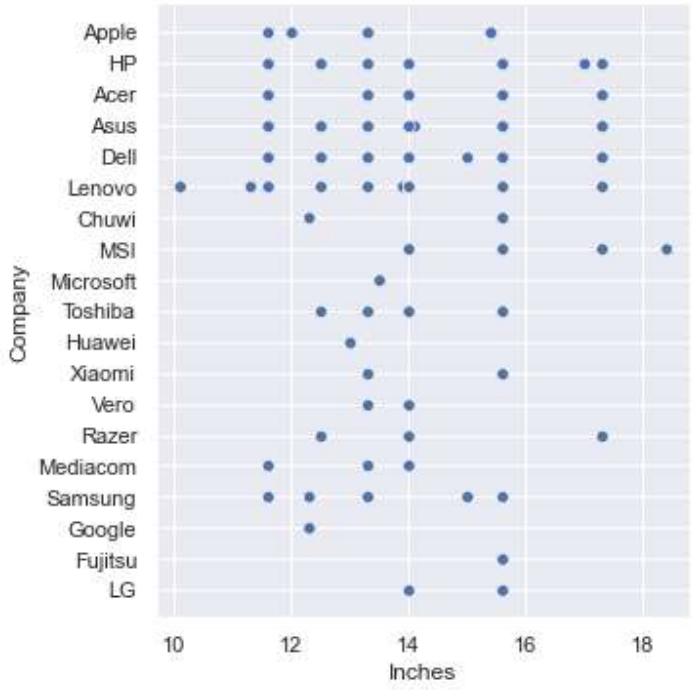
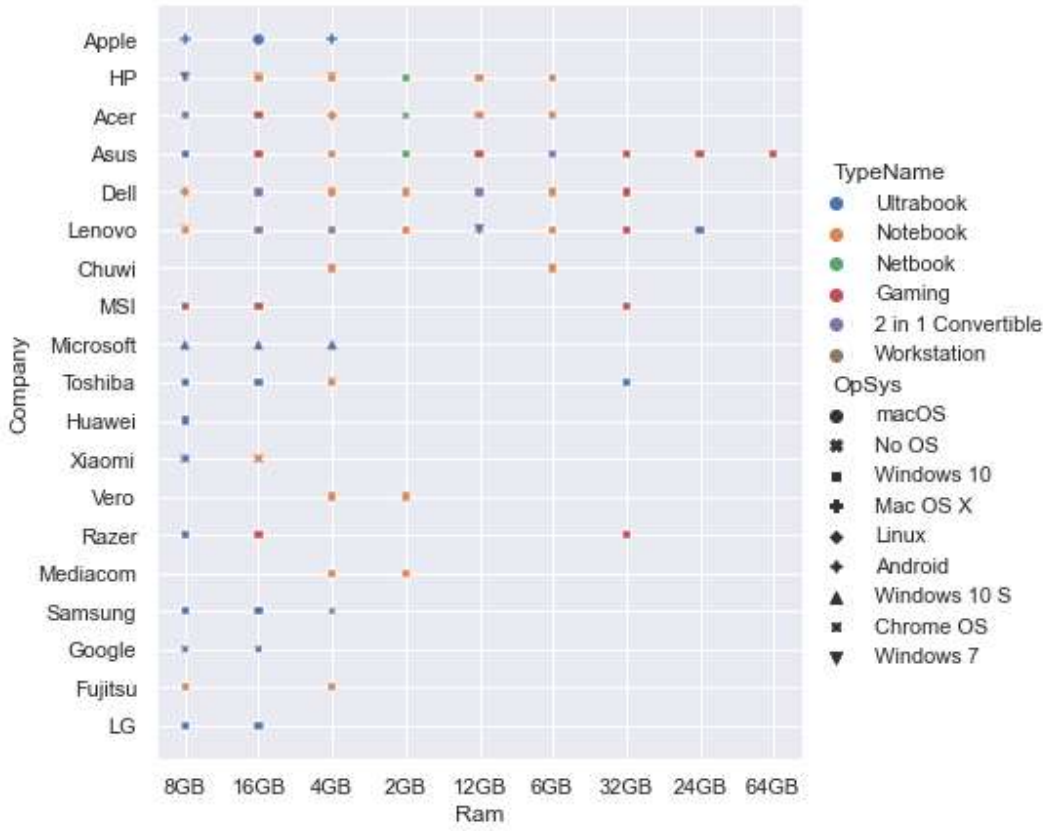
ASSIGNMENT - 2

