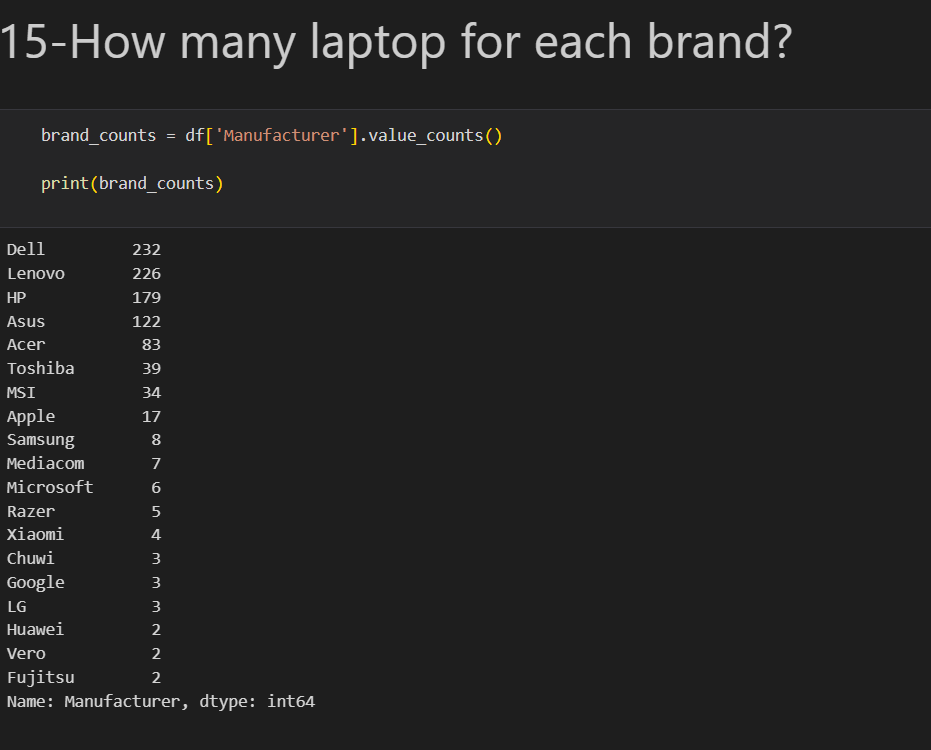
****

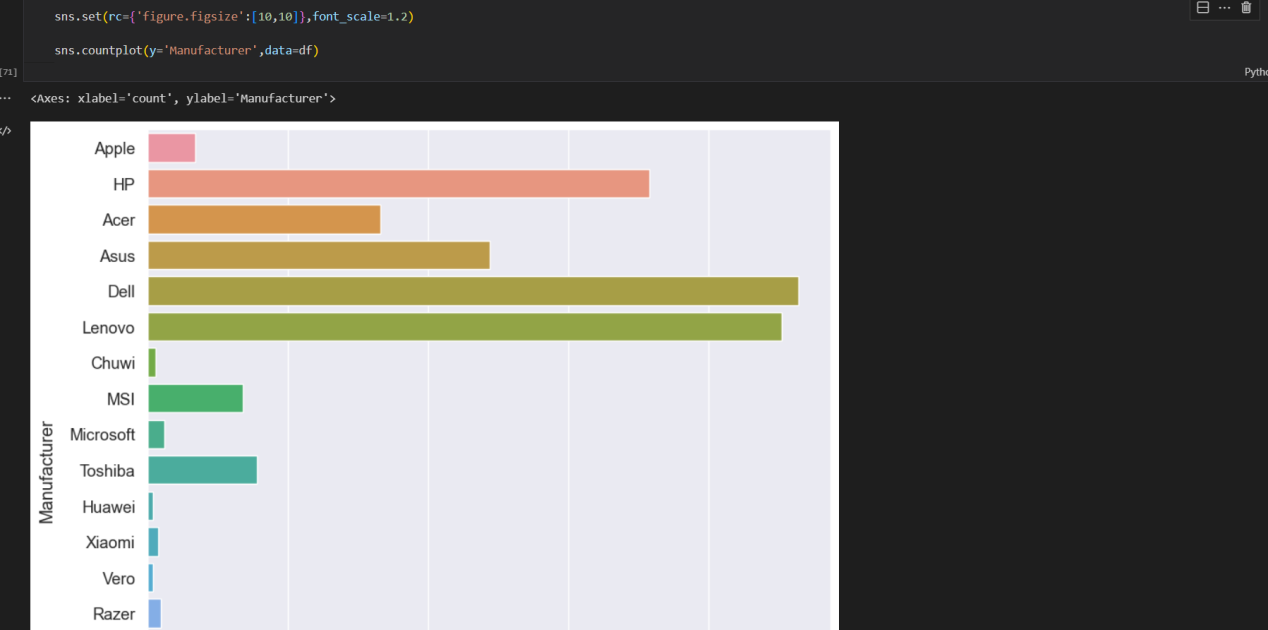
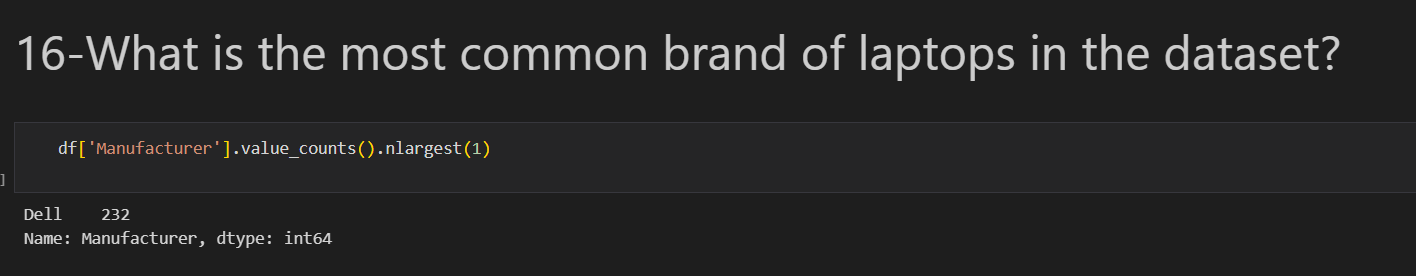
****

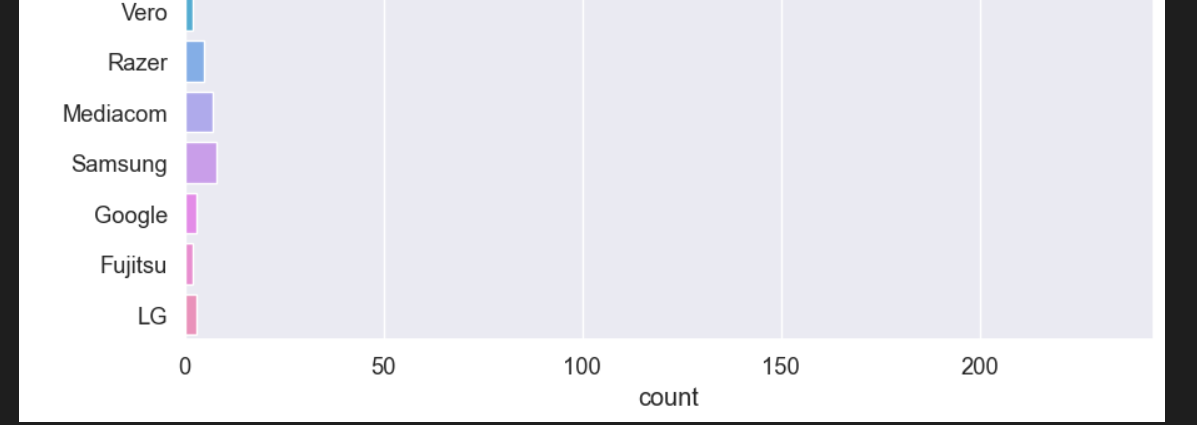
