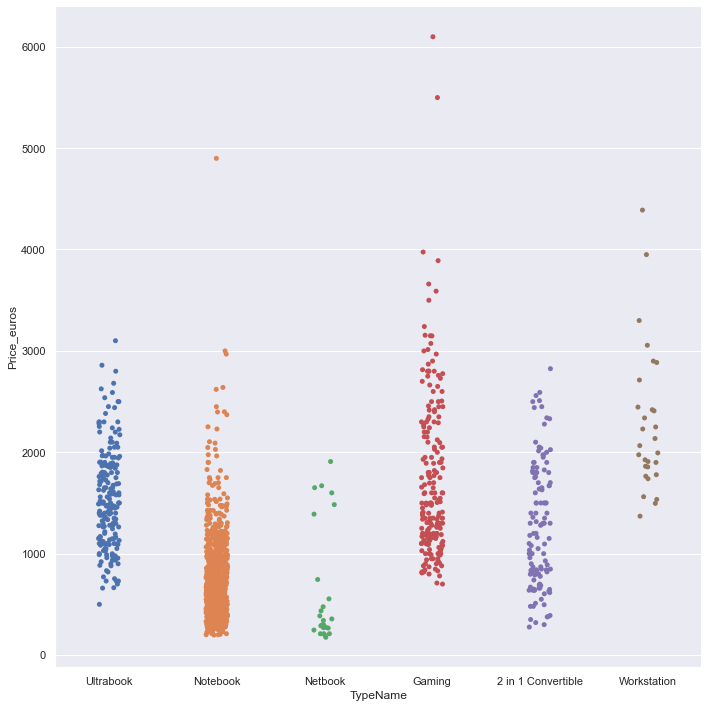
**ASSIGNMENT – 2**

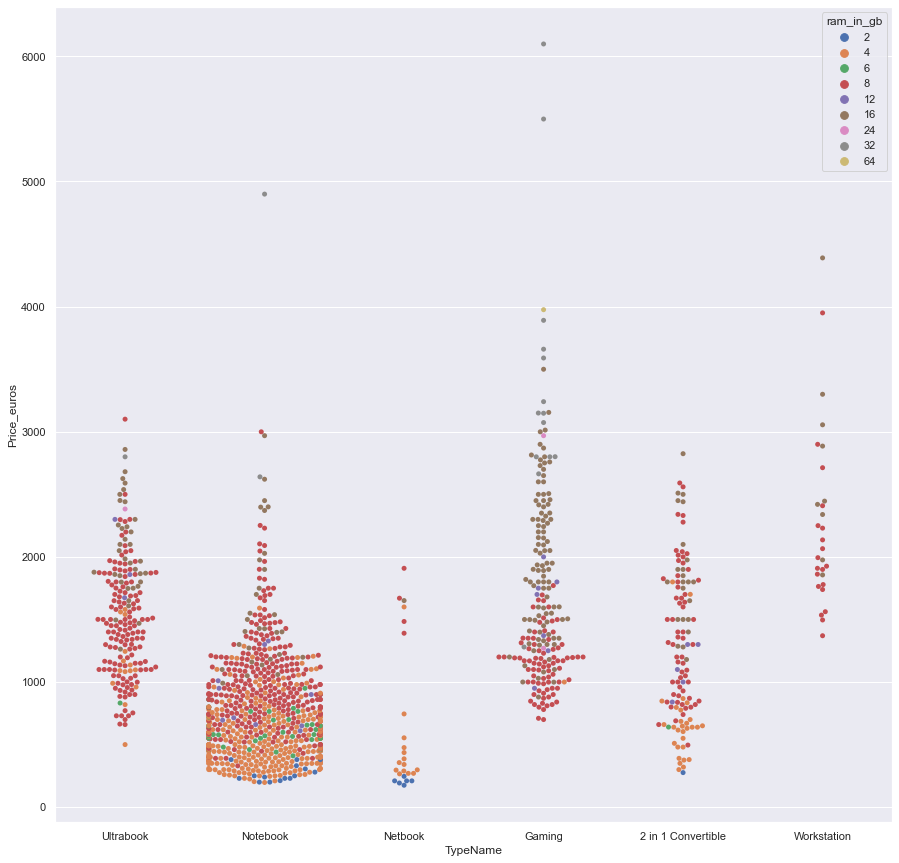
1. Document 5-6 key insights from EDA and support each point with a visualization.



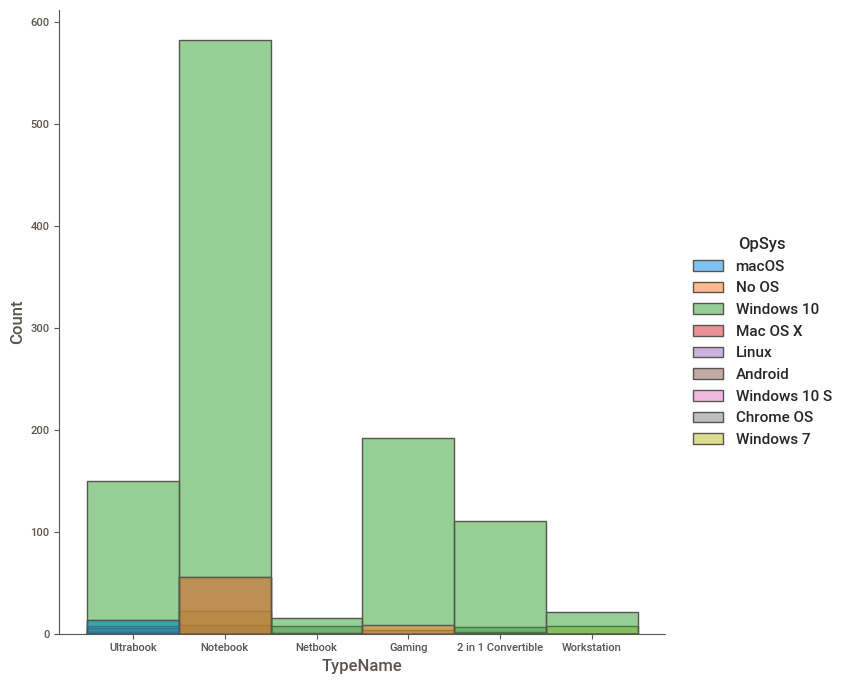
There is a positive relationship between Price and Memory.



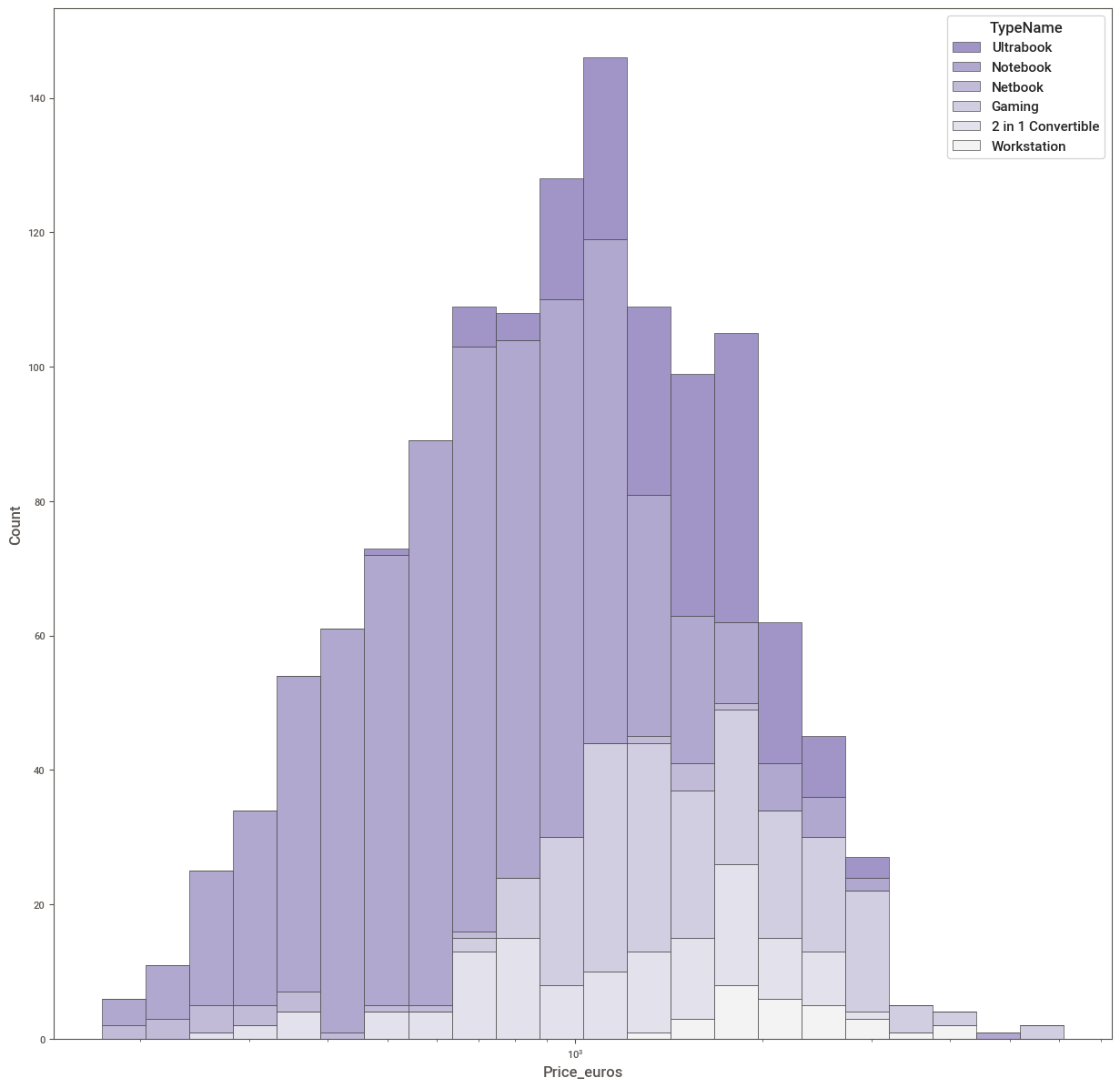
The Starting Price of the Workstation Laptops looks the highest. We can see that there are outliers in Gaming and Notebook Computers.