HITESH S
21BDA47

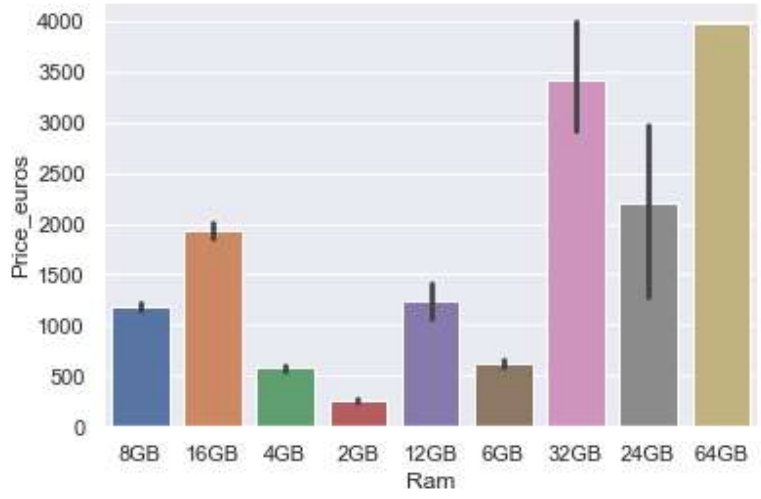
1.Document 5-6 key insights from EDA and support each point with a visualization.



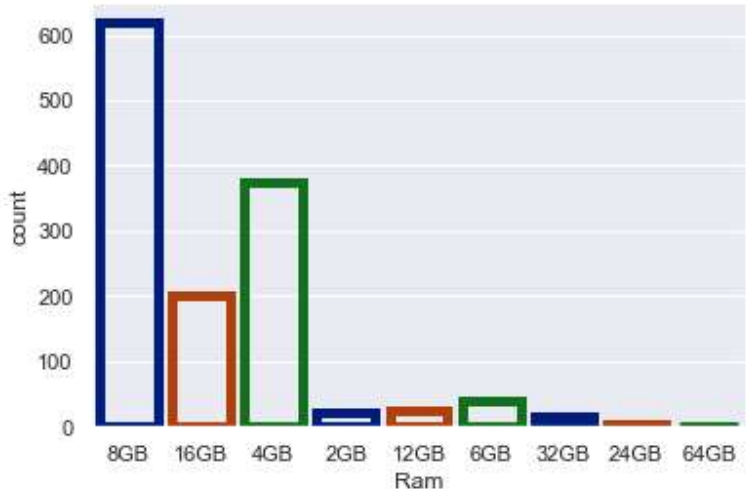
Here we are looking at Company and inches so we can MSI is the company which produces the highest 18 inch of display for the sales and Lenovo produces the lowest display of 10 inch.