****

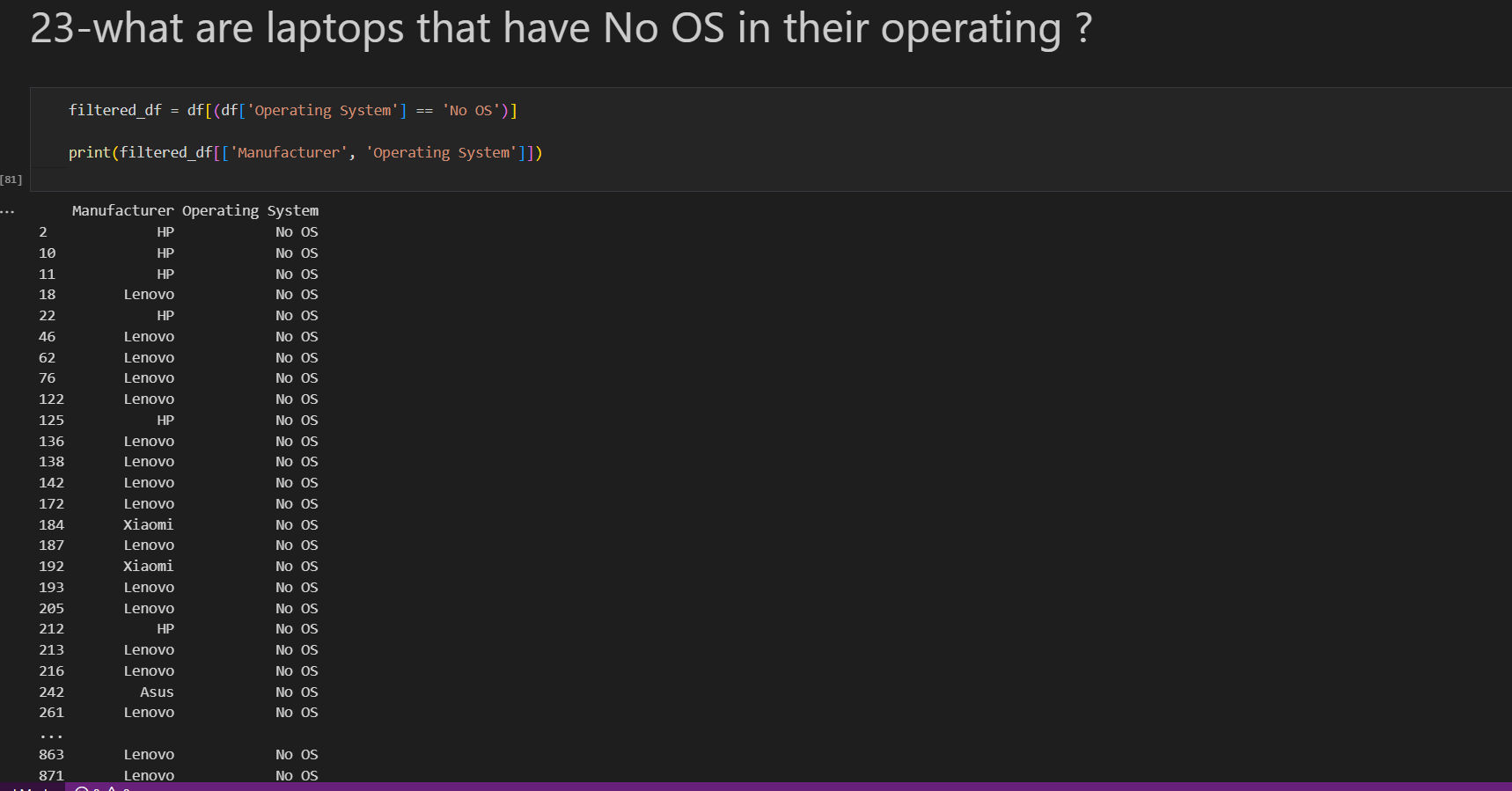
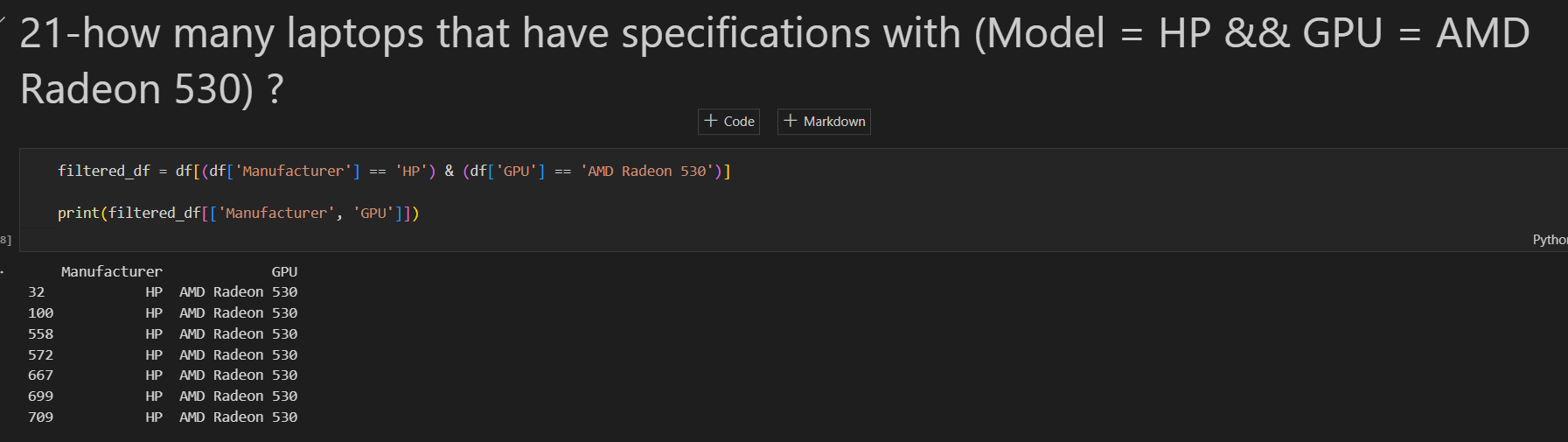
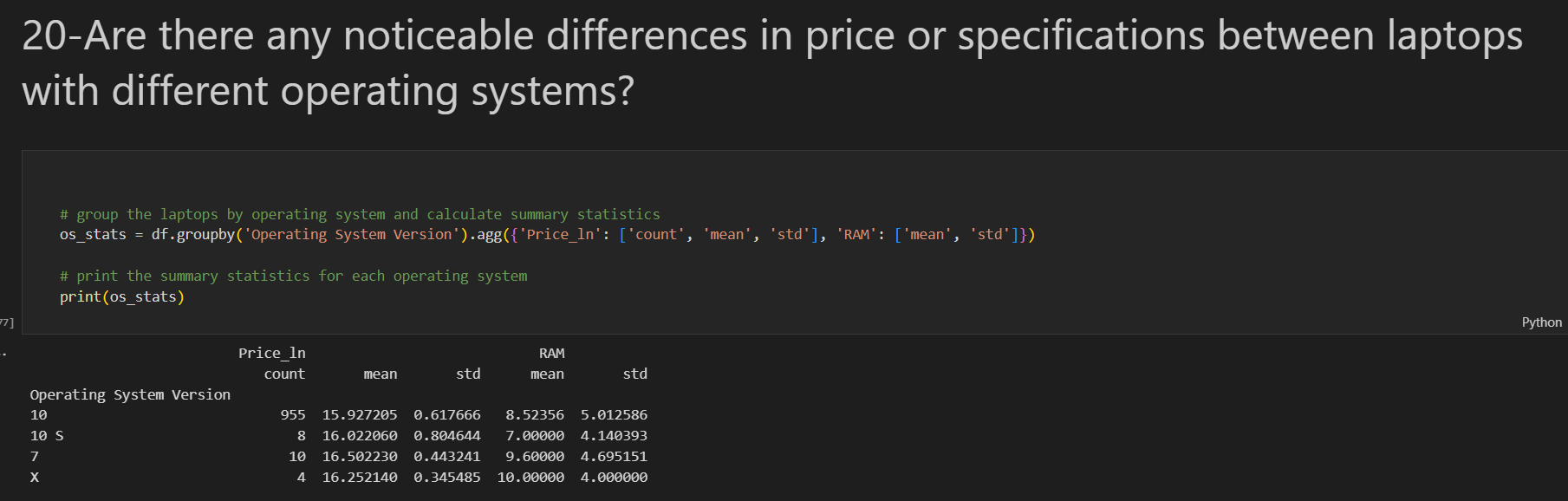
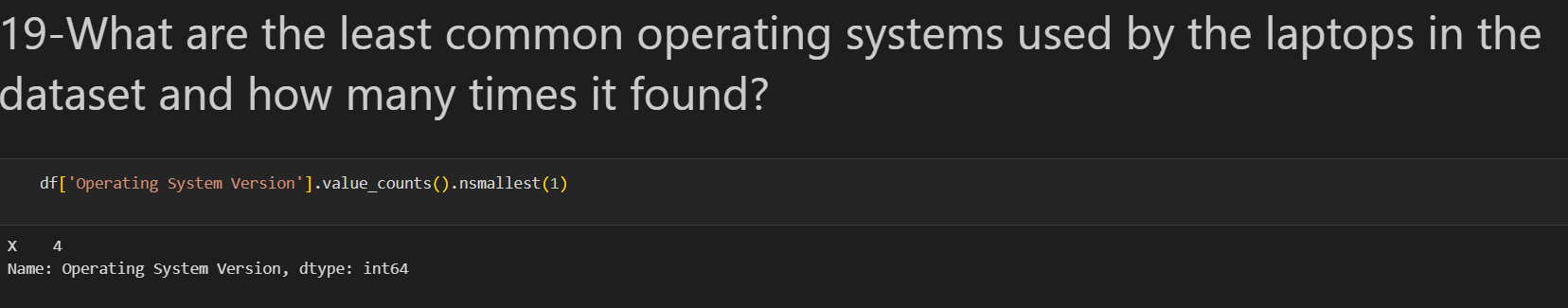
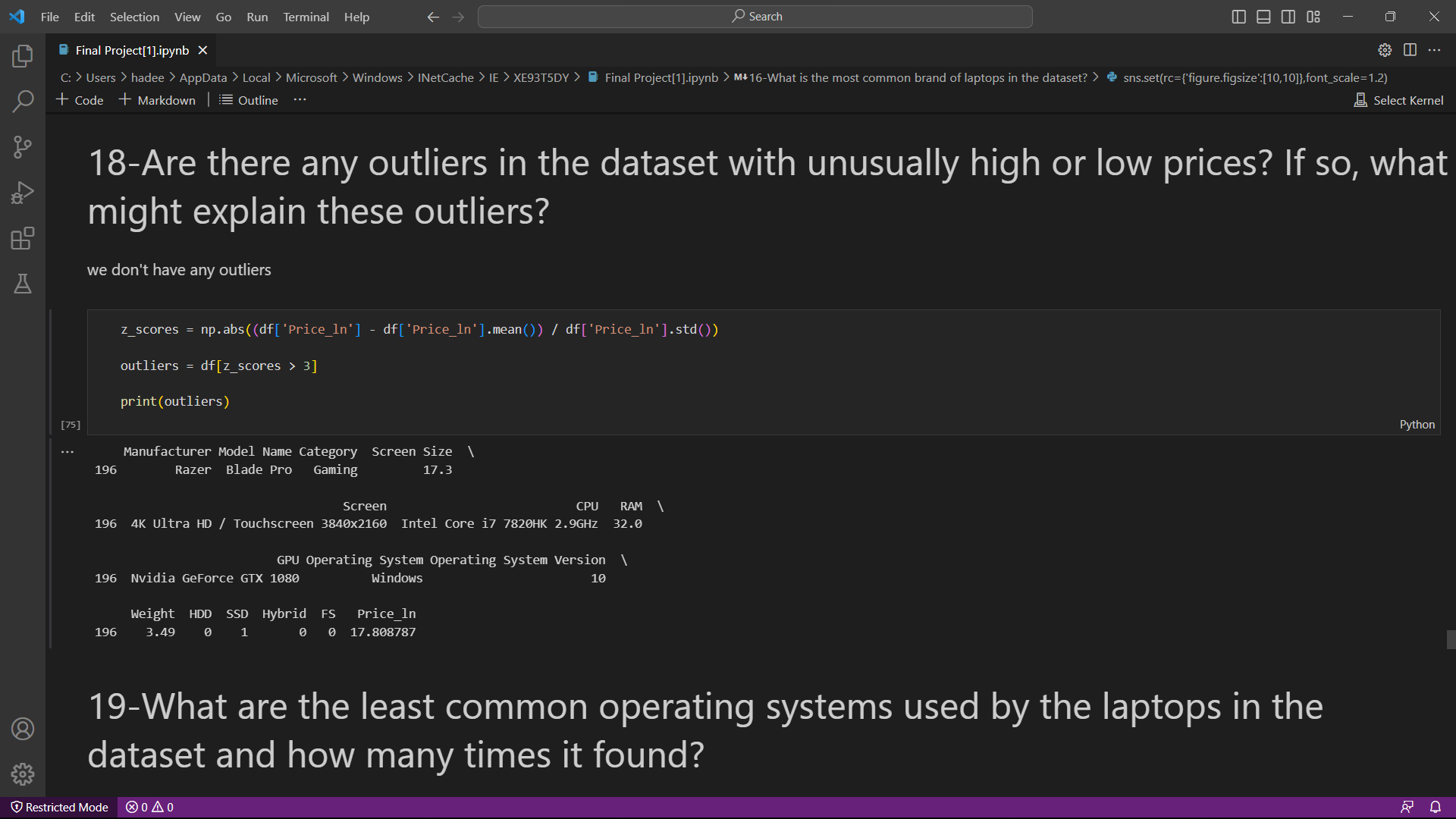
