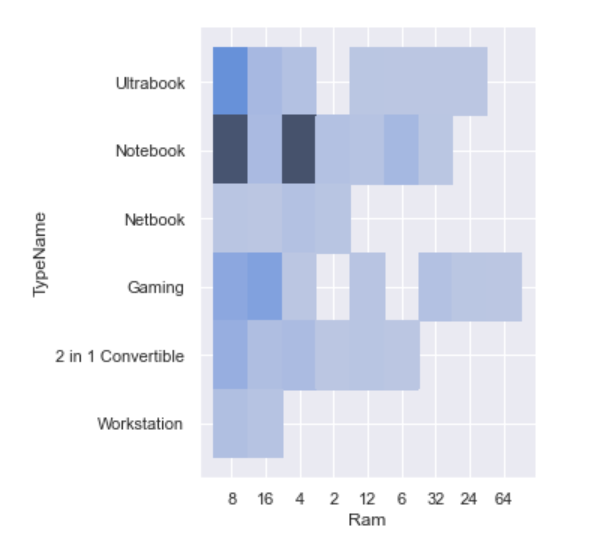
## **PART B - Exploratory Data Analysis**

**Conclusion 1 -**

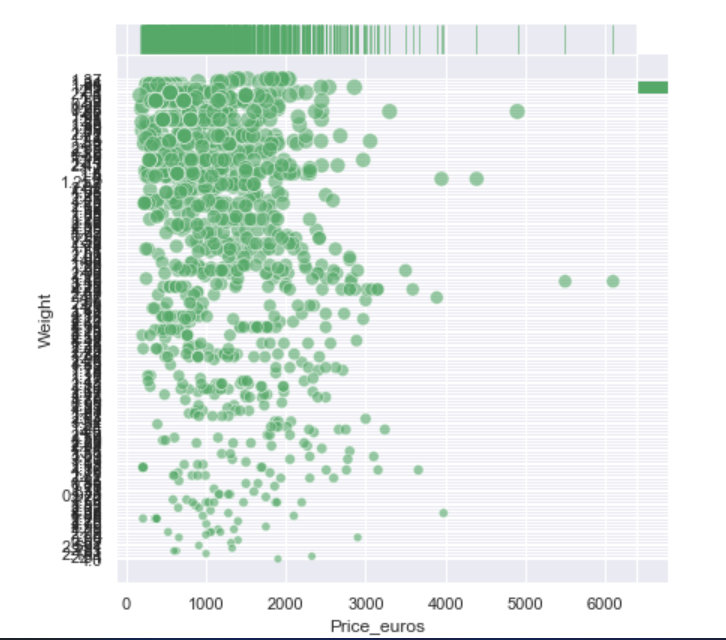
1) We can observe from the graph that the higher ram which is 32-64GB , they are mostly found in Gaming laptop.

2) Laptop having 8 and 4 GB Ram are mostly present in Notebook Type Laptop

****

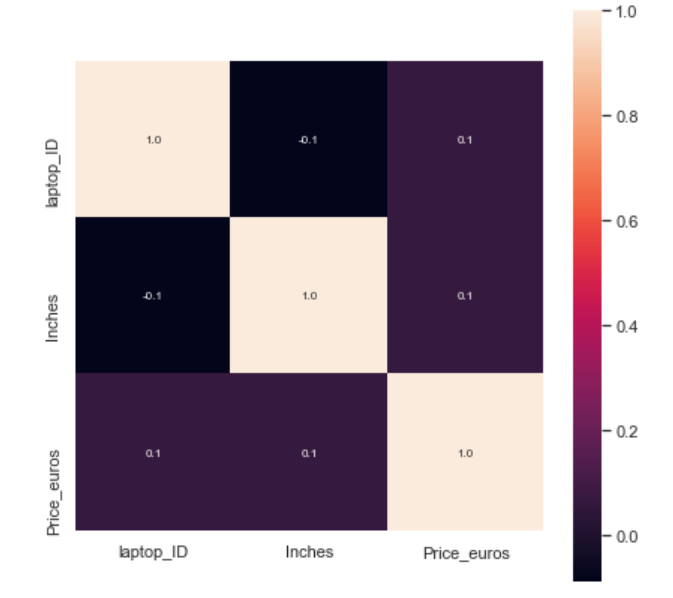
**Conclusion 2 -**

If the price is more then weight of laptop is less , vice versa we can see that laptops having higher weight are cheaper in price.

* ****

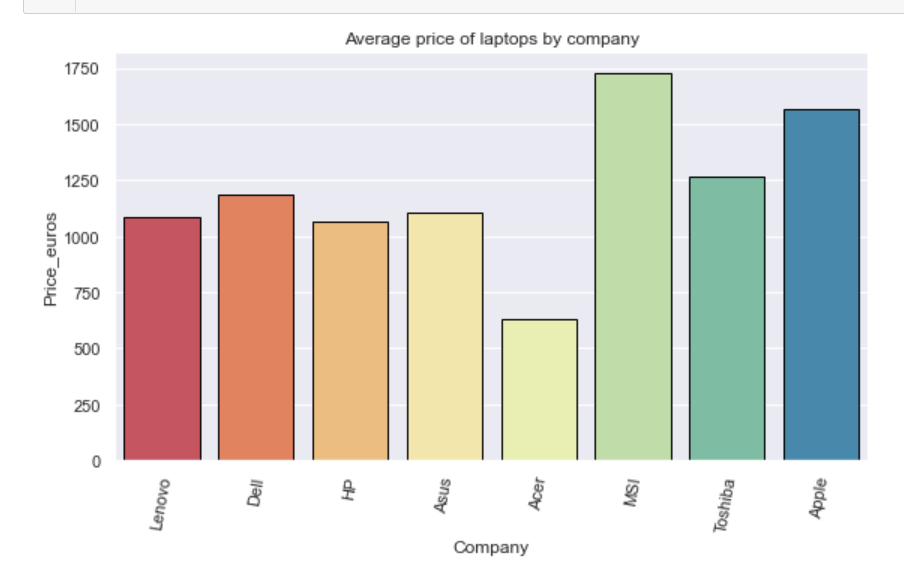
**Conclusion 3 -**

1. Heat Map here shows the correlation between different numerical varibale.
2. Graph here depects that correlation between inches and price is 0.1 which is negligible

