**23907-MDSC-102P-ESE-REPORT**

The dataset I selected for the MDSC-102P is on Laptop information.

This dataset contains the following columns:

Manufacturer – provides the name of the manufacturing company, for instance Apple

Model Name – gives the model name of the laptop, for instance MacBook Air

Category – gives the details of the laptop to which category it belongs to, for instance Notebook

Screen Size – gives size of laptop’s screen in inches, for instance 13.3 inches

Screen – provides more information on laptop’s screen, for instance Full HD 1920 x 1080

CPU – provides information about the processor, for instance Intel Core i7

RAM – gives information about RAM, for instance 16GB

Storage – gives information about the storage, for instance 256 GB SSD

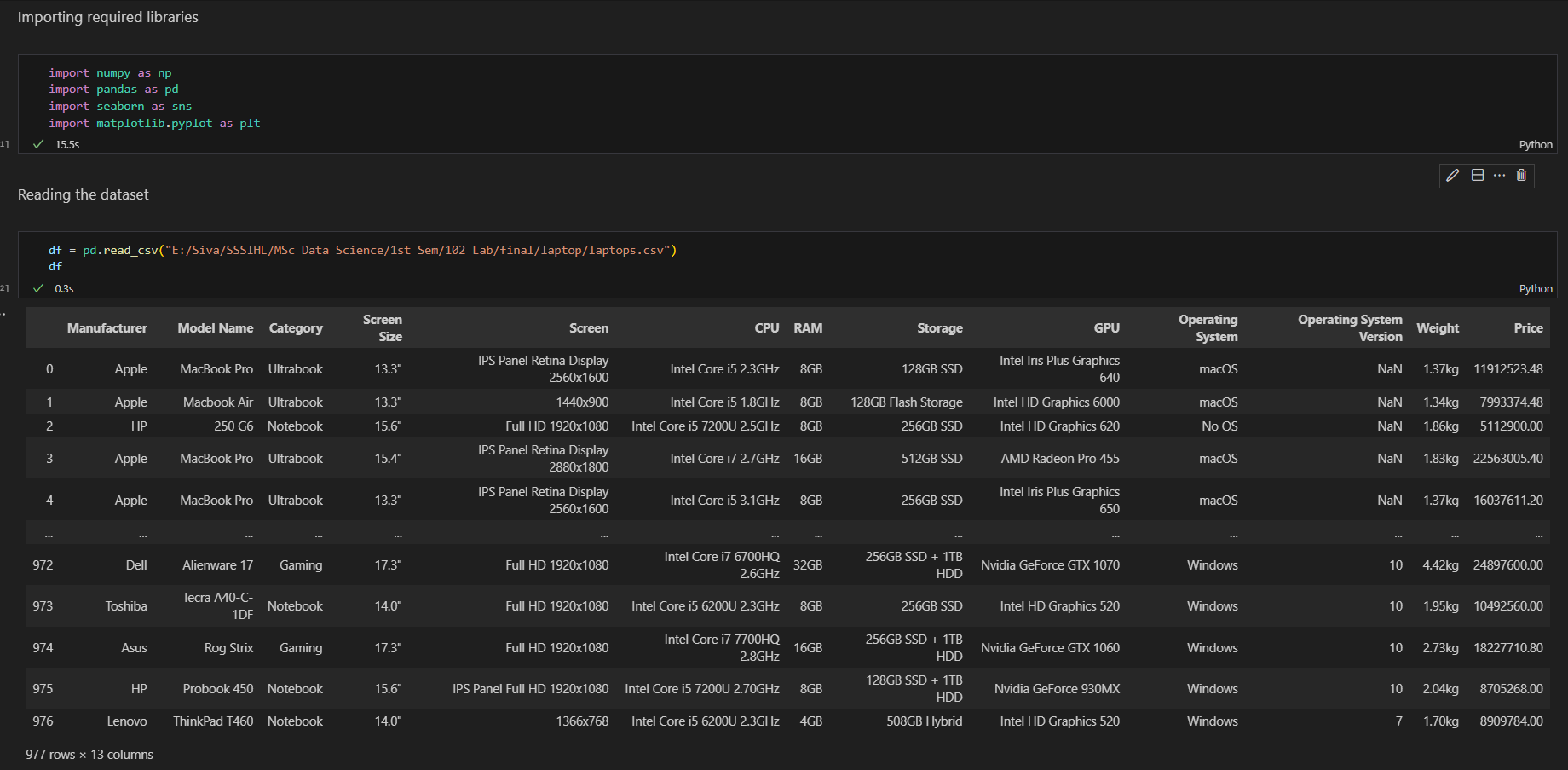
GPU – provides more information about GPU, for instance Nvidia GeForce GTX 1070

Operating System – provides information about the OS, for instance Windows

Operating System Version – gives OS Version information, for instance 10 i.e., Windows 10

Weight – gives the weight of the laptop in kg’s, for instance 1.83kg

Price – gives the price of laptop in lakhs, for instance 248900

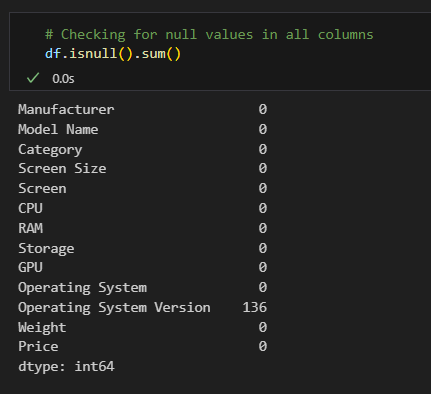


**Data Preprocessing**

By looking at the above data frame, it looks complex with their respective data types. So, let’s try to minimize the complexity and do data preprocessing to understand the dataframe with more ease.