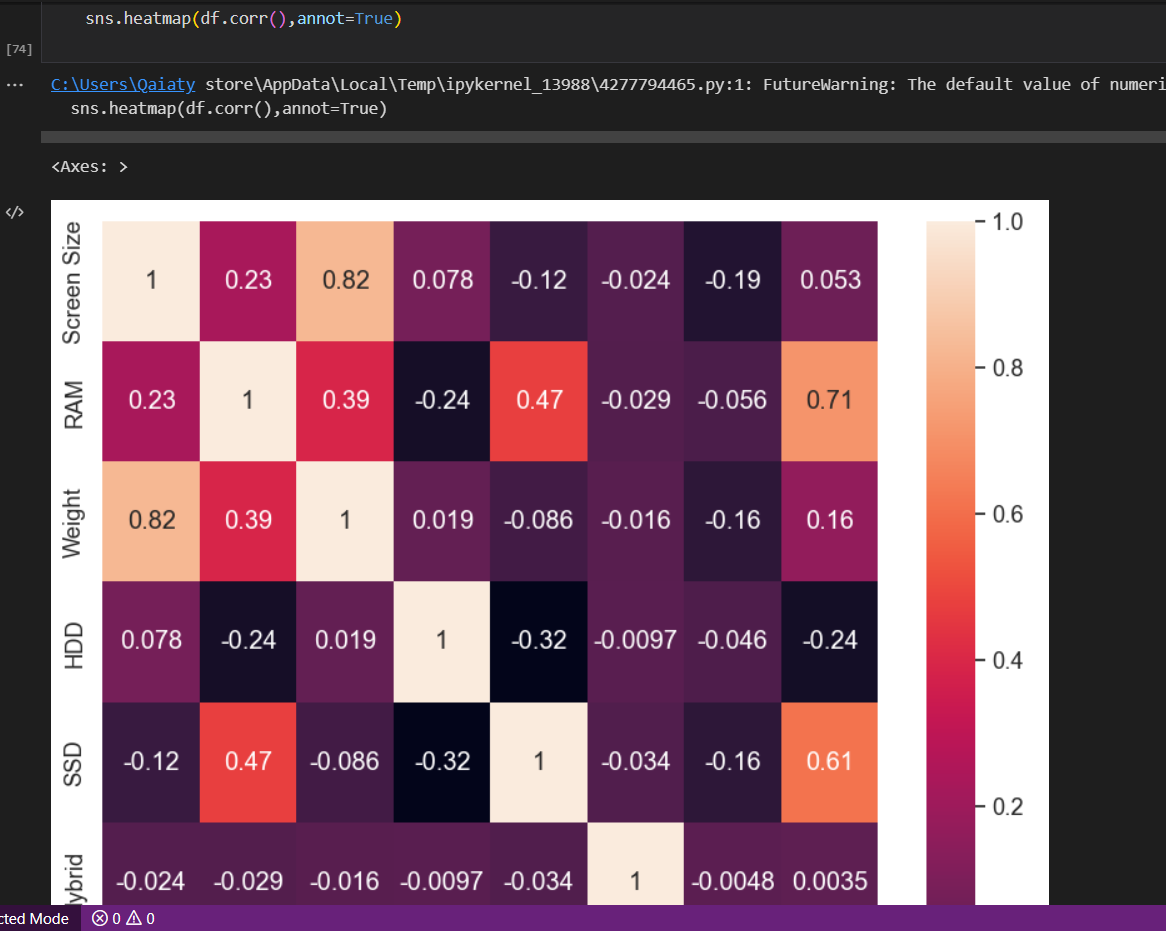
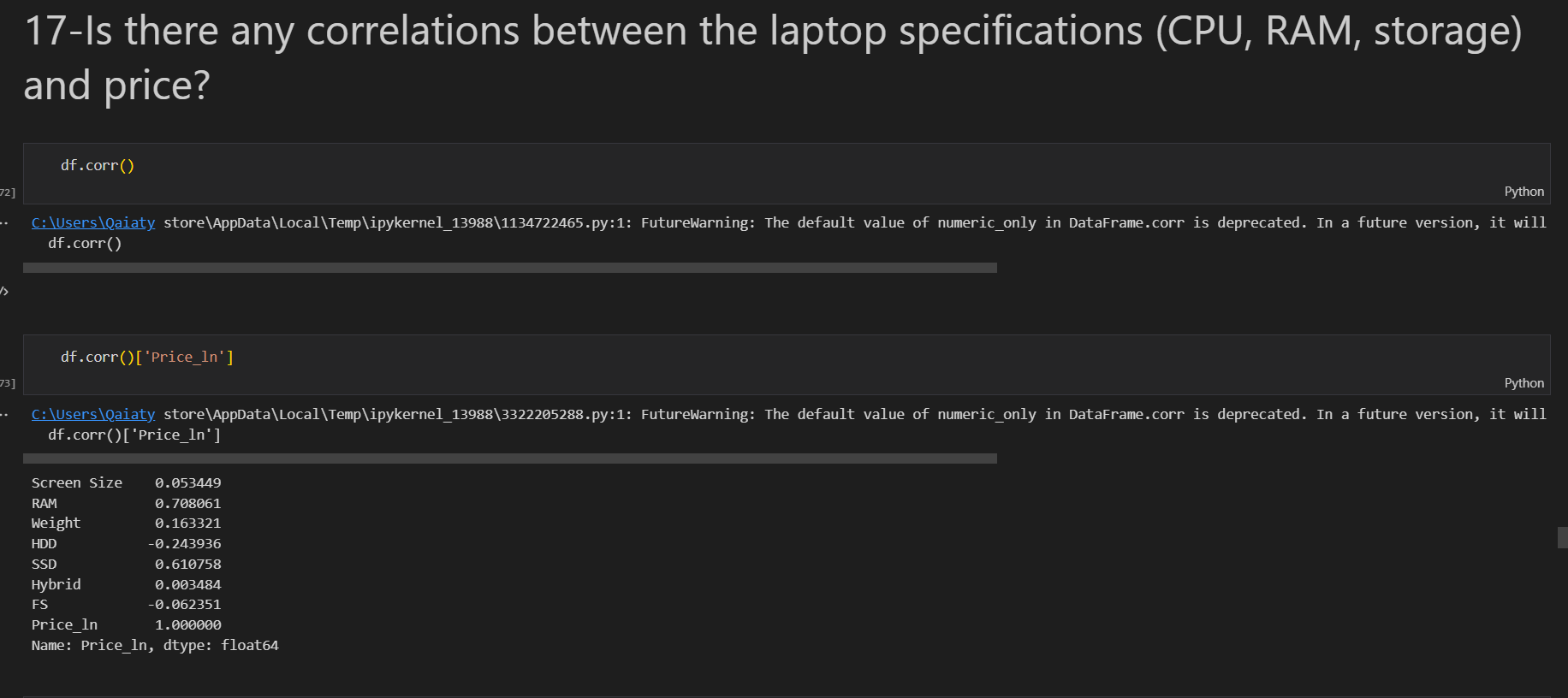
****

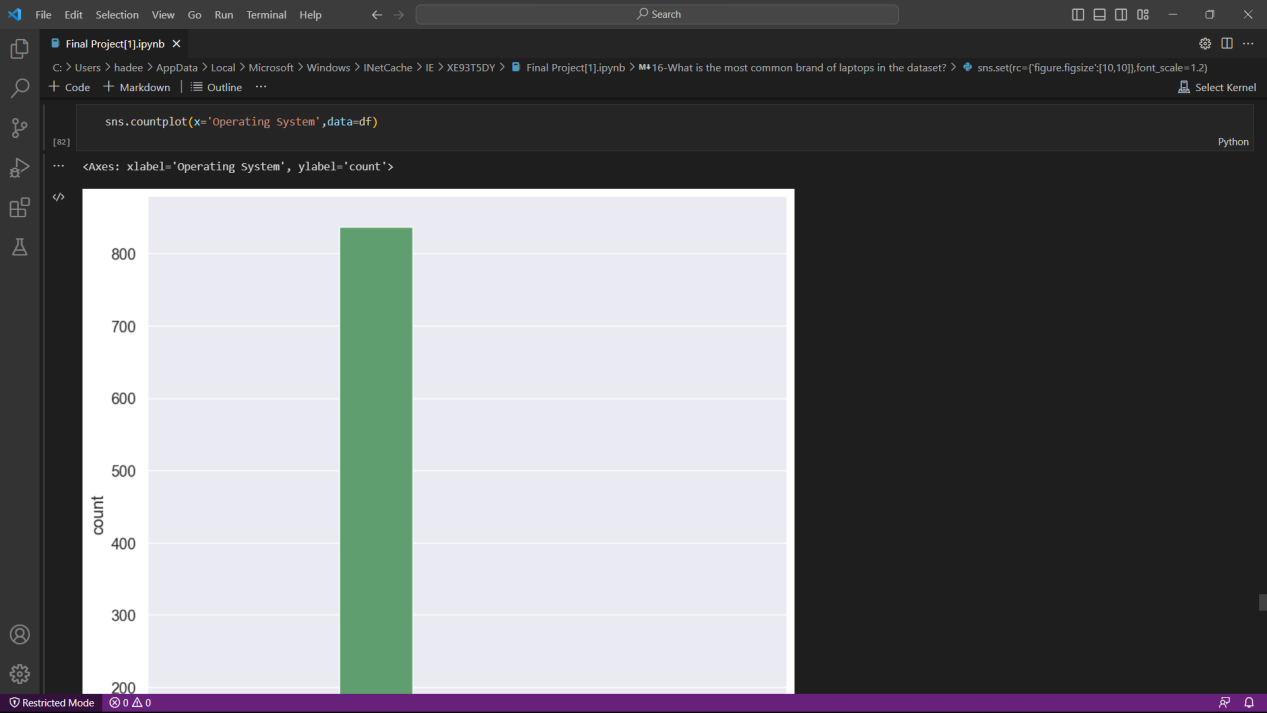
****

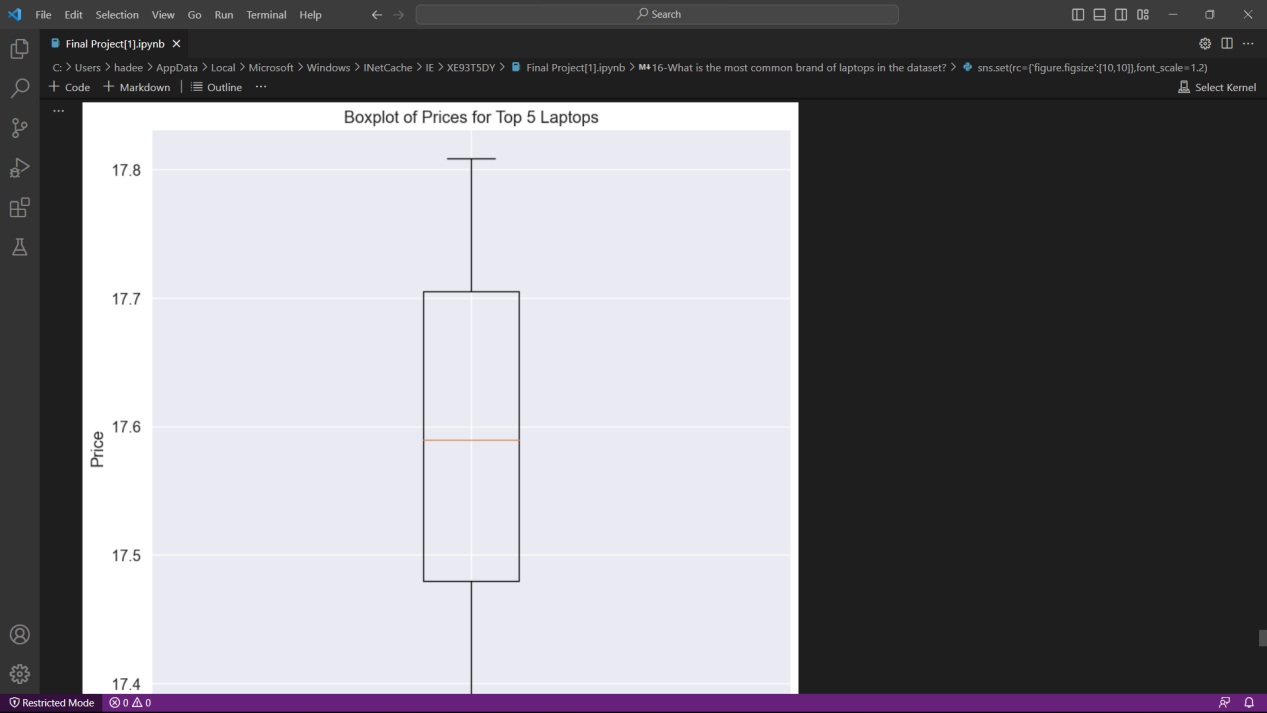