This Plot shows that most laptops have 8gb ram. With the increase in ram, price also increases.



Windows 10 is the most popular Operating System.



Notebook laptops are the most in demand and extend over a range of prices.

Netbook has the least listings.

Gaming Laptops also extend over a range of prices.

1. What are the assumptions of linear regression?

There are four assumptions associated with a linear regression model:

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed.

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

Evaluation metrics are a measure of how good a model performs and how well it approximates the relationship.

## There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square

2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)

3. Mean Absolute Error(MAE)

## Mean Squared Error (MSE)

The most common metric for regression tasks is MSE. It has a convex shape. It is the average of the squared difference between the predicted and actual value. Since it is differentiable and has a convex shape, it is easier to optimize.

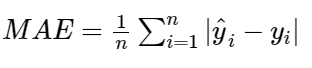


Mean squared error. Image by the author.

MSE penalizes large errors.

## Mean Absolute Error (MAE)

This is simply the average of the absolute difference between the target value and the value predicted by the model. Not preferred in cases where outliers are prominent.



Mean absolute error. Image by the author.

MAE does not penalize large errors.

## R-squared or Coefficient of Determination

This metric represents the part of the variance of the dependent variable explained by the independent variables of the model. It measures the strength of the relationship between your model and the dependent variable.

To understand what R-square really represents let us consider the following case where we measure the error of the model with and without the knowledge of the independent variables.

1. Can R squared be negative?

it is possible to get a negative R-square for equations that do not contain a constant term. Because R-square is defined as the proportion of variance explained by the fit, if the fit is actually worse than just fitting a horizontal line then R-square is negative. In this case, R-square cannot be interpreted as the square of a correlation. Such situations indicate that a constant term should be added to the model.

1. What is dummy variable trap?

The Dummy variable trap is a scenario where there are attributes that are highly correlated (Multicollinear) and one variable predicts the value of others. When we useone-hot encodingfor handling the categorical data, then one dummy variable (attribute) can be predicted with the help of other dummy variables. Hence, one dummy variable is highly correlated with other dummy variables. Using all dummy variables for regression models leads to a dummy variable trap. So, the regression models should be designed to exclude one dummy variable.

1. Is One Hot Encoding different from Dummy Variables?

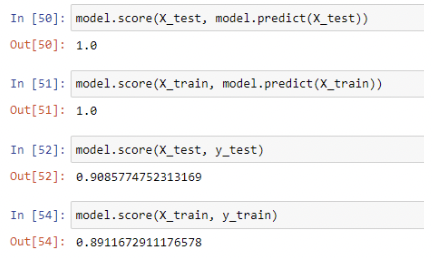
One-hot encoding and dummy encoding are different ways of encoding the categorical data.

If we have a categorical variable that has k different categories then one-hot encoding will convert it into k variables whereas dummy encoding will convert it into k-1 variables.

1. How is polynomial regression different from linear regression?

Simple linear regression has only one x and one y variable. Polynomial Regression is a one of the types of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E (y |x).

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



The .score method is used to check the accuracy.

The first line shows accuracy between the X\_test values and its predicted values, the value 1 means accurate predictions.

The Second line of code shows accuracy between the independent variables that will be used to train the dataset and the predicted ones.

The function used to calculate accuracy against the text data and training data.

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

The other options are dtale, dataprep and pandas profiling

Dataprep was the one that I found best.