**ML 1 Lab – 02**

Date: 23-04-2022

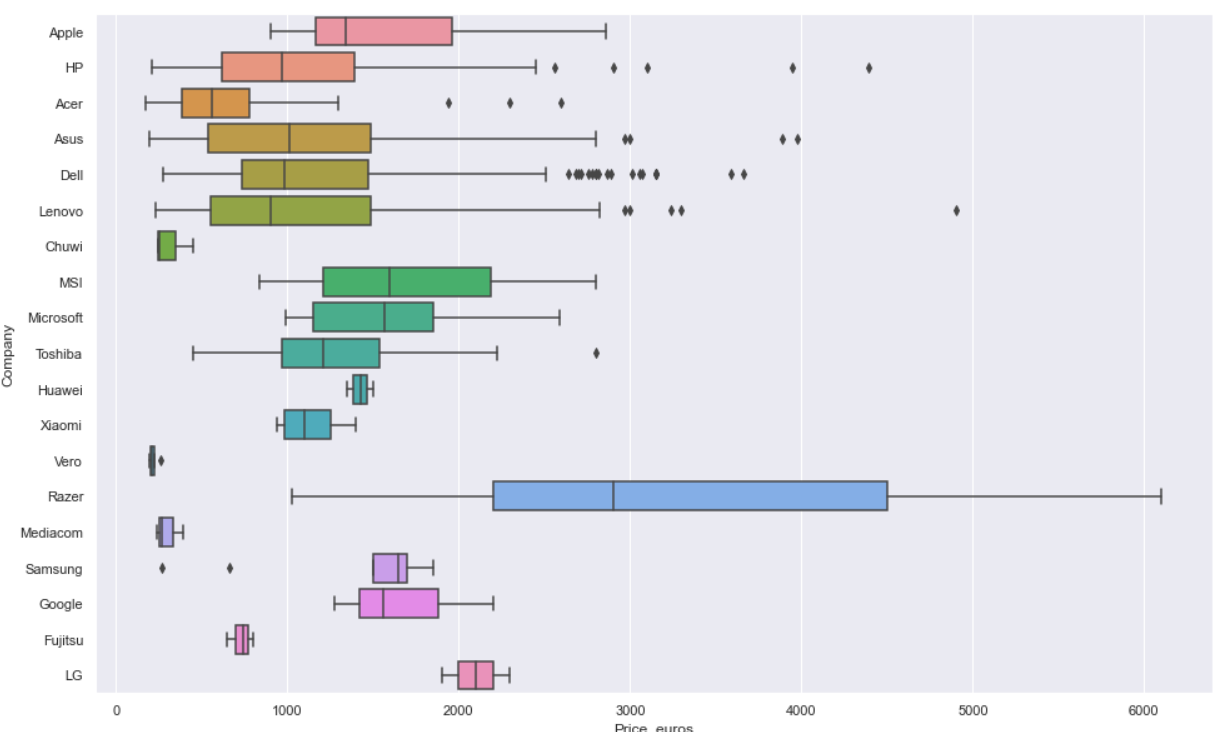
By: 21BDA10 – Aishwarya CS

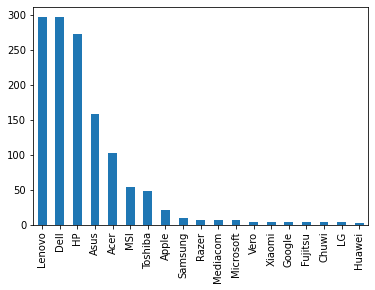
**Question 1: 5 insights from EDA**

My dataset is about Laptop where it has 1303 rows and 13 columns.