We can achieve this by applying certain techniques for each of the attributes available.

First, let’s look at the missing values (null values) present in the dataframe.

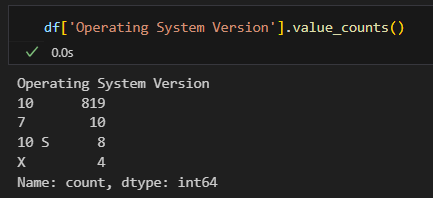


So, we can observe that there are no null values present in the dataframe except for Operating system version attribute.

Now, here it's really tough to handle the missing values as each product is of different manufacture, OS and depends on several other factors.

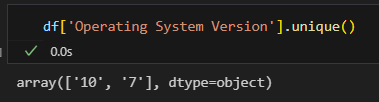
We might drop the missing values, but that may lead to data loss. So, to keep the data available, let’s do the following:

For Operating System version:

Looking at the value counts of OS version, we got to know that most of the laptops have version 10

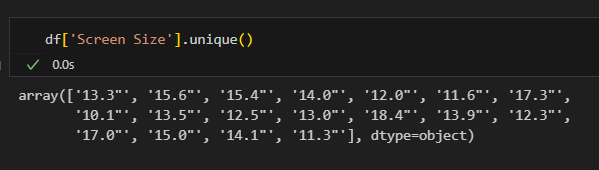
So, to keep the data available, and the most used OS Version is 10, so let's replace missing values with 10

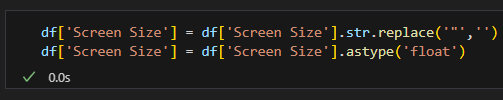
From above, we observe that roman language X is used, which is 10 and also 10 S is used. So, let’s convert both of them into 10 and now let’s check at the unique values present in the column.



Unique values in the column are 10 and 7.

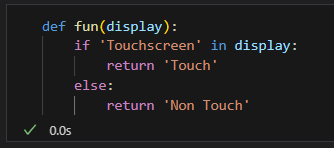
For Screen Size:

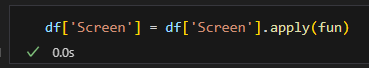
It’s in object type, we need to minimize this to int or float now.



For Screen:

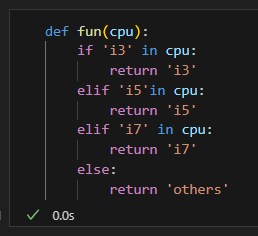
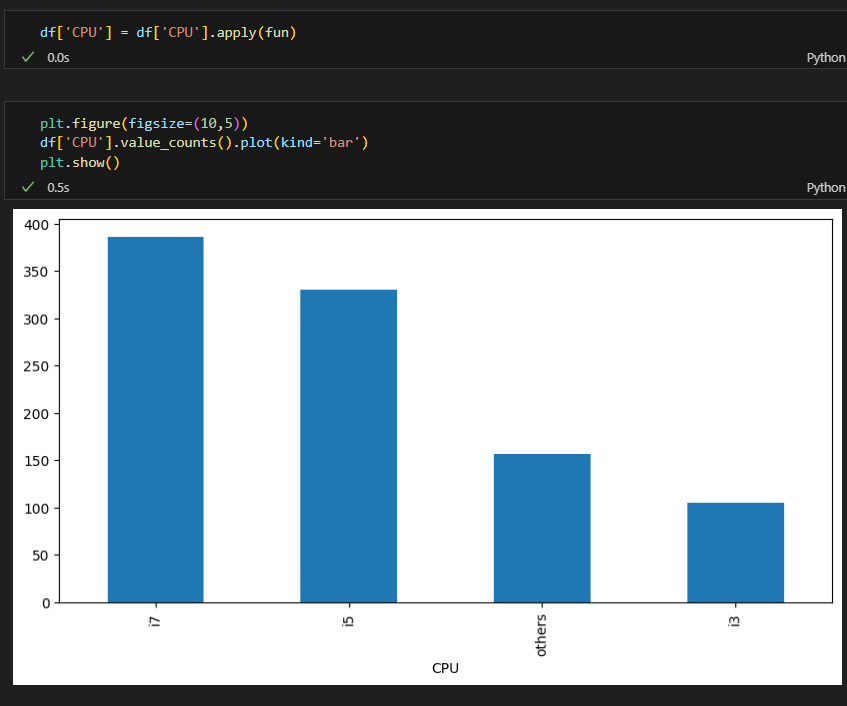
Upon checking the unique values, we can segment the values of Screen to smaller unique values basing on Touch Screen and Non-Touch Screen, so let’s write a function to achieve that.



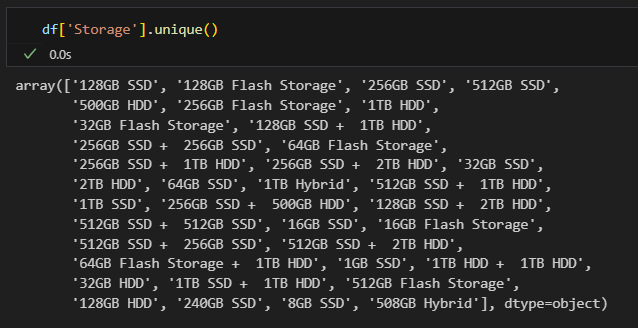


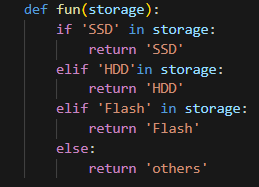
For CPU:

Upon checking the unique values, we can segment the values of CPU to smaller unique values such as i3, i5, i7 and others. Let’s write a function for it and plot a bar plot for it.

