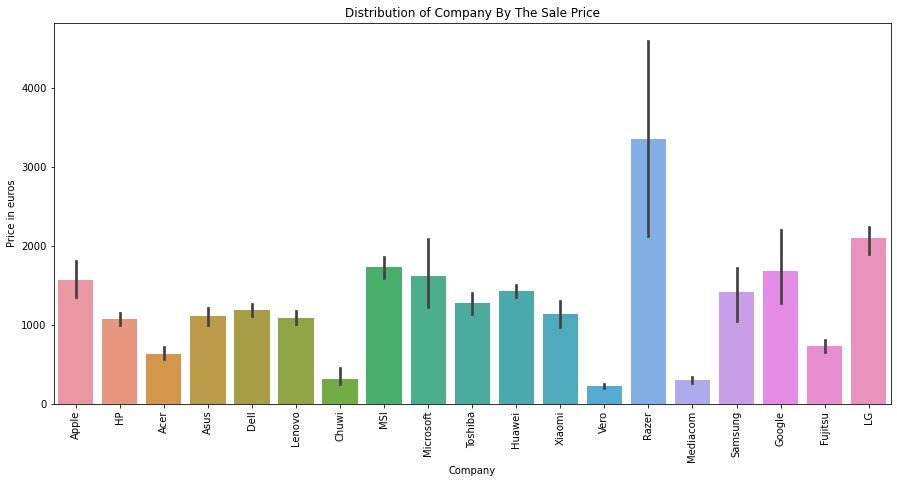
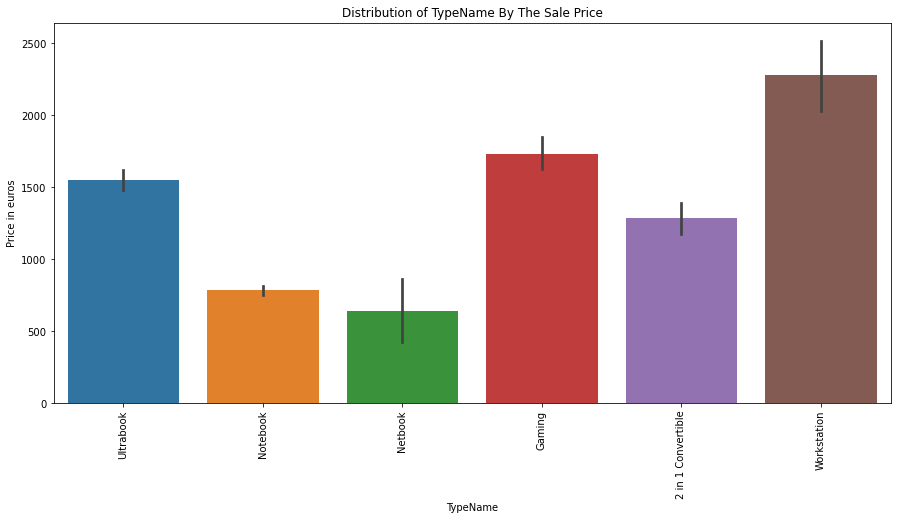
Question -2

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

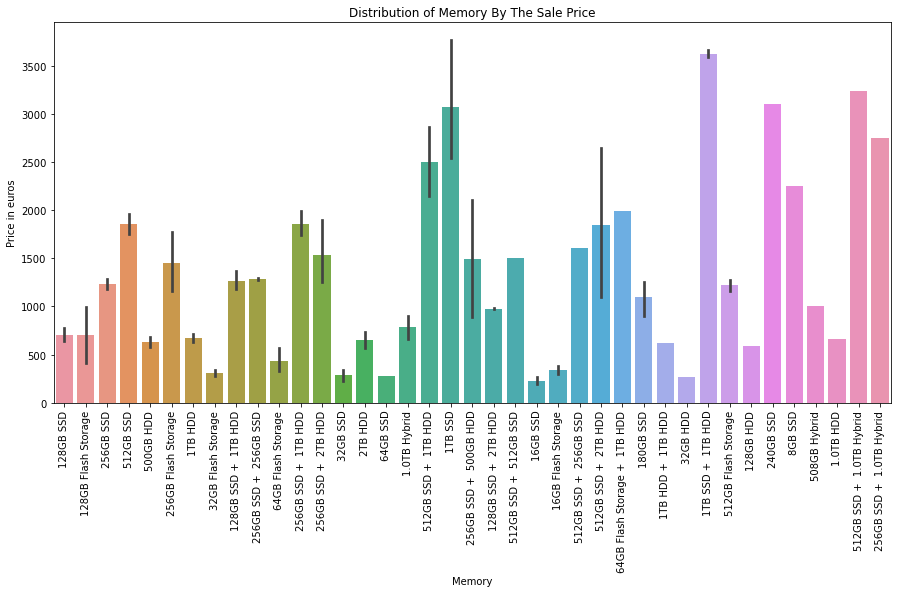
Razer products are costlier than other ones