Attributes are:

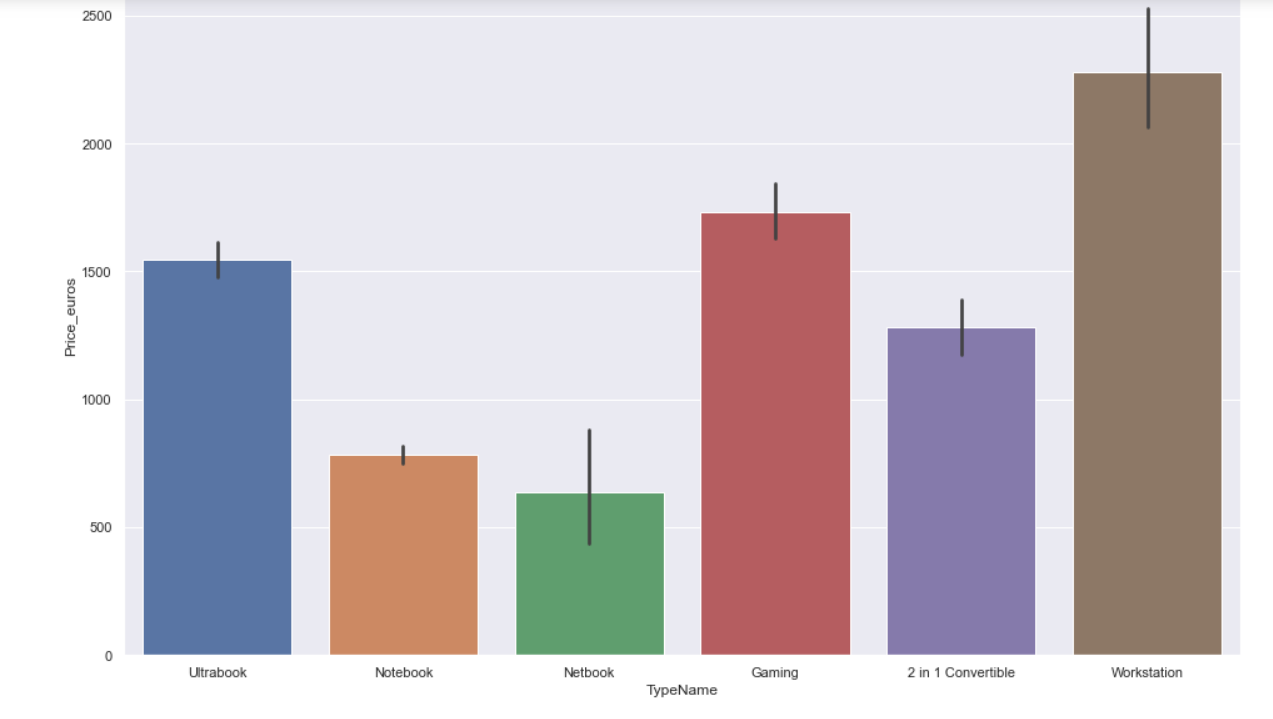
1. laptop\_ID
2. Company – Laptop Manufacturer
3. Product – Brand and Model
4. TypeName – Type
5. Inches – Screen Size
6. ScreenResolution – Screen size
7. Cpu – Central Processing Unit
8. Ram – Laptop RAM
9. Memory – Hard disk / SSD memory
10. Gpu – Graphics Processing Units
11. OpSys – Operating Sysytem
12. Weight – Laptop Weight
13. Price\_euros – Price in Euro
14. Price vs Company