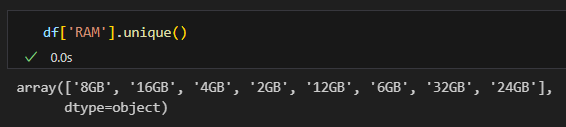
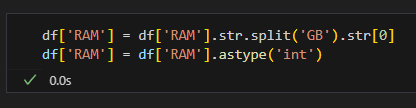
For Storage:

Let’s write a function to segment the values into unique values like SSD, HDD, Flash, and others.



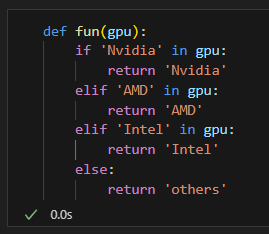
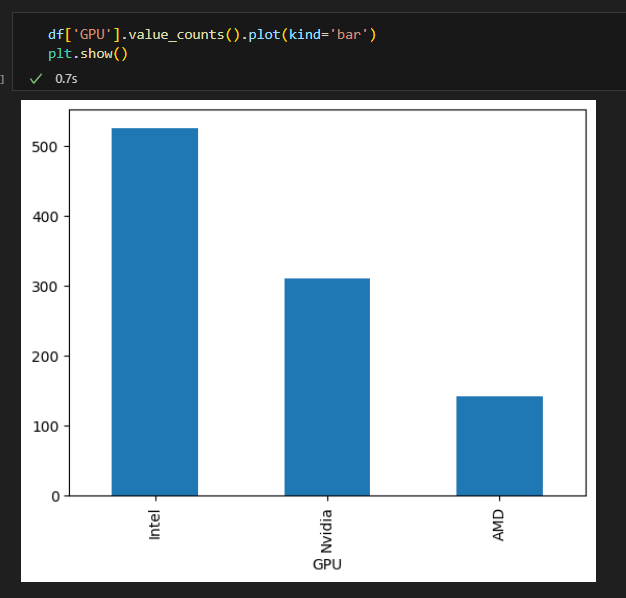
For RAM:

The unique values present are 2GB, 4GB, 8GB, 12GB, 16GB, 24GB so on. As we are aware that RAM values are of GB, let’s convert them into numerical and into integer data type.

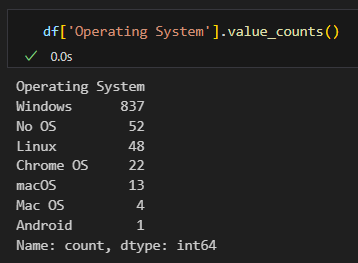
 

For GPU:

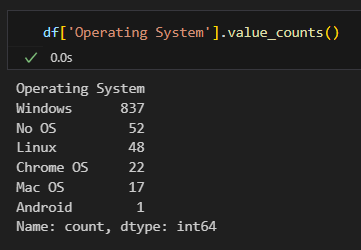
Upon checking the unique values, we can segment the values of GPU to smaller unique values such as Nvidia, AMD, Intel, and others. Let’s write a function for it and plot bar plot for it.

For Operating System:

As we can observe macOS and Mac OS, let's combine them both with Mac OS by replacing macOS with Mac OS.



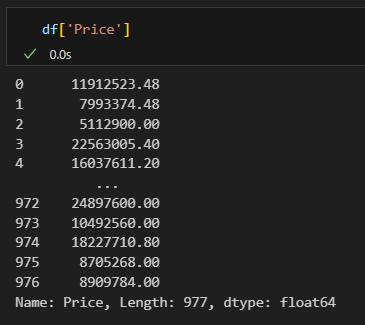


For Weight:

As we know that all laptops weigh in kg's only, so let's extract the numerical value & convert to float.

For Price:



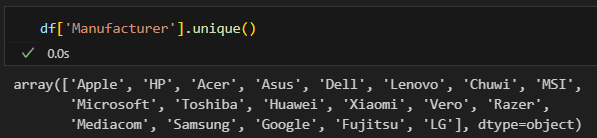
As most of the laptop prices will be in lakhs and not crores, so let’s get price in lakhs.

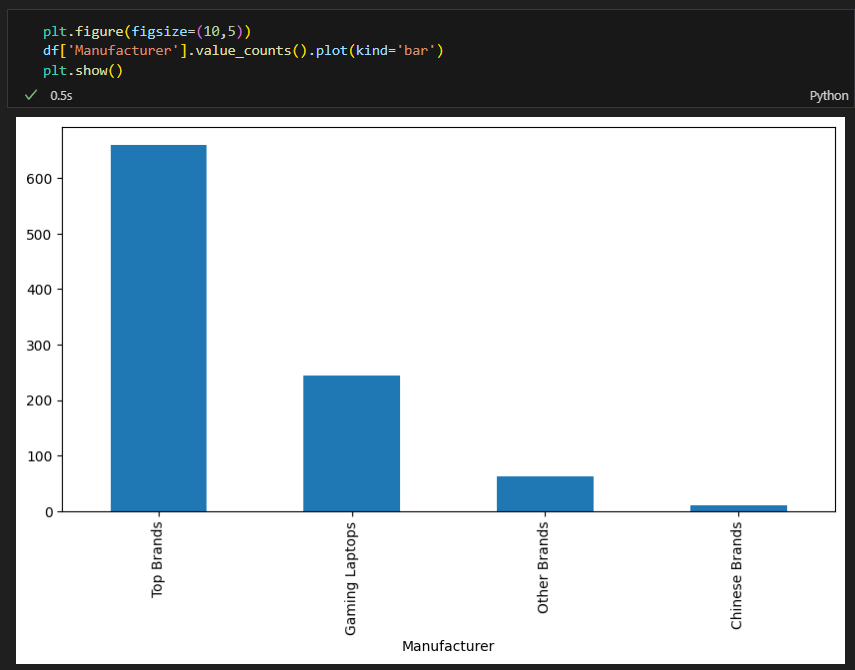
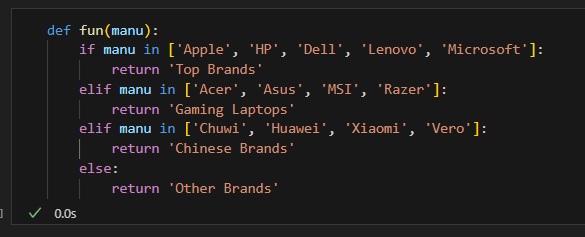