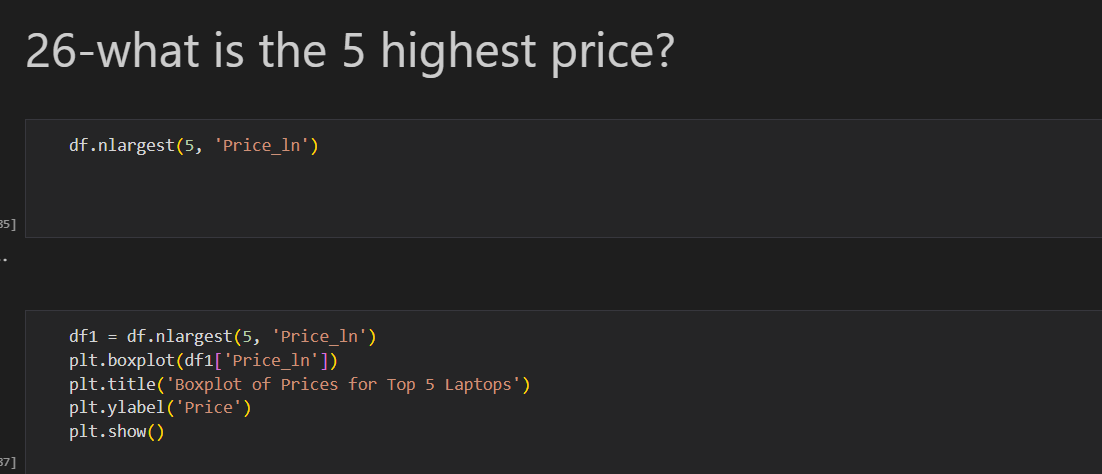
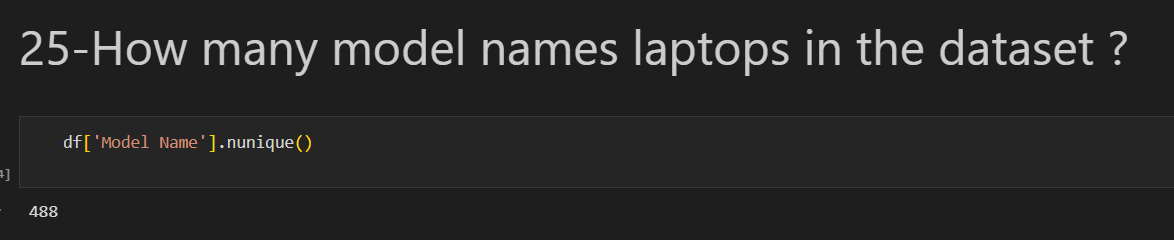
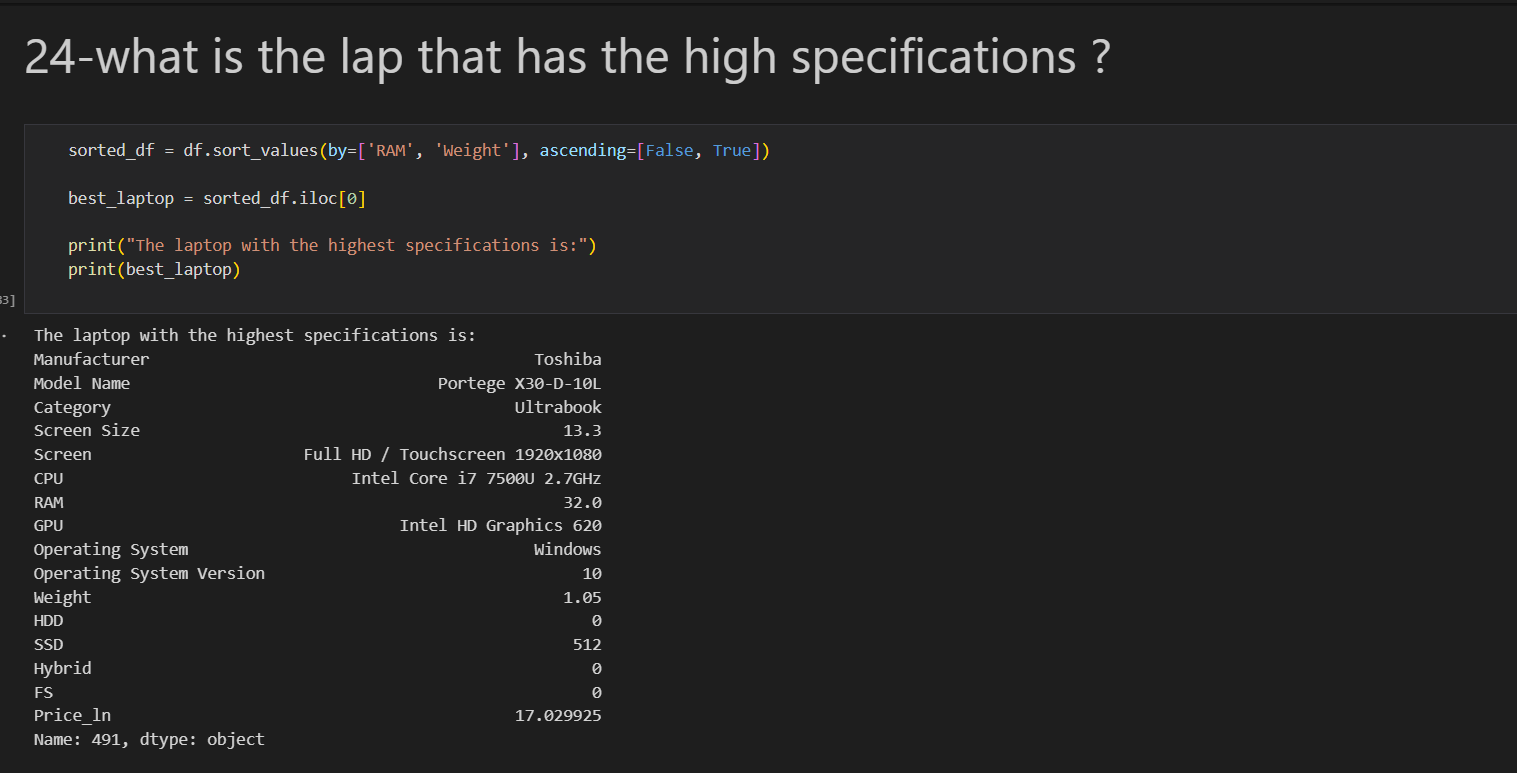
****

****

****

****

****

****

**Machine learning :**

The main idea of machine learning is to train a linear regression model to predict the price of a laptop based on its specifications. The code uses the Pandas library to read in a CSV file containing data about laptops and selects a subset of the columns to use as features for the model. It then uses the Scikit-learn library to split the data into training and testing sets, transform the categorical variables using a one-hot encoder, and fit a linear regression model to the training data. Finally, it evaluates the performance of the model on the training and testing sets using the R-squared score. The overall goal of this process is to build a model that can accurately predict the price of a new laptop based on its features.