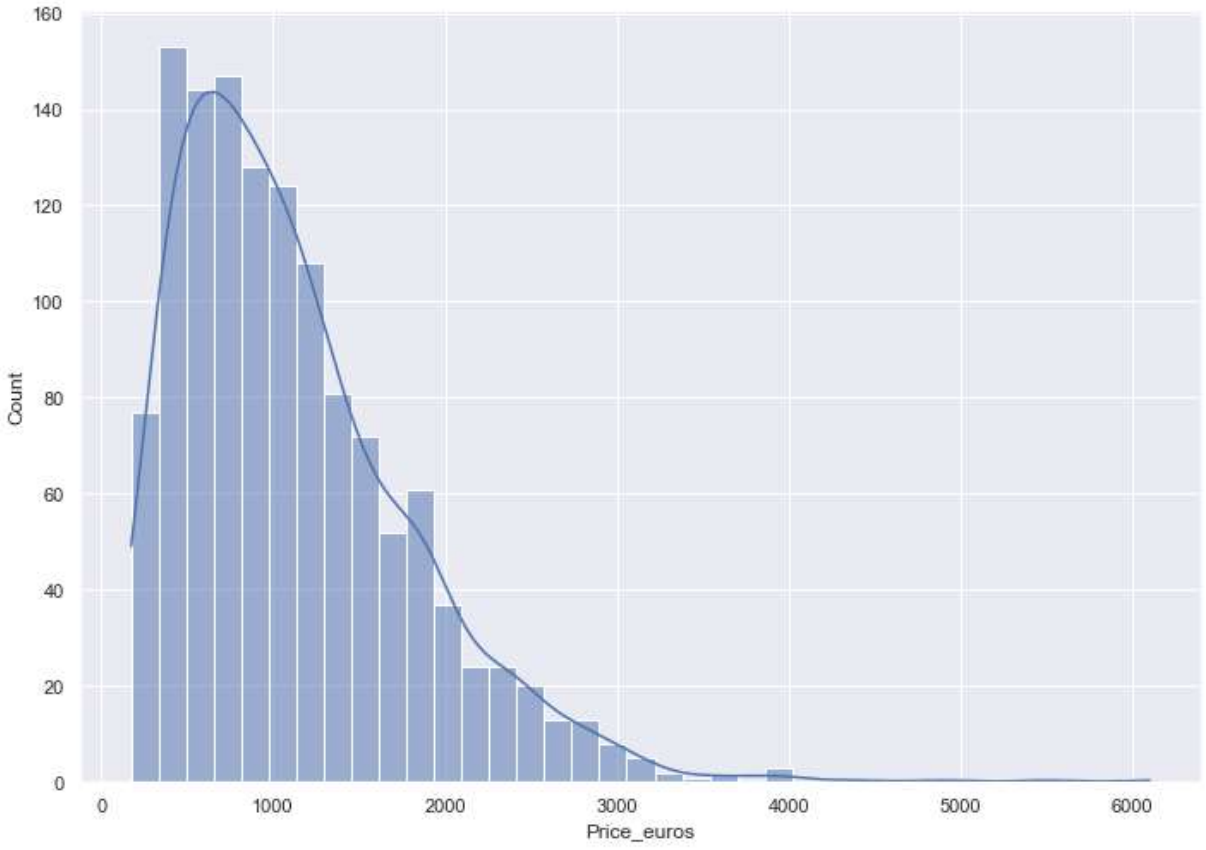
If we look at both ends of the RAM spectrum, we can see that Asus and a few other firms offer gaming laptops with larger RAM capacities, while the majority of companies offer laptops with lower RAM capacities, such as two in one convertibles and Ultrabooks along with which types of operating system is there in particular producers.



As the above bar plot clearly says, as the RAM increases, the price also increases due to RAM increases, the storage capacity increases and the laptop works smoothly without any trouble.



In this box plot we can see that 8GB RAM users are more compared to any other RAM, and we can see that 64GB RAM has least count due to Price affecting factor.



So in this histogram we see that as the Price increases the buyers count are very less. This might be a factor of not being able to afford it so companies should try to focus on the 500-1500 price range.

2. Answer the following questions:

i. What are the assumptions of linear regression?

Linearity: refers to the relationship between X and the mean of Y.
Homoscedasticity: For every value of X, the variance of the residual is the same.
Independent observations: Observations are not reliant on one another.
Normality: Y is normally distributed for any fixed value of X.

ii. How can we evaluate a Regression model? Define each metric and its interpretation.

R Square/R Square Adjusted
R Square is a metric for how well a model can explain variability in a dependent variable. It's called R Square since it's the square of the Correlation Coefficient(R).

MSE (Mean Square Error)/Root Error in the Mean Square (RMSE)
While R Square is a relative measure of the model's ability to fit dependent variables, Mean Square Error is an absolute measure of the fit's quality.

Mean Absolute Error (MAE)
Is a measure of how accurate a (MAE) The Mean Absolute Error (MAE) is compared to the Mean Square Error (MSE) (MSE). MAE, on the other hand, takes the total of the absolute value of error rather than the sum of squares of error like MSE does.

iii. Can R squared be negative?

For equations that do not contain a constant term, a negative R-square is feasible. Because R-square is defined as the proportion of variation explained by the fit, it is negative if the fit is worse than fitting a horizontal line.

iv. What is a dummy variable trap?

The Dummy variable trap occurs when multiple qualities are highly correlated (Multicollinear), and one predicts the value of others. When categorical data is handled using one-hot encoding, one dummy variable (attribute) can be predicted using other dummy variables. As a result, one dummy variable is

substantially associated with the others. When all dummy variables are used in regression models, a dummy variable trap occurs.

v. Is one hot encoding different from Dummy Variables?

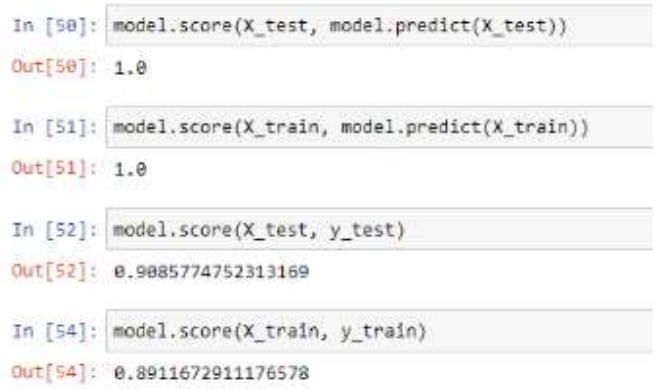
Assume that a category variable has n values. With one-hot encoding, it's transformed into n variables, while with dummy encoding, it's translated into n-1 variables. If there are k categorical variables, each having n values, Dummy encoding generates kn-k variables, whereas hot encoding generates kn variables.

vi. How is polynomial regression different from linear regression?

Polynomial regression is a type of linear regression that estimates the connection as an nth degree polynomial. It is a specific example of multiple linear regression.

Linear regression is a supervised learning machine learning algorithm. It carries out a regression task. Based on independent variables, regression models a goal prediction value.

vii. Interpret the screenshot below from the notebook we discussed in class today:



There is a model accuracy of roughly 0.9085 among the test data, indicating that the model is extremely accurate with the given test data. Test data often makes up a small portion of the data collected in order to conduct tests and acquire reliable findings.

There is a model accuracy of 0.8911 among the trained data, indicating that the model is also extremely accurate using the given trained data. The majority of the data used to execute particular tests and acquire results is usually trained data.

The training dataset is used to teach a machine learning programme to recognise patterns or perform to your criteria, whereas the testing or validation dataset is used to assess the accuracy of your model.

> viii. Bonus: We saw Sweetviz as an Automated EDA option. What are the other options? Try a few of them and share which one you find the best.

Dtale, pandas profiling, sweetviz, autoviz are the four types of automated EDA.

AutoViz is another favorite of mine. With just one line of code, AutoViz can automatically visualize any dataset. Only those automatically picked features can be used by AutoViz to determine the most relevant features and create effective visuals. AutoViz is also extremely fast, thus visualizations are created in a matter of seconds.