Now let’s convert it into integer data type.

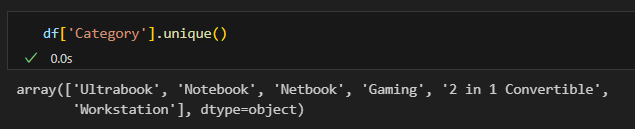


For Manufacturer:

We can observe many company names here, so let’s classify them as following using a function and plot it using bar plot.

For Category:

We need not perform any changes to the values present in Category.

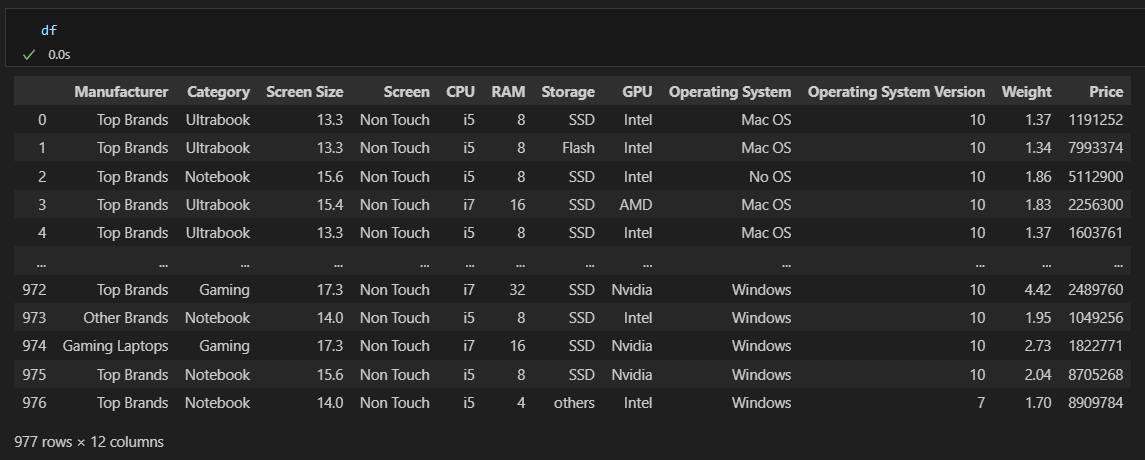
For Model Name:

Since we have all features required (manufacturer and category) and there are a greater number of unique values in Model Name and also this feature is of less importance when compared to others. So, upon this basis, we don’t need this column, so let’s drop it from the dataframe.

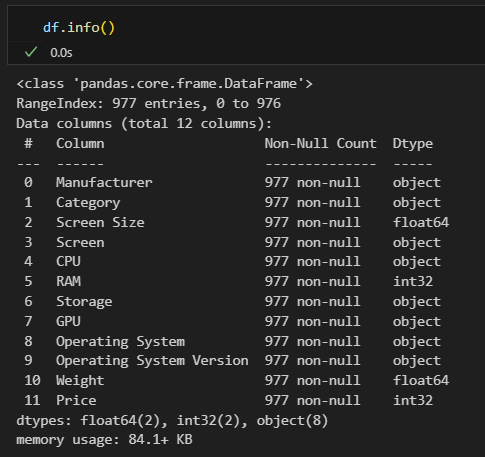


So, all the columns have been minimized for better understanding and converted to optimal data types, let’s look at the dataframe and the data types of the columns.

***Dataframe:***



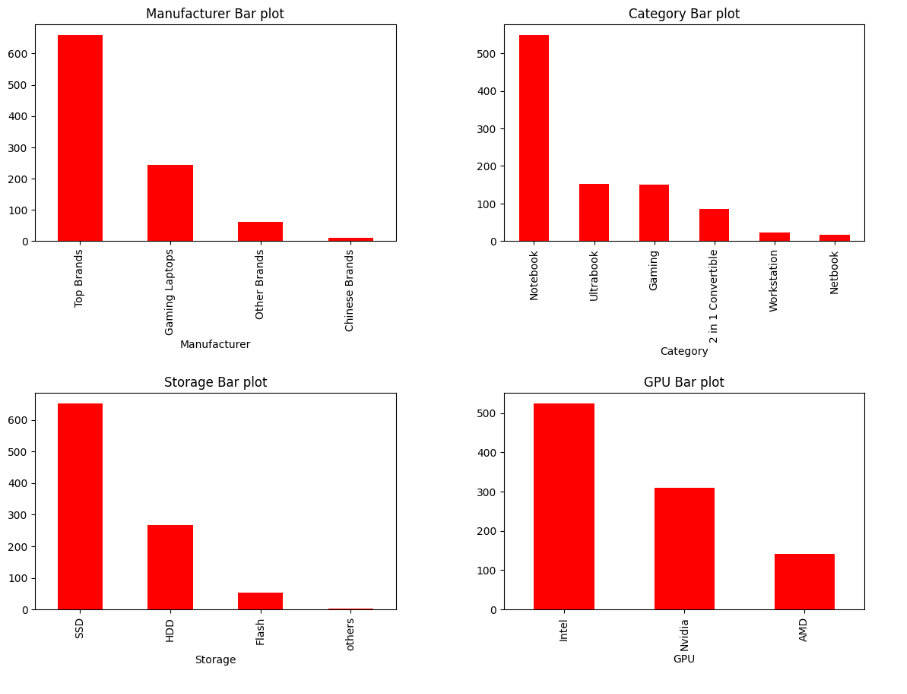
Data types of the columns:



Now, as we are done with preprocessing, let’s perform Exploratory Data Analysis using certain visualizations like bar plot, histogram, pie chart, box plot, scatter plot etc.

**Exploratory Data Analysis**

***Bar Plot*** for all the categorical attributes(object) along with their inferences.

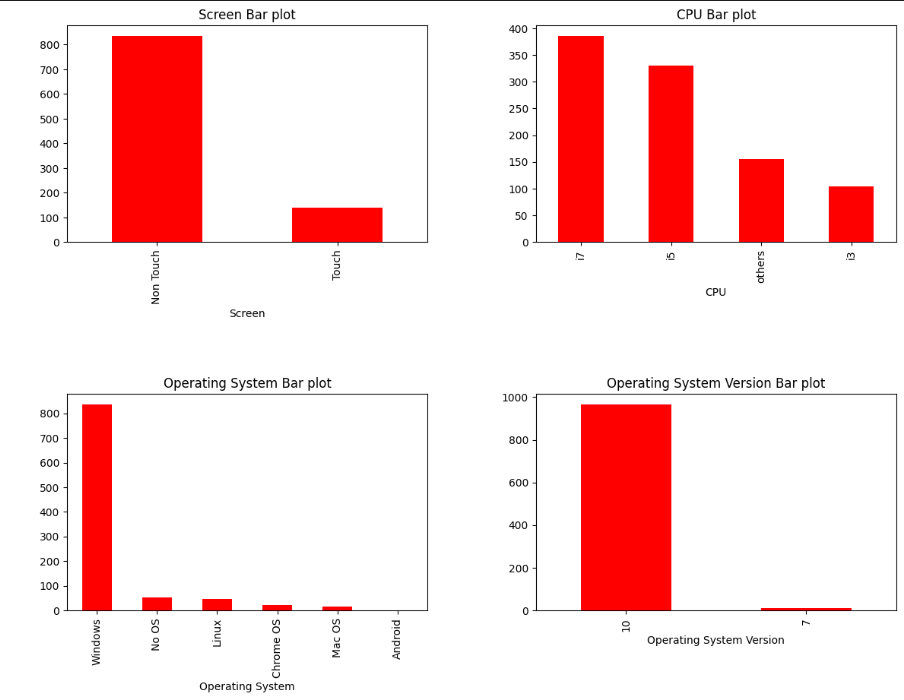


From the above snippet, we can observe that currently in the market, manufacturing of the top brand laptops (Apple, HP, Dell, Lenovo) are more than the gaming laptops or the other brands.

Also, more people are preferring notebooks with Intel processor which has SSD storage.

We can also infer that less population is going for AMD when compared with Nvidia, whereas when we compare Nvidia and Intel, more people are going for Intel Processor.

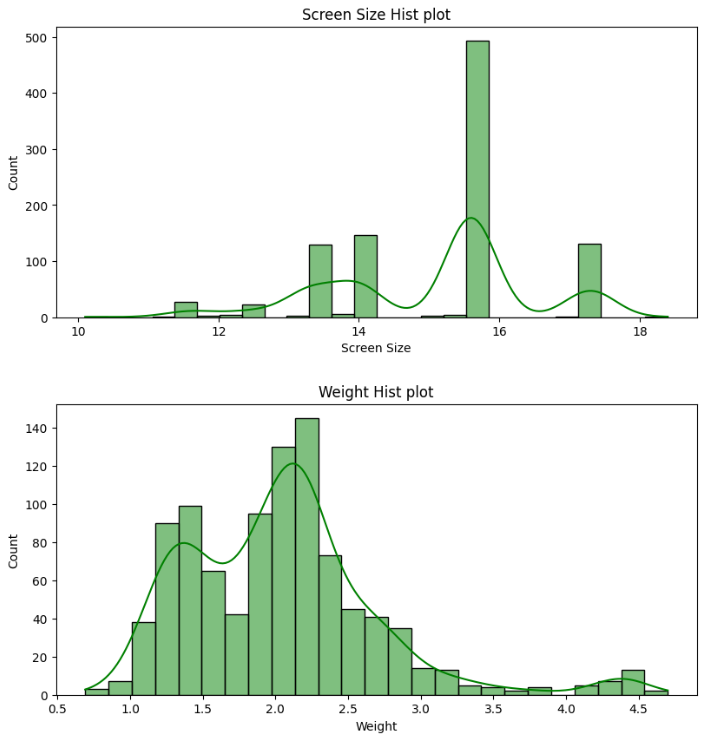
Day by day, the usage of Flash as storage is being reduced, and good amount of people are preferring for external storage like HDD, but most of the people are opting for SSD.



