

Exploratory Data Analysis for Machine Learning

IBM Skills Network

Project Report

By:

Fahd Seddik

Table of Contents

[Brief Description 3](#_Toc113036489)

[Initial plan 3](#_Toc113036490)

[Data cleaning & Feature engineering 4](#_Toc113036491)

[Key Findings and Insights 7](#_Toc113036492)

[Hypothesis 9](#_Toc113036493)

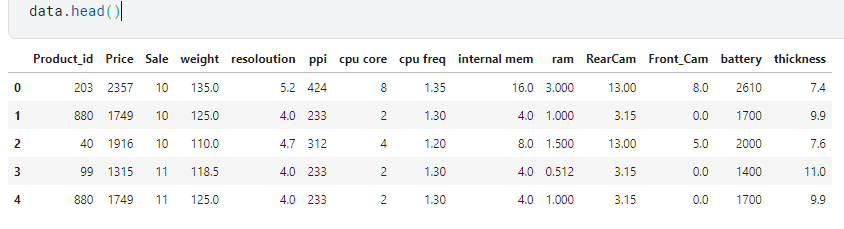
[Suggestions 10](#_Toc113036494)

[Summary 10](#_Toc113036495)

# Brief Description

Mobile price depends on various factors such as resolution, brand, size, weight, imaging quality, RAM, battery and CPU power. In this dataset, we want to estimate the price of mobile phones using the above features.

## Columns:



