APPLIED BUSINESS STATISTICS - ASSIGNMENT 1
Phone Prices in India

This Case Study will investigate what determines, predicts, and explains Mobile Phone Prices in India. We use the following Dataset downloaded from Kaggle.
This is the following Description of the present variables:

1. **Brands:** The brands of smartphones included in the dataset.

2. **Colors:** The colors available for the smartphones.

3. **Memory:** The storage capacity of the smartphones, typically measured in gigabytes (GB) or megabytes (MB).

4. **Storage:** The internal storage capacity of the smartphones, often measured in gigabytes (GB) or megabytes (MB).

5. **Rating:** The user ratings or scores assigned to the smartphones, reflecting user satisfaction or performance.

6. **Selling Price:** The price at which the smartphones are sold to consumers (In Indian Rupees).

7. **Original Price:** The original or list price of the smartphones before any discounts or promotions (In Indian Rupees).

8. **Mobile:** Indicates whether the device is a mobile phone.

9. **Discount:** The discount applied to the original price to calculate the selling price.

10. **Discount percentage:** The percentage discount applied to the original price to calculate the selling price.

Problem Definition

The goal of this research is to understand what determines price in an eastern country. With linear regressions, we will analyze the correlations between the characteristics of the phone, and its original price.

This study can be useful to understand, at the end of the day, what matters most in determining the price of a phone. As a smartphone manufacturer, this would be useful to understand how to handle the different characteristics of a product when targeting a specific selling price for our phone.

Methodology

1 - Cleaning and adjusting Data

It is necessary to clean and engineer data for analysis, so we have provided a Jupyter notebook called '*Dataset_Engineering.ipynb*'

After doing, and cleaning all the datasets for future usage, we decided to use only the following variables:

1. **Brands:** The brands of smartphones included in the dataset.

2. **Memory:** The storage capacity of the smartphones, typically measured in gigabytes (GB) or megabytes (MB).

3. **Storage:** The internal storage capacity of the smartphones, often measured in gigabytes (GB) or megabytes (MB).

4. **Rating:** The user ratings or scores assigned to the smartphones, reflecting user satisfaction or performance.

5. **Selling Price:** The price at which the smartphones are sold to consumers (In Indian Rupees).
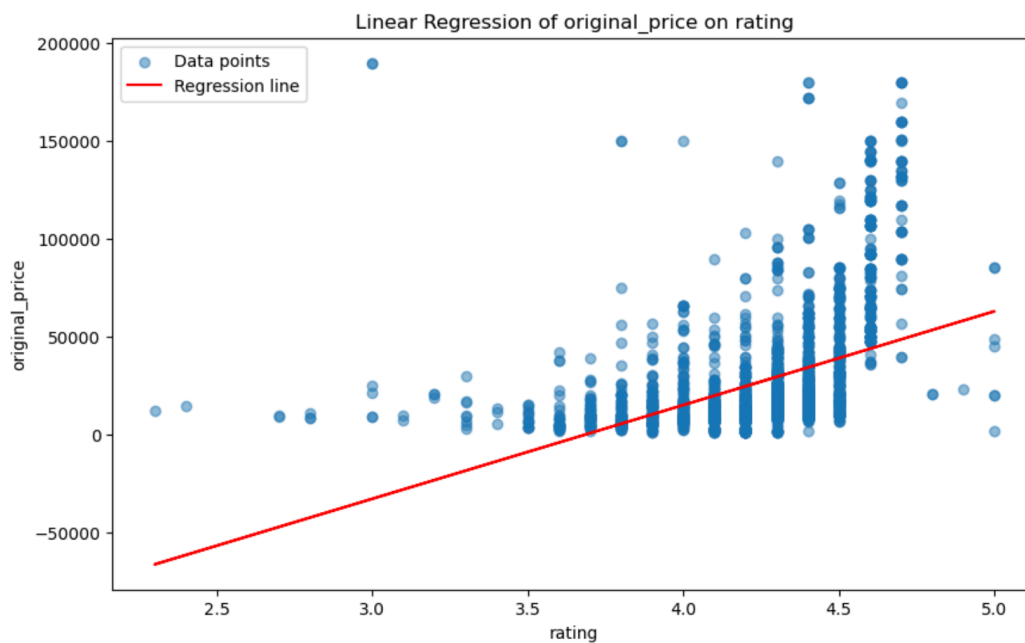
6. *Pro:* **Pro version of the phone.**