****

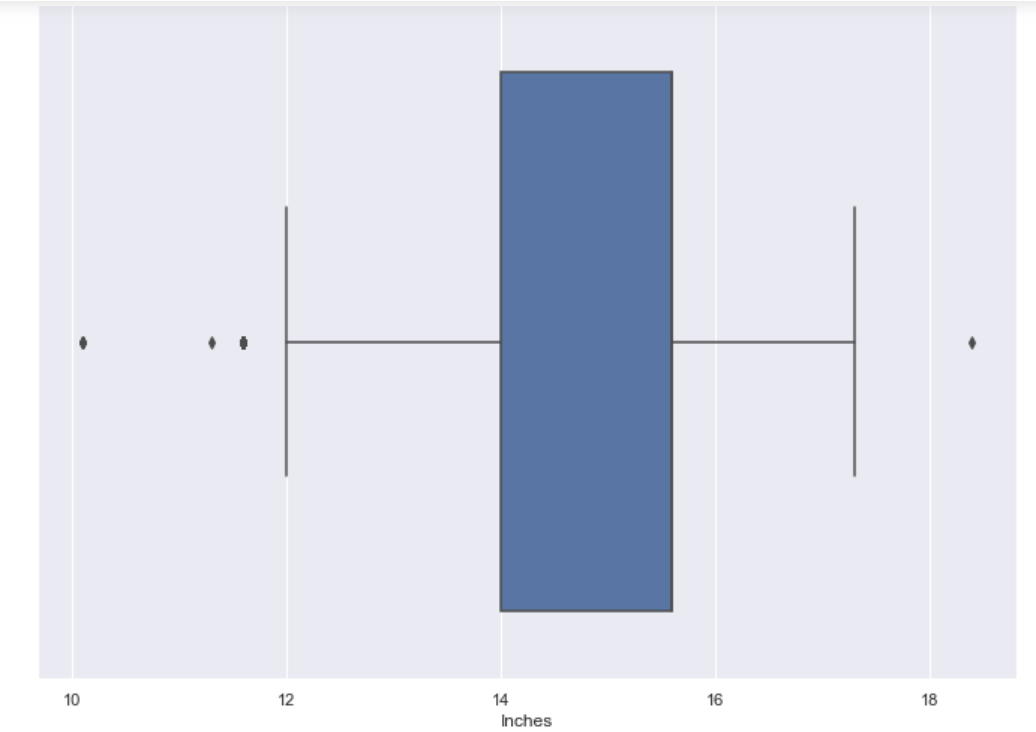
**Conclusion 4 -**

1. Graph here shows that cost of Acer Company laptops are lowest with the price being 600 euros (approx)
2. MSI is most expensive with the cost around 1750 euros (approx)

****

**Conclusion 5 -**

Average Inches of Laptops are 15.5 inches (approx)

****

**Question 1 -**

**Assumption of Linear Regression are -**

**1. Linear relationship:** There exists a linear relationship between the independent variable, x, and the dependent variable, y.

**2. Independence:** The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

**3. Homoscedasticity:** The residuals have constant variance at every level of x.

**4. Normality:** The residuals of the model are normally distributed.

**Question 2 -**

**Parameters -**

1. **R Square/Adjusted R Square**

R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

R Square is a good measure to determine how well the model fits the dependent variables. However, it does not take into consideration of overfitting problem. If your regression model has many independent variables, because the model is too complicated, it may fit very well to the training data but performs badly for testing data. That is why Adjusted R Square is introduced because it will penalize additional independent variables added to the model and adjust the metric to prevent overfitting issues

## **Mean Square Error(MSE)/Root Mean Square Error(RMSE)**

While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

MSE gives you an absolute number on how much your predicted results deviate from the actual number. You cannot interpret many insights from one single result but it gives you a real number to compare against other model results and help you select the best regression model.

Root Mean Square Error(RMSE) is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily. Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and makes it easier for interpretation.

1. **Mean Absolute Error(MAE)**

Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error. Compare to MSE or RMSE, MAE is a more direct representation of sum of error terms. MSE gives larger penalization to big prediction error by square it while MAE treats all errors the same.

**Question 3-**

The coefficient of determination, R2, is used to analyze how differences in one variable can be explained by a difference in a second variable. Yes , 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

**Question 4 -**

In linear regression models, to create a model that can infer relationship between features (having categorical data) and the outcome, we use the dummy variable technique. Dummy Variable is an artificial variable created to represent an attribute with two or more distinct categories/levels. The dummy variable trap is a scenario in which the independent variables become multicollinear after addition of dummy variables.

**Question 5 -**

**Difference** -

One**-hot encoding converts into n variables**, while dummy encoding converts it into n-1 variables. If we have k categorical variables, each of which has n values. One hot encoding ends up with kn variables, while dummy encoding ends up with kn-k variables.

**One-hot encoding**

In one-hot encoding, we create a new set of dummy (binary) variables that is equal to the number of categories (k) in the variable. For example, let’s say we have a categorical variable *Color* with three categories called “Red”, “Green” and “Blue”, we need to use three dummy variables to encode this variable using one-hot encoding. A dummy (binary) variable just takes the value 0 or 1 to indicate the exclusion or inclusion of a category.

**Dummy encoding**

Dummy encoding also uses dummy (binary) variables. Instead of creating a number of dummy variables that is equal to the number of categories (k) in the variable, dummy encoding uses k-1 dummy variables. To encode the same Color variable with three categories using the dummy encoding, we need to use only two dummy variables.

**Question 6 -**

Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial regression.

**Question 7 -**

The coefficient of determination is the square of the correlation(r), if R2 is equal to 1, then the dependent variable can be predicted from the independent variable without any error. The predicted values of X\_test are very close to the real values of X\_test, that’s why we got value 1 for coefficient of determination.