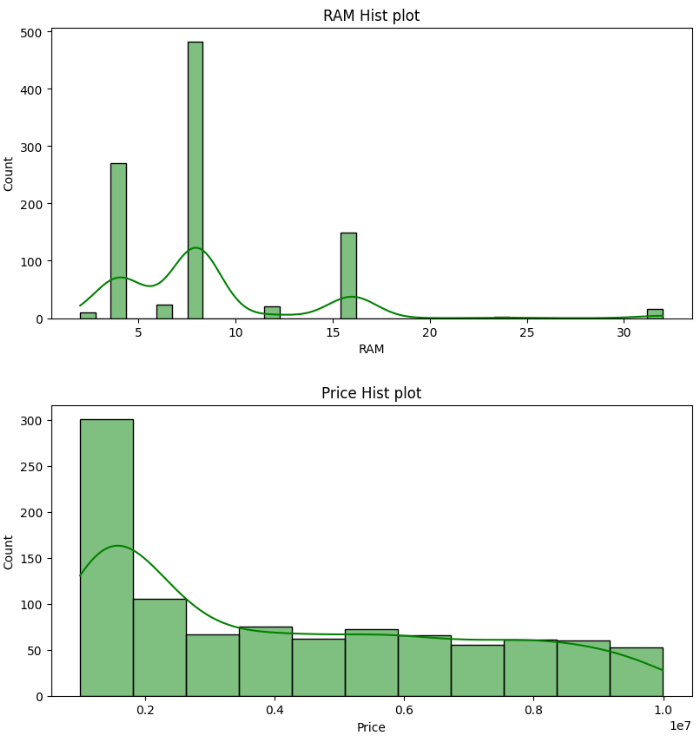
From the above snippet, we can observe that currently in the market, more preference of laptop screen is for non-touch when compared to touch screen laptops and with i7 processor as the latest trend and promising processor when compared to others. At start i3 use to be promising, as technology advancements taken place, i5 emerged and now its i7era and the i9 will be in the near future. This will continue gradually as new promising processor comes in to market.

To talk about the popular operating system and its version, as we all are aware and from the above bar plot is true that Windows is the leading trader in Operating systems with Mac Os in the second place and the stable and promising version of Operating System Version of Windows is 10 today. As days passes, new versions will be coming into the picture, and this will be carried forward.

***Histogram*** (Hist Plot) for all the continuous attributes (int and float) along with their inferences.



From the above snippet, it’s clear that the Screen Size of laptops is from 10 to 19 approximately in today’s market. And most consumers are preferring 15inches laptop as their top preference where people who require minimal screen are going for 13inches laptop and people who prefer larger screen size are preferring around 17inches laptop and the weight of the laptop is ranging from 1kg to 5kg, where most customers are preferring ideal weight around 2.3kgs.

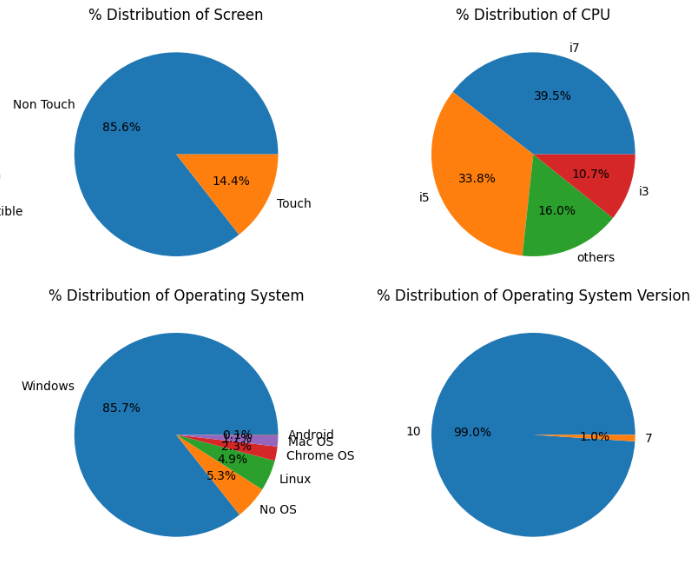


From the above snippet, we can infer that RAM Capacity in the laptop generally ranges from 4 to 32, and we can see that few customers who have brought early has RAM storage around 4GB, and it moved on to 8GB where currently most users are using and few prefer 8GB and extra 4GB in the second slot to make it 12, and others who need high performance for instance gaming, these category of people go for 16GB of RAM, also for people if necessary they go for 32 too.

Most of the buyers prefer budget friendly laptop for general usage, and the people who require for a specific usage, will go for a high-end laptop where it costs more.

***Pie-chart*** for all the categorical attributes:

As we have discussed earlier, we can draw similar inferences but now in percentage for each of the item using a pie chart.

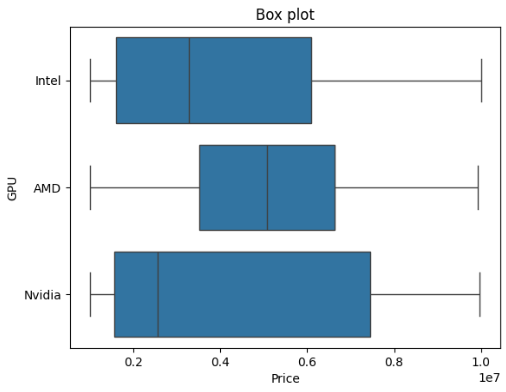


Here, the percentage distribution of screen is 85.6% Non-Touch Screen and the other 14.6% who go for Touch Screen.

Many buyers are preferring i7 than all other CPU Processors, where they go for Windows Operating System of version10.

Coming to manufacturing, more top brands laptops are getting manufactured than any other laptops where more notebooks are getting manufactured for daily usage with SSD storage with Intel Processor.