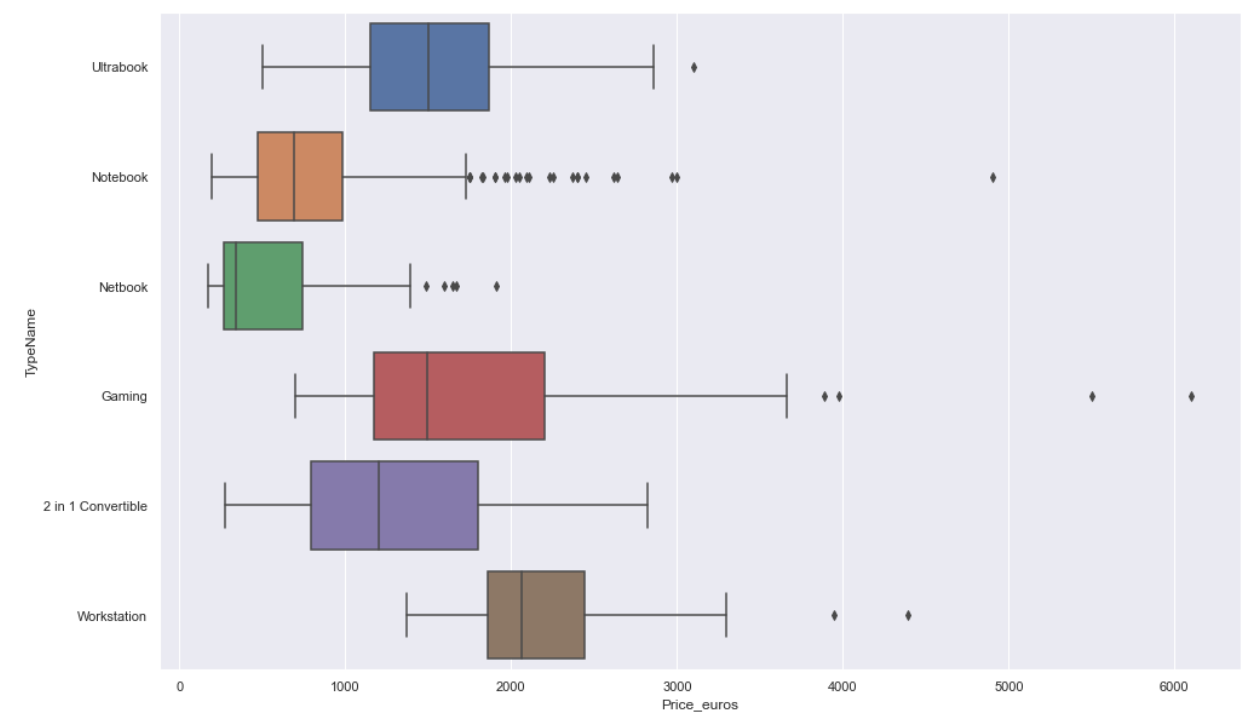




* From the above graphs, we have 19 different types of laptop manufacturers.
* Out of 19, Lenovo and Dell have 297 devices each.
* From the boxplot, we can see that Razer has the highest average price almost 3000 euros but it has less number of laptops.
* Chuwi, Vero, and Mediconhaves less average prices compared to other laptops.
* Lenovo, Dell, and Hp are the most used laptops and the average price of the laptop is less than 1000 which can be afforded by many customers.