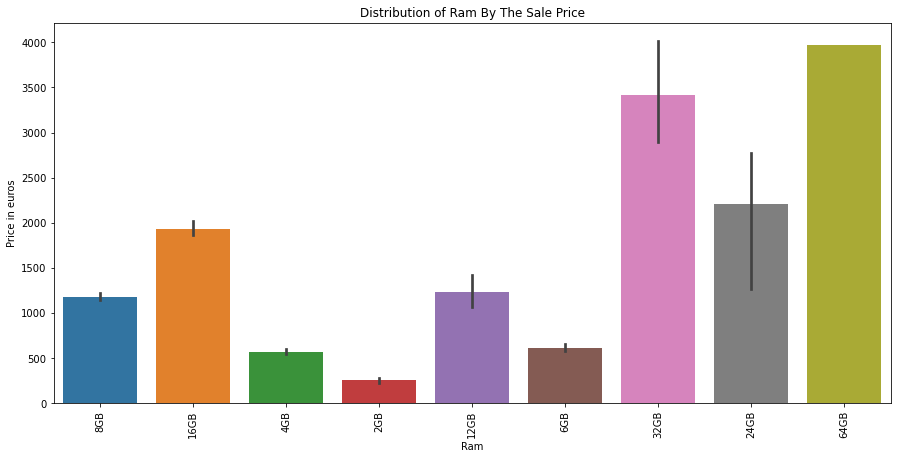


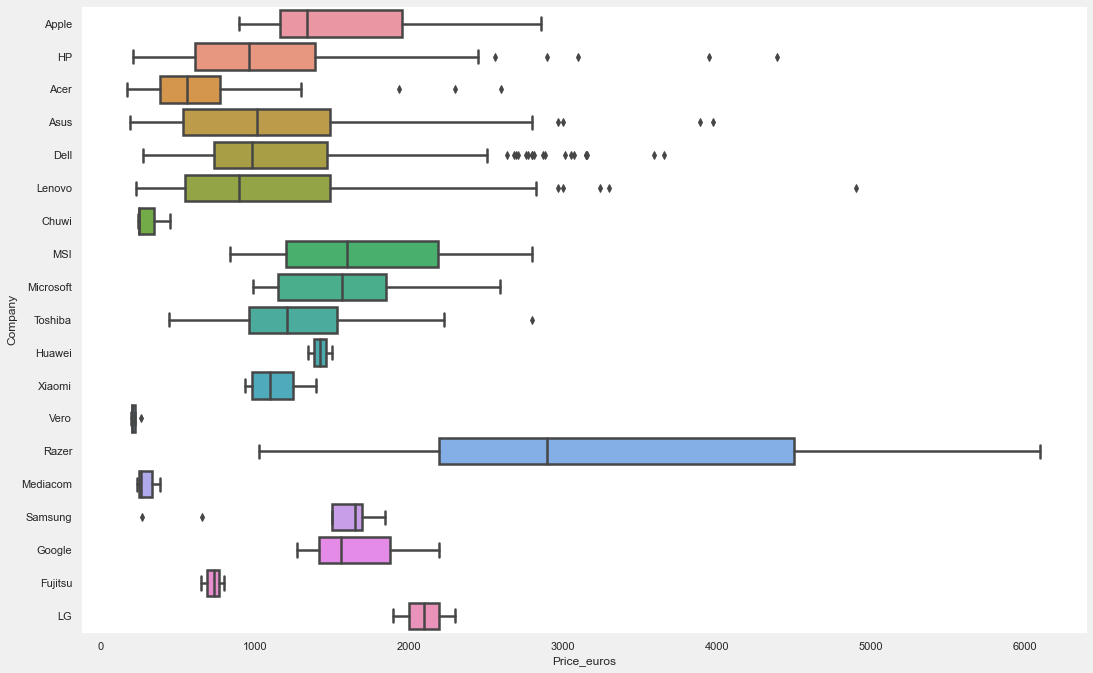
Workstation laptops tend to be costly