***Box Plot*** for few attributes w.r.t price and their inferences:



From this box plot, we can observe that generally laptops with Intel GPU starts from low price to mid-price, with more products of comparatively less price, and AMD maintains certain standards with it’s price ranges at a good cost, whereas Nvidia starts with it basic variants to the high end one’s.

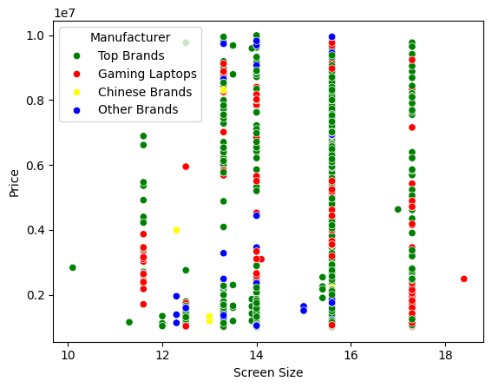
From this box plot, the major observation is w.r.t the Chinese brands where the cost generally ranges low, but there are a few with high prices too.

***Scatter Plot*** for few attributes w.r.t price and their inferences:

Scatter Plot is done for two continuous variables by taking one on the x-axis and the other on y-axis based on other variable (hue).

This scatter plot is based on Weight and Price of the laptops based on hue Screen.

We can infer that, generally the weight of the touch screen laptop generally ranges low, as they are used for commercial purpose for instance office etc. And Non touch laptops are distributed across many weights and price ranges

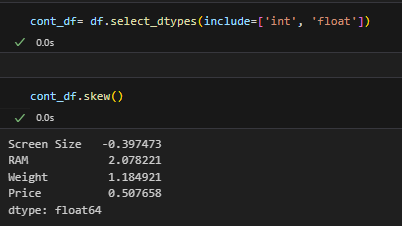


This scatter plot is based on Screen Size and Price of the laptops based on hue Screen.

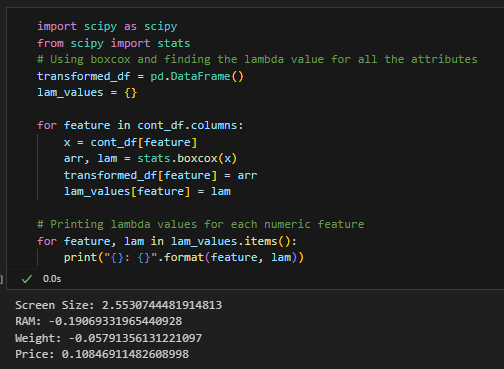
We can infer that, most of the top brands laptops Screen Size is nearly 15inches, but there also equivalently many distributed between 11,13 and 17inches.

Now, let’s try to perform some **statistical tests**, but before doing that let’s select the continuous attributes and normalize them using boxcox transformation.

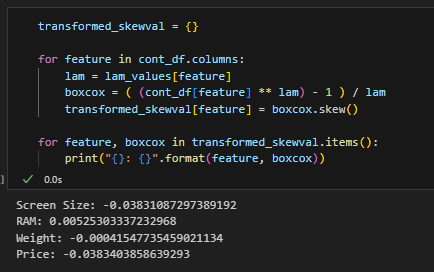
Before performing boxcox transformation, let’s check the skewness.



Here, we can see the skewness of each continuous attribute. Now, we should try to find the lambda values and then apply the boxcox transformation.

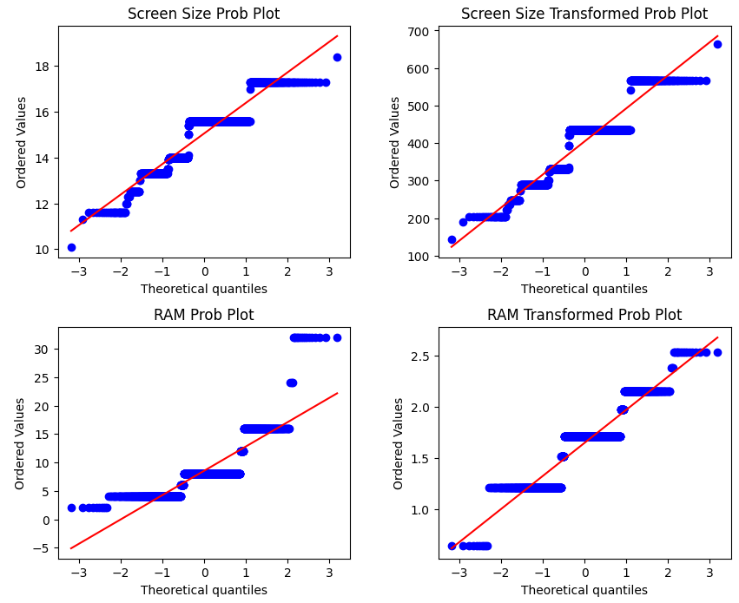


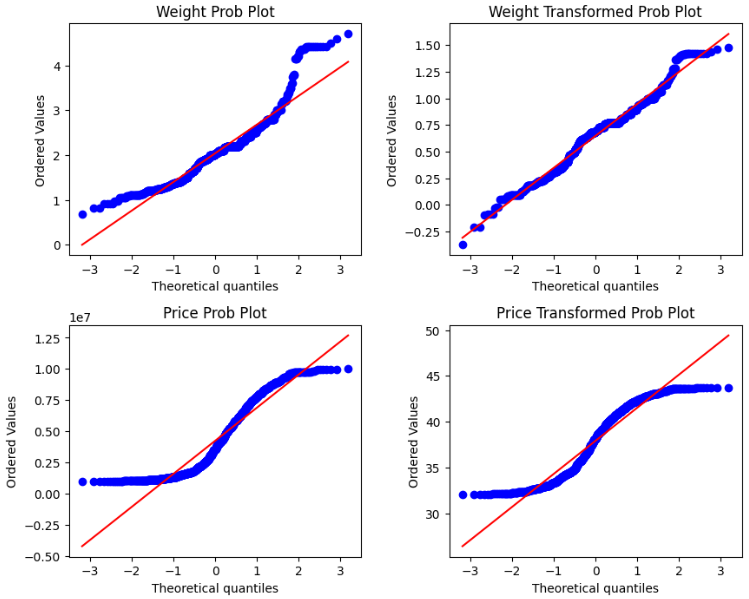
We got the lambda values, now let’s perform boxcox transformation to find the required values.



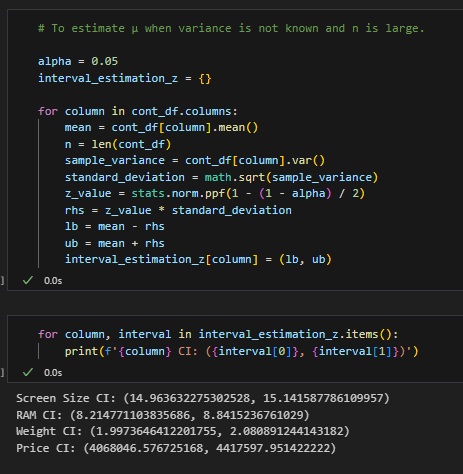
Normalized values

Let’s look at the ***probability plot*** below to see the difference before and after transformation.

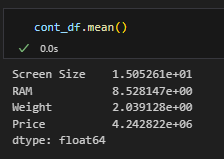


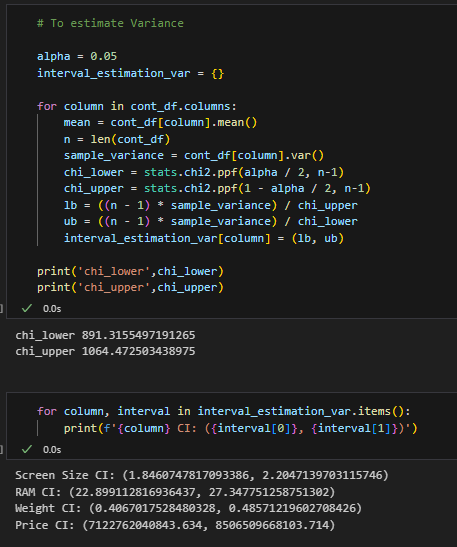


Here, we can clearly see the difference between before and after transformation of each attribute with its probability plot.

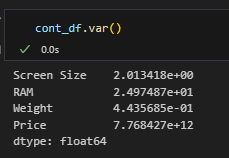


Now, let’s find the interval in which the mean ranges for each of the attribute. And we cross check it with mean of each attribute.

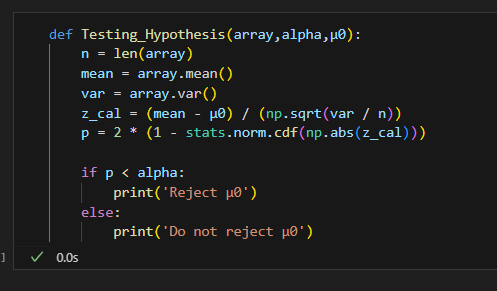




Now, let’s find the interval in which the variance ranges for each of the attribute and display the chi square values (lower & upper). And we cross check it with variance of each attribute.

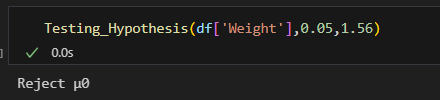
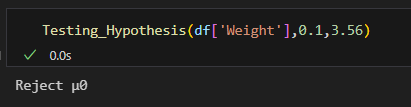


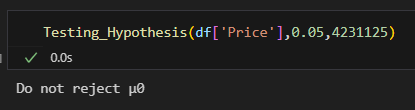
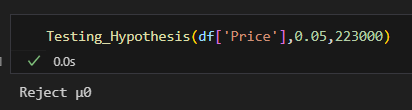
Now, let’s write a function for the following ***testing hypothesis***:

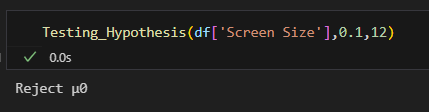
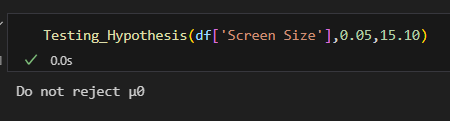


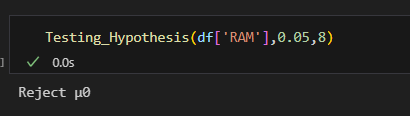
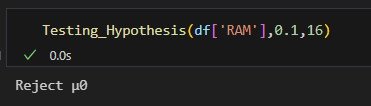
This function takes the array as the input where we will pass the values of an attribute with level of significance alpha and with µ0

Now, let’s perform testing hypothesis for the following:



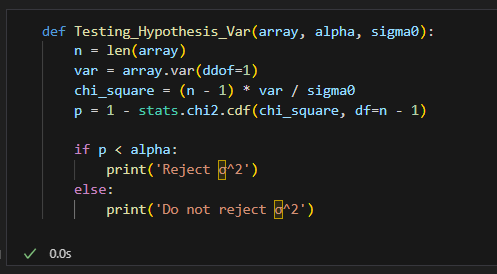
 

Now, let’s write a function for the following ***testing hypothesis***:





This function takes the array as the input where we will pass the values of an attribute with level of significance alpha and with σ0

Now, let’s perform testing hypothesis for the following:

