# **Introduction**

# **Data Understanding**

For this project the dataset that was chosen is named “Laptop Price” available on Kaggle. Using the “df.shape” and “df.info” functions, is it possible to see that the data size consists of 1303 and 13 columns, of which eleven columns are represented by ‘object’ data type, and only

three column with numerical data, being consisted of float64 and int64 type.

By using the “df.describe” to obtaining the summary statistics for this dataset. We can the summary statistics for only numerical data, which are laptop\_ID, Inches and Price\_euros.

A screenshot of a computer

Description automatically generated

Figure 1. Summary (describe).

The measures of dispersion evaluate how scattered the collected data are. They are standard deviation, variation and interquartile range. If the value of the standard deviation is high, it means that the set of data is very spread out. In this case, the result of standard deviation is lower than the mean. So, it’s mean the set not very spread out.

# **Data preparation**

With the function of “df.head” it was possible to observe that in the categorical variables it was necessary to make some modifications.

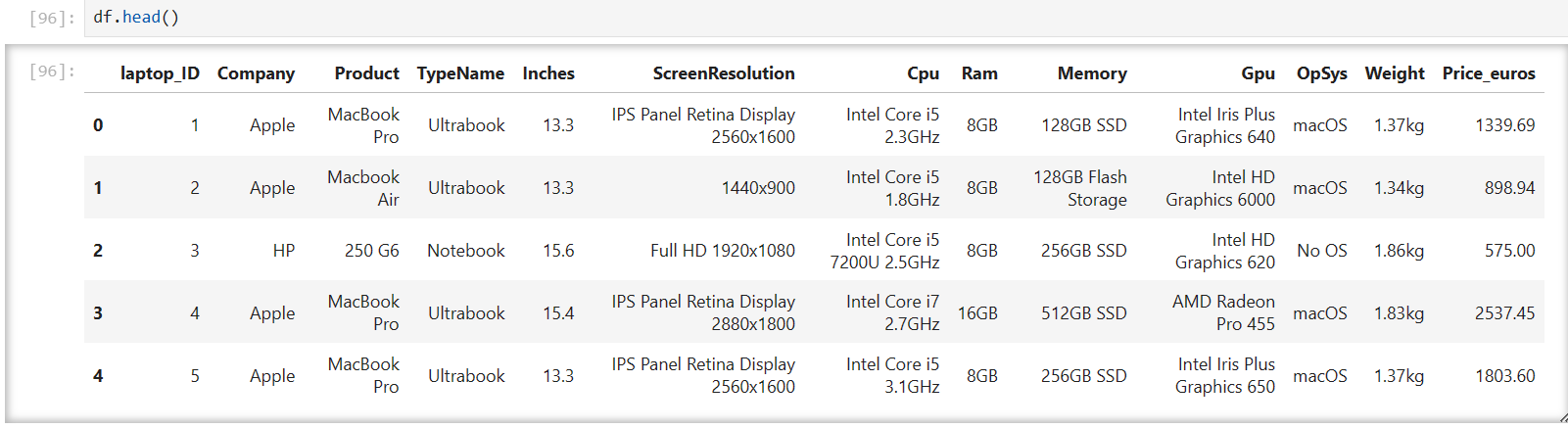


Figure 2. df.head

Starting by removing str from the RAM, Weight, Cpu\_ frequency columns, such as, GB, kg, GHz. as well as changing make replacements of dtype of the columns. Ram was transformed into int32, weight into float32. And the same thing happened on screenR, Cpu\_frenquency.

By applied the lambda function to a single column. New columns is formed in ‘ScreenResolution’, Appearing Two New Columns ‘screen\_width’ and 'screen\_height'.