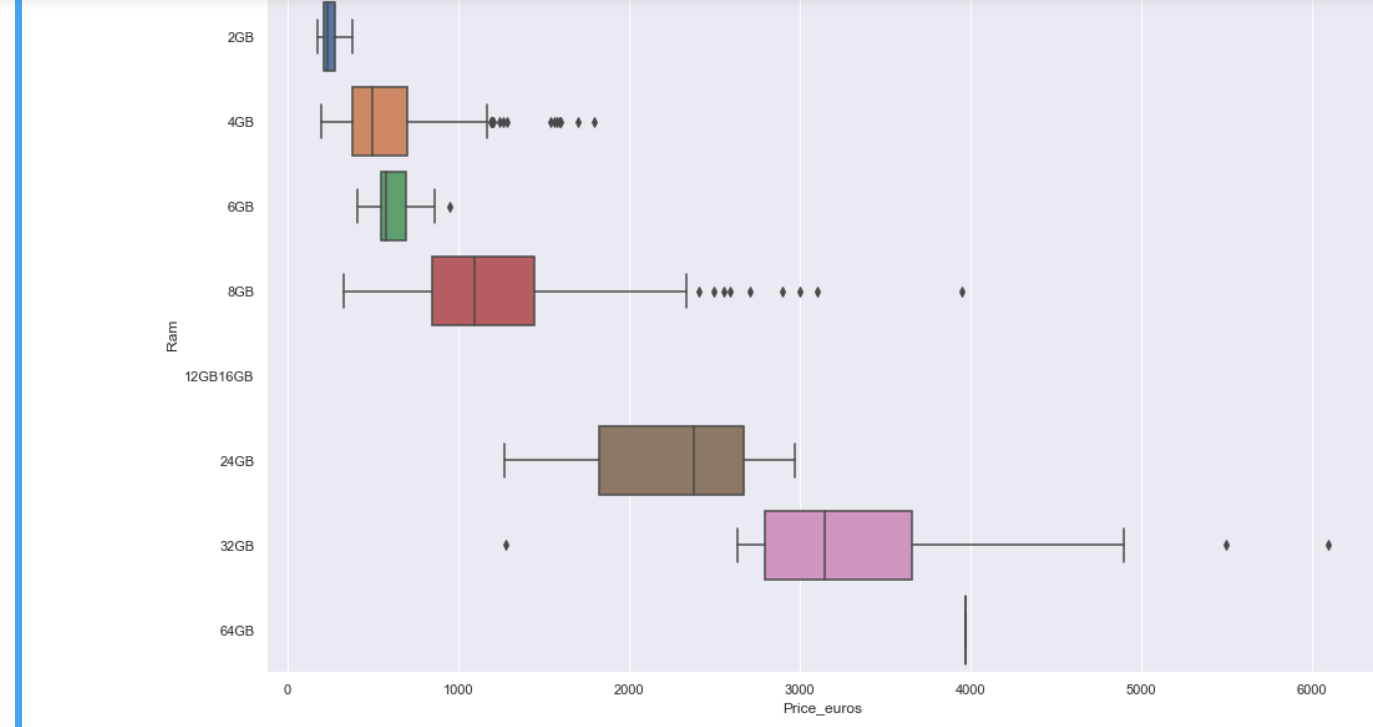
1. TYPENAME VS PRICE





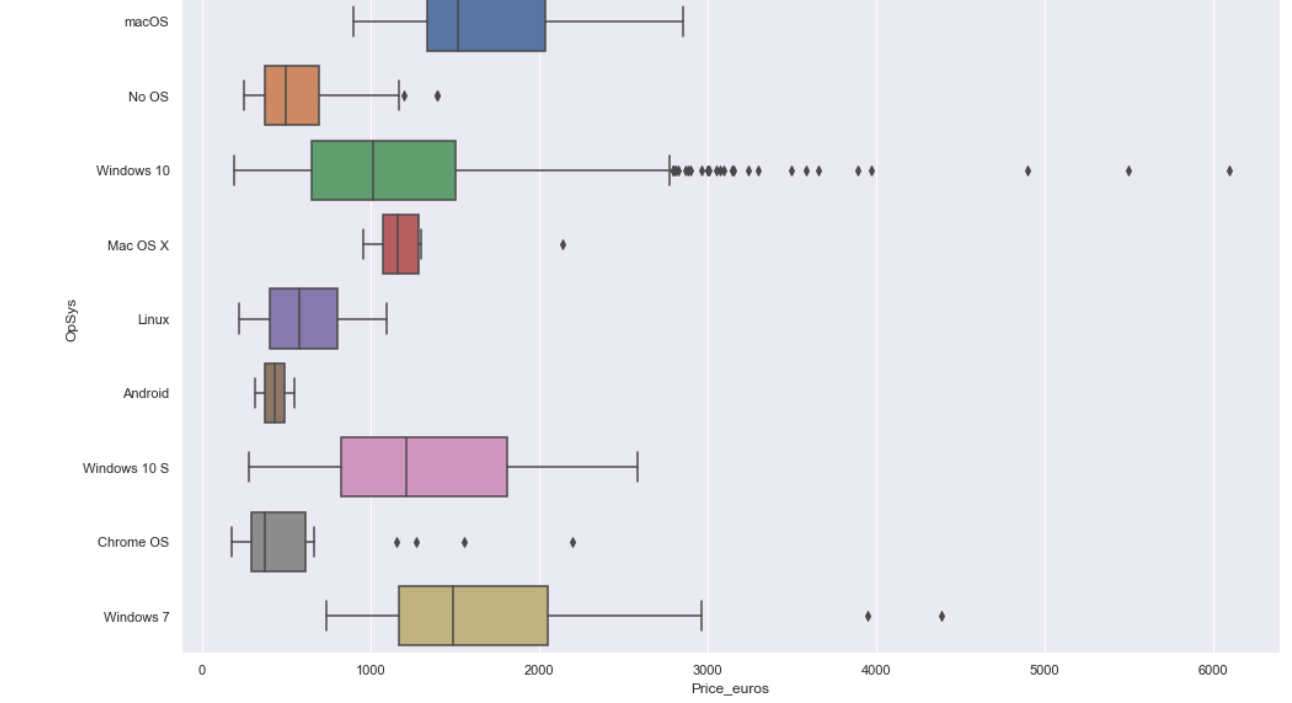
* There are 6 types of Laptops, and most of them are Notebook-type laptops.
* Most of the laptops are Notebook type (almost 50 percent).
* Workstation laptop type has the highest average price.
* Notebooks and Netbooks type has the lowest average price compared to other types.
* Gaming and Ultrabook are the second and third highest average prices.
* We can say that notebooks and Gaming types of laptops are used by most of the customers.

1. PRICE VS RAM



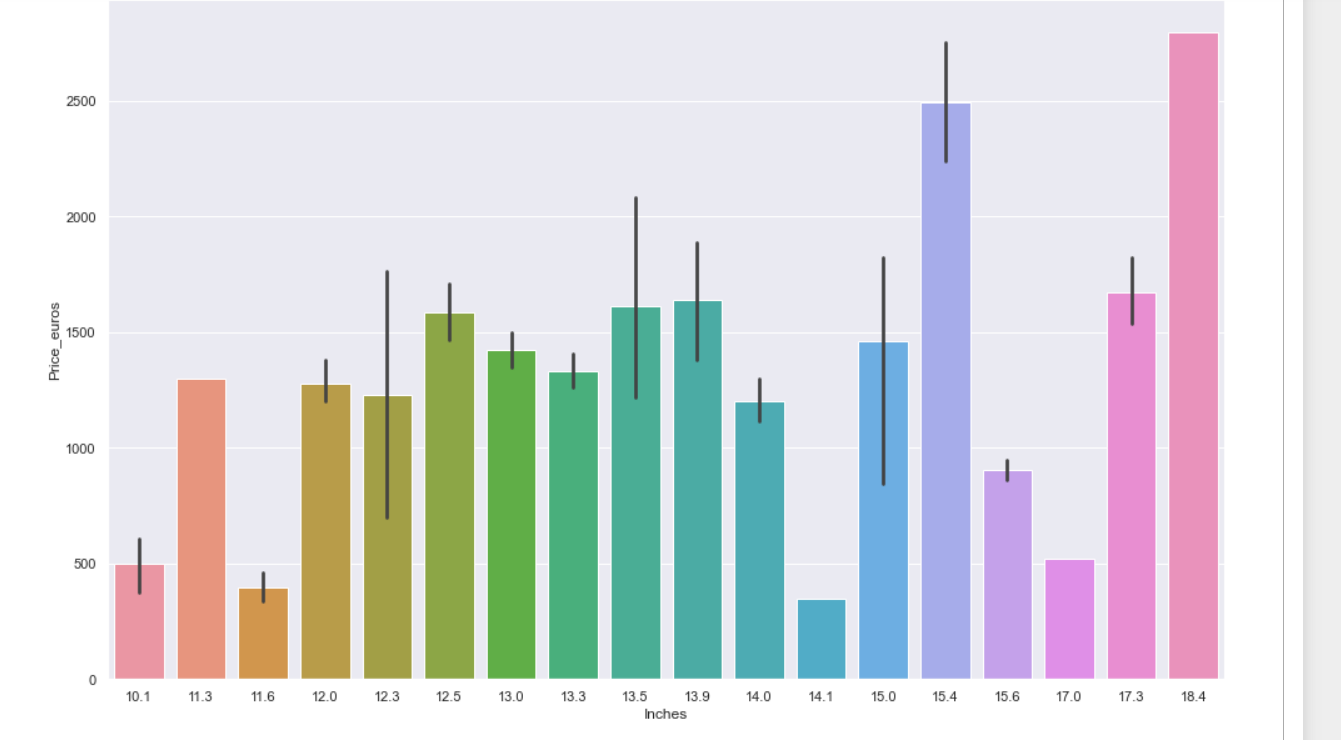
* In this dataset, the size of the laptop’s RAM range from 2GB to 64GB. It has 9 different sizes of RAM.
* From the graph we can see that as the size of RAM increases the price of the laptop also increases.
* 2GB RAM has the lowest average price and 64GB RAM has the highest average price.
* 8GB RAM laptops are used most by customers.