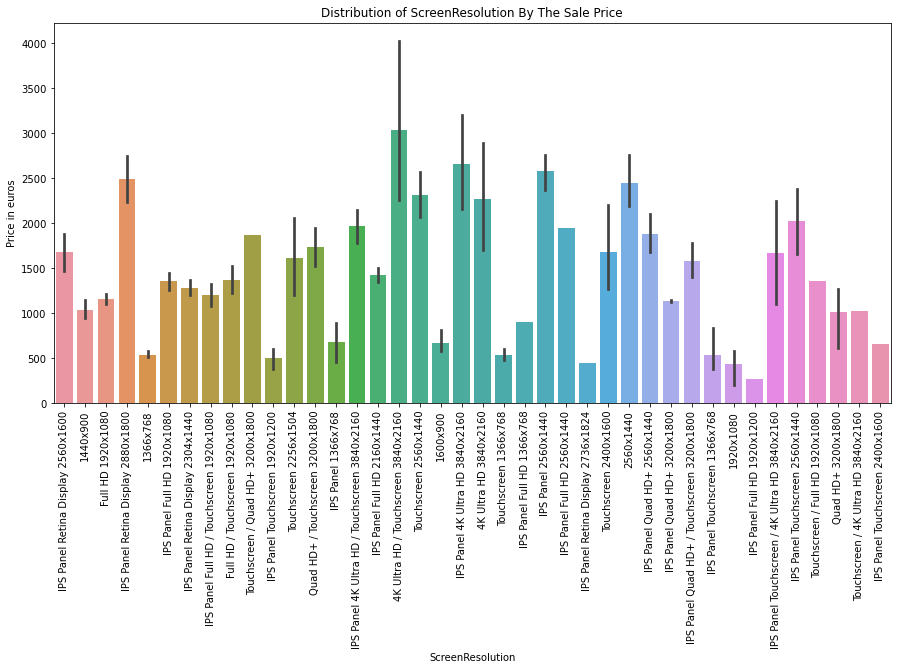


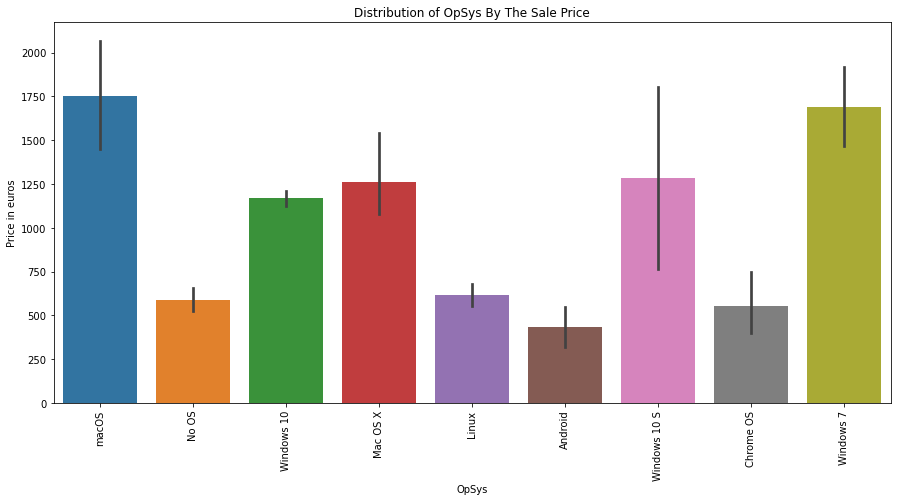
Price is directly proportional to storage



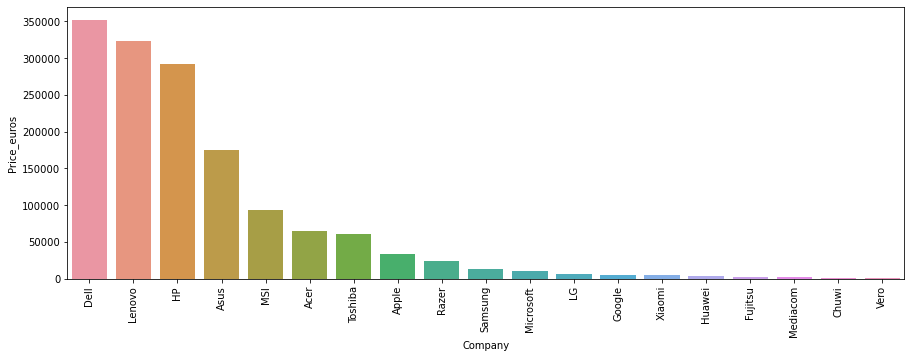






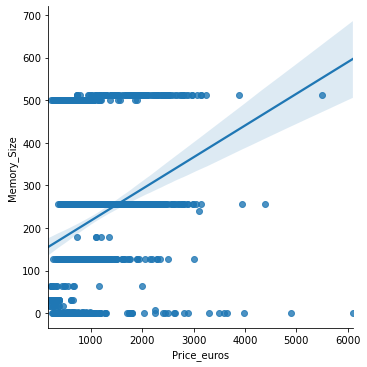


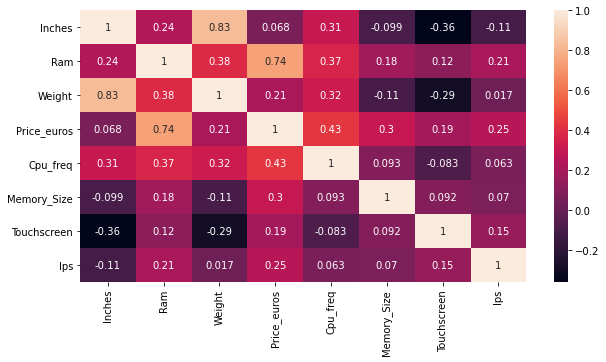
Most Sold product

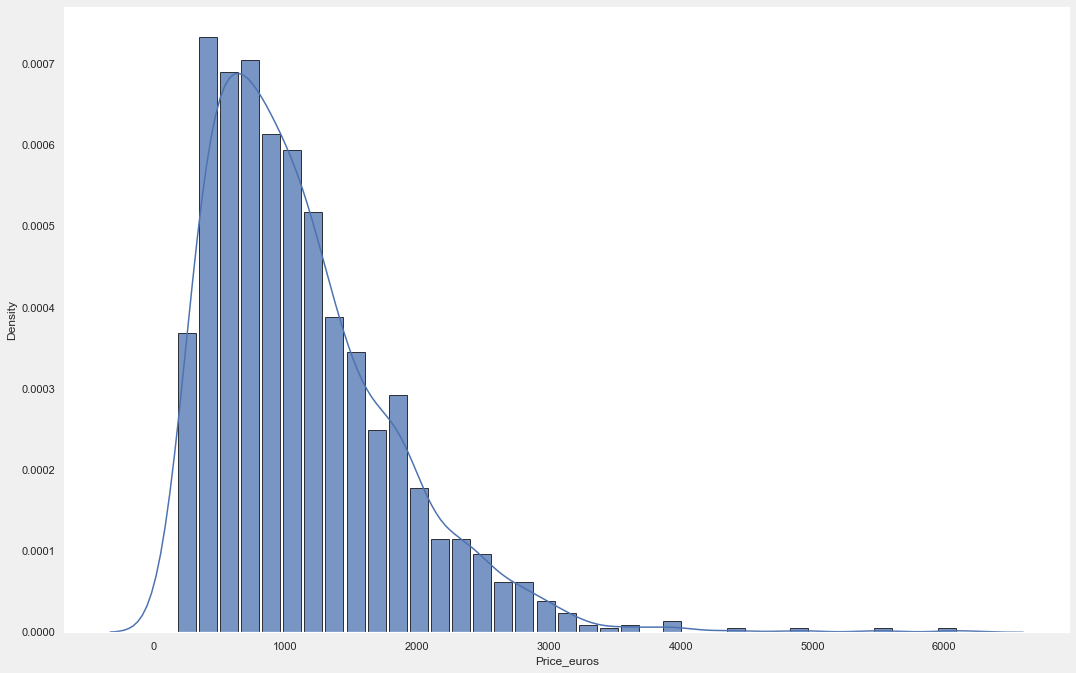


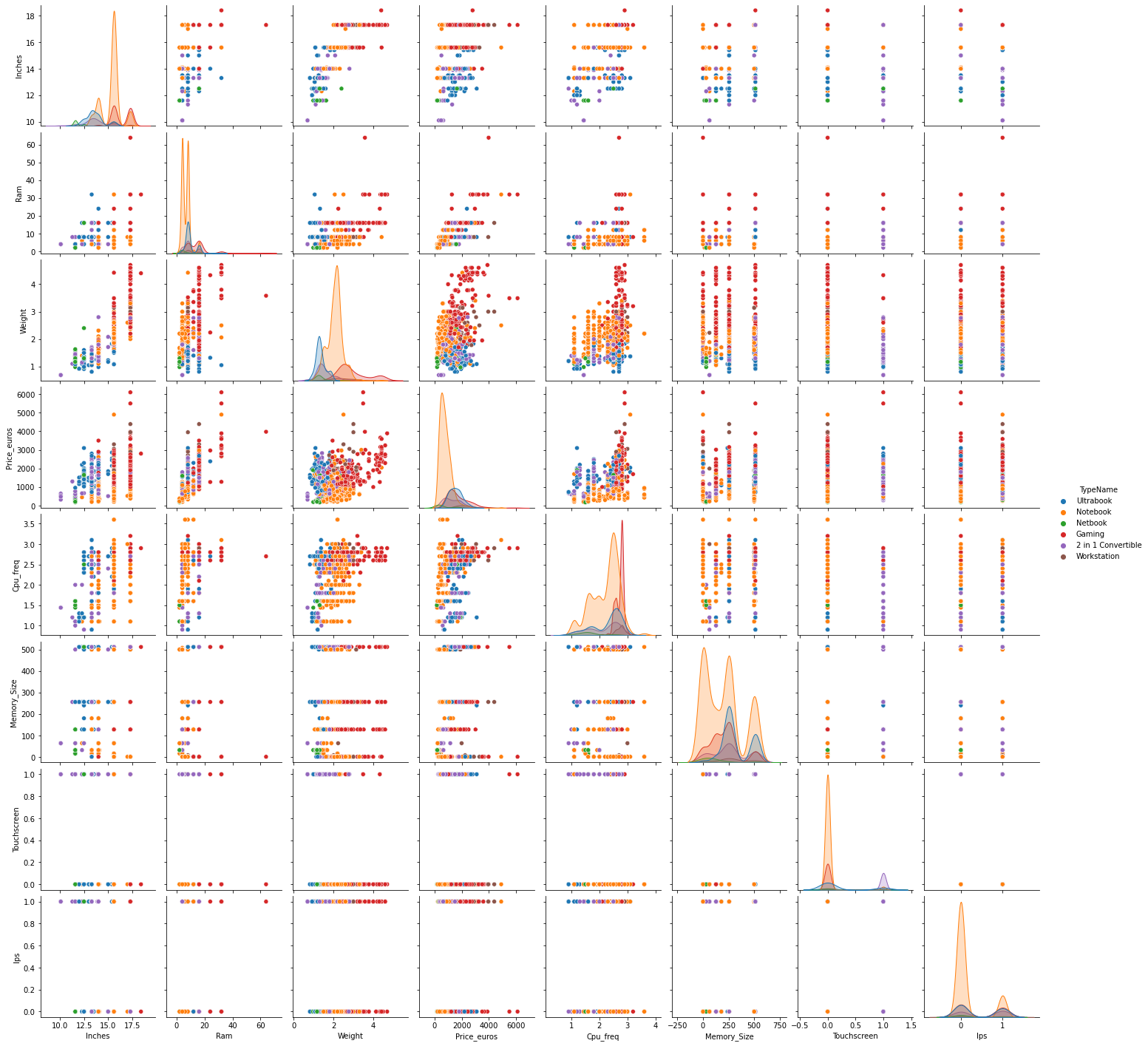
Resolution that people prefer











**Insights:**

* On average, the most expensive laptop brand is Razer and the most popular is Dell and Lenovo,
* Among the types of equipment, the notebook is the most popular, while gaming equipment and workstations are the most expensive,
* Size of the laptop screen is neither stimulant nor destimulant, as on average laptops with a medium-sized screen (between 14 and 16 inches) have the lowest price,
* Basically, the higher the screen resolution, the higher the price for a laptop, and by far the most popular resolution is FullHD 16:9,
* On average, the most expensive laptops have a 7th generation Intel Core processor, while the cheapest ones have the old version of Intel or Intel Core i3 processors and AMD processors,
* Most graphics cards are manufactured by Nvidia and usually have the highest prices, while many devices do not have a graphics card and use graphics from Intel,
* RAM definitely affects the model - the greater its amount expressed in GB, the more expensive the laptop's price,
* A similar relationship to the one mentioned above applies to hard memory, but the tendency is not so clearly visible there,
* Operating system has a relatively low impact on the prices of the laptop as the most expensive average laptops have Mac OS (but the most expensive laptops have Windows),
* There is no clear correlation between the weight and price of a laptop, but the heaviest laptops tend to be expensive in terms of their components,