2 - Linear Regressions

We will try to use linear regressions to see if any numerical variable, representing a structural feature, is able to predict the original selling price.
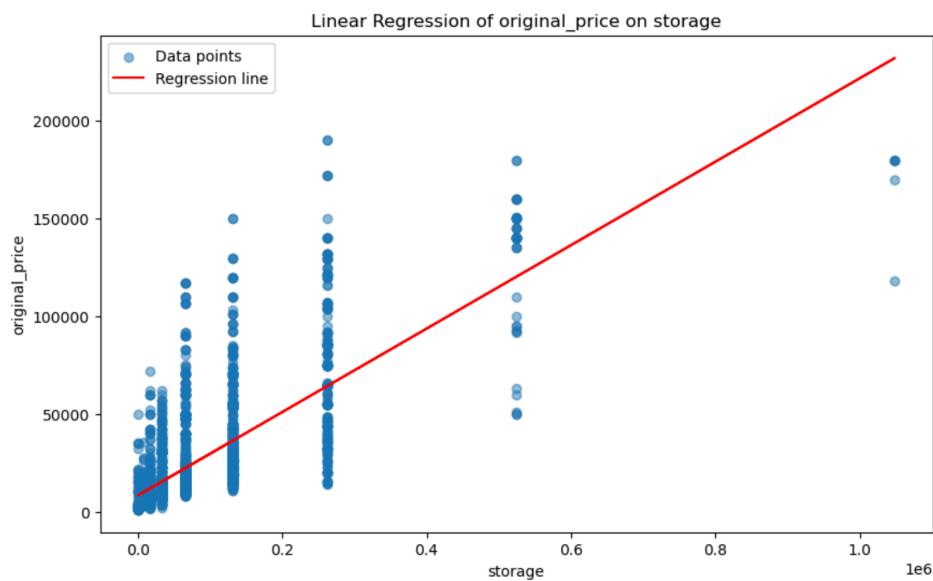
- RAM

-Storage

- Rating

Linear Regression of original_price on storage

We can see the results on '*Reg_Analysis.ipynb*', but all of these regressions yield very unsatisfactory results. Not only graphically, but also the highest $R^2 = 0.520$, which means that only half of the variance is explained.

4 - Multiple Regression

We were expecting that, although it is a different culture than the one we live in, the physical characteristics of the phones were not crucial in determining the price, and, in general, the demand of a product. We believed that the brand would have a huge impact. So, we decided to combine the two data types we had, and one hot encoded brands, to perform a Multiple regression, and these are the graphical results.

Actual vs. Predicted Price
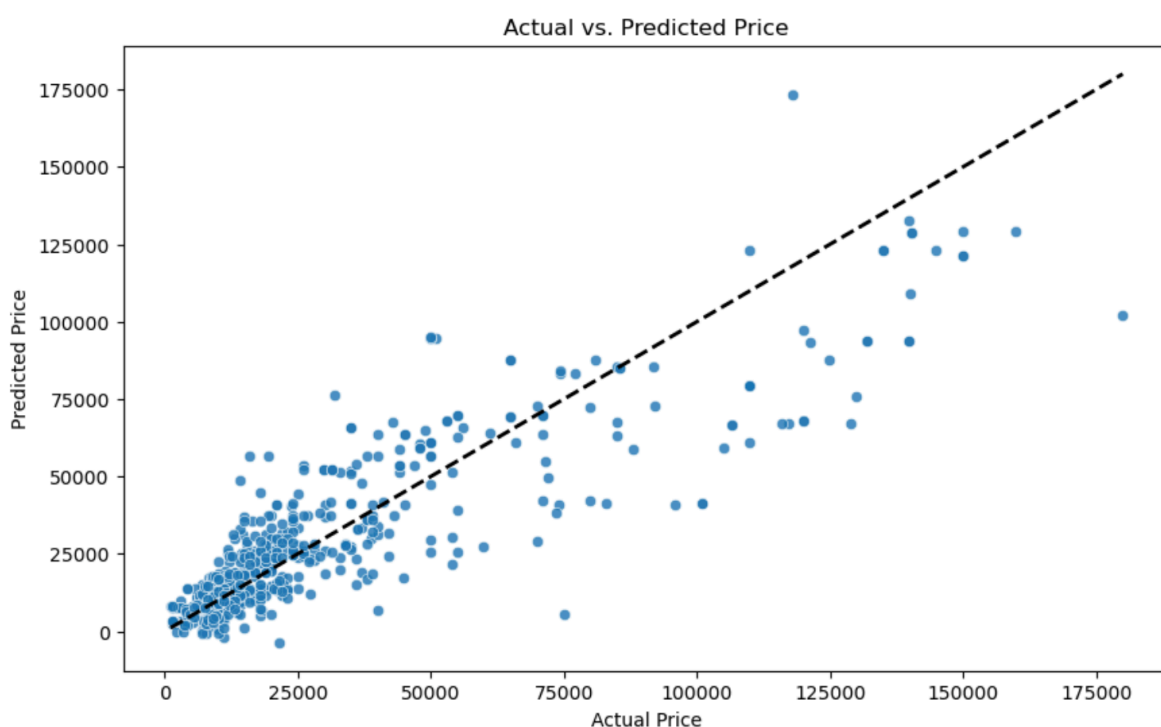
With an $R^2 = 76.77\,\%$, we can see, and verify, that this is a much better predictor for the price, although still not great.

Just to have a better sense of the results, the Rsme is 15132.81, which, in euros, means that on average the result is 166,99 € off the real price. It is still a lot, but on higher-end phones, which the dataset substantially included, it is less relevant overall.

The question could have also been: is the brand, alone, capable of achieving similar results?
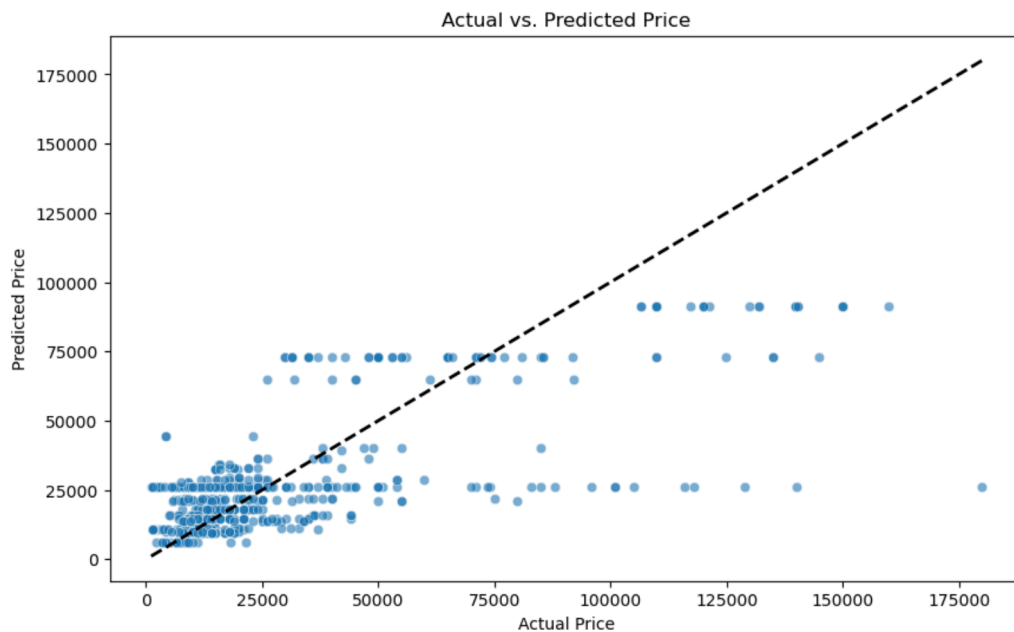
The $R^2 = 55.03\,\%$ drops back down again.

Conclusions:

As a bottom line, only by mixing characteristics, and branding, we can have a good estimation of the selling price of a product. In general, it makes sense, as brands in the industry, especially the bigger ones, have developed a wide product differentiation, that targets segment clients based on income and needs, and adapts the features of the product based on that.