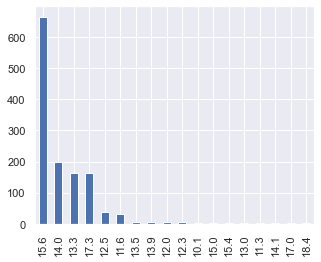
1. PRICE VS OPERATING SYSTEM



* There are 9 different types of operating systems.
* macOS and Windows have the highest average price.
* Chrome OS and Android have the lowest average price.
* Out of these many operating systems, windows 10 OS is the most used by customers. (about 1072)

1. PRICE VS INCHES





* We have 18 different inches of laptop.
* From the second graph we can see that 15.6 laptops are used most by customers.
* 18.4-inch laptops have the highest average price but it is by only one customer.
* 14.1-inch laptop has the lowest average price.
* From the first and second graph we can say that 15.6-inch laptops are used most and their price is less than 1000 euros and can be afforded by all the customers

**Question: 2**

1. **What are the assumptions of Linear Regression?**
2. The relationship between dependent variable y and independent variable x exists is LINEAR.
3. For every value of the independent variable x, there is an expected value of the dependent variable y and these values are normally distributed.
4. Y is a continuous random variable and x values are fixed and not random.
5. The sampling errors associated with the expected value of y are associated to be an independent random variable distributed normally with mean 0 and constant standard deviation.
6. **How can we evaluate Regression model? Define each metric and its interpretation**

There are 3 main metrics for model evaluation in regression:

1. R Square / Adjusted R Square
2. Mean Square Error (MSE)
3. Mean Absolute Error(MAE)

🡺 R Square / Adjusted R Square

R square measures how much variability in the dependent variable can be explained by the model. it is a good measure to determine how well the model fits the dependent variables. Its value is between 0 to 1 and a bigger value indicates a better fit between prediction and actual value.

In python we can calculate R square using Statsmodel or sklearn package.

🡺 Mean Square Estimate

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

It gives us an absolute number on how much your predicted results deviate from the actual number. We 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

🡺 Mean Absolute Error

MAE is similar to 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, 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.

1. **Can R Squared can be negative?**

Yes, R Squared can be negative if regression is a worse fit.

1. **What is a 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 use one-hot encoding for handling the categorical data, then one dummy variable(attribute) can be predicted with the help of other dummy variables. Using all dummy variables for regression models leads to a dummy variable trap.

1. **Is One Hot Encoding different from Dummy variable?**

Both of them are the same.

Dummy Variable is a binary variable that indicates whether a separate categorical variable takes on a specific value.

In one hot encoding technique is a function or code to create a dummy variable. let us take gender as an example which is a categorical variable with parameters like male and female. So by using this technique we can give a numerical value to male as 0 and female as 1 or vice versa.

Here dummy variables are equal to a number of categories.

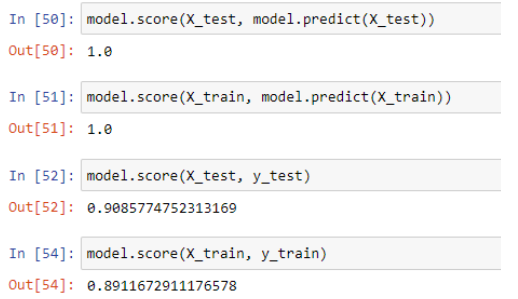
A dummy variable is a numeric variable that represents categorical data.

1. **How is polynomial regression different from linear regression?**

A simple linear regression measures the relationship between two variables by fitting a linear equation to the data. For linear regression, the relationship between the data should be linear.

But suppose if we have non-linear data then linear regression will not capable to draw the best fit line and it fails in such conditions.

To overcome the above problem we introduce polynomial regression which helps to identify the curvilinear relationship between independent and dependent variables.



Model.score() is used to get the accuracy of a model whereas model.predict is used to get predict variable.

In the first two lines of code, we can see that the r squares value is 1 which is the best fit. The first, code shows the variation of x\_test is explained by the predicted value of x\_test. the In second code, the variation of x\_train is explained by the predicted x\_train variable.

In the third and fourth code, the accuracy of x\_test and y\_test is 0.90 i.e 90% and the accuracy of X\_train and y\_train variables are 0.89 i.e 89%