* Model database contains 1415 rows and 51 columns, 1/4 of which was allocated to the test set and the remaining 75% to the validation training set (also divided in the proportion of 3 to 1),
* Neural network model has been optimized with regard to the following parameters: activation function w and number of neutrons in 1 value, learning rate and batch size and number of epochs,
* Mean error (RMSE) on the training set was 383 euros and on the test set 441 euros, and the model was not over-learned in the learning process,
* Most influential variables in the model turned out to be RAM, screen resolution and hard memory.

2.

i. 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

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

The evaluation metrics are:

* **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**.

* **Mean Squared Error(MSE)**

MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points. It 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.

* **RMSE**

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.

* **RMSLE**

Taking the log of the RMSE metric slows down the scale of error. The metric is very helpful when you are developing a model without calling the inputs. In that case, the output will vary on a large scale.

To control this situation of RMSE we take the log of calculated RMSE error and resultant we get as RMSLE.

* **R squared**

R Square is calculated by the sum of squared of prediction error divided by the total sum of the square which replaces the calculated prediction with mean. R Square value is between 0 to 1 and a bigger value indicates a better fit between prediction and actual value.

* **Adjusted R Squares**

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.

iii.Can R squared be negative?

Note that **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.

iv. What is dummy variable trap?v.Is One Hot Encoding different from Dummy Variables?

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 use one-hot encoding for handling the categorical data, then one dummy variable (attribute) can be predicted with the help of other dummy 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

**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.

vi. How is polynomial regression different from linear regression?

**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.

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

An R2 of 1 **indicates that the regression predictions perfectly fit the data**. Values of R2 outside the range 0 to 1 occur when the model fits the data worse than the worst possible least-squares predictor (equivalent to a horizontal hyperplane at a height equal to the mean of the observed data). R2 always increases as you add additional parameters. An R2=1 indicates perfect fit. That is, you've explained all of the variance that there is to explain. In ordinary least squares (OLS) regression (the most typical type), your coefficients are already optimized to maximize the degree of model fit (R2) for your variables and all linear transforms of your variables .You can always get R2=1 if you have a number of predicting variables equal to the number of observations, or if you've estimated an intercept the number of observations - 1.

viii.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.

Other options for Sweetviz are :

1. **dtale**
2. **pandas profiling**
3. **sweetviz**
4. **autoviz**

Out of these dtale seemed to be more user-friendly and easy to understand. It works more or less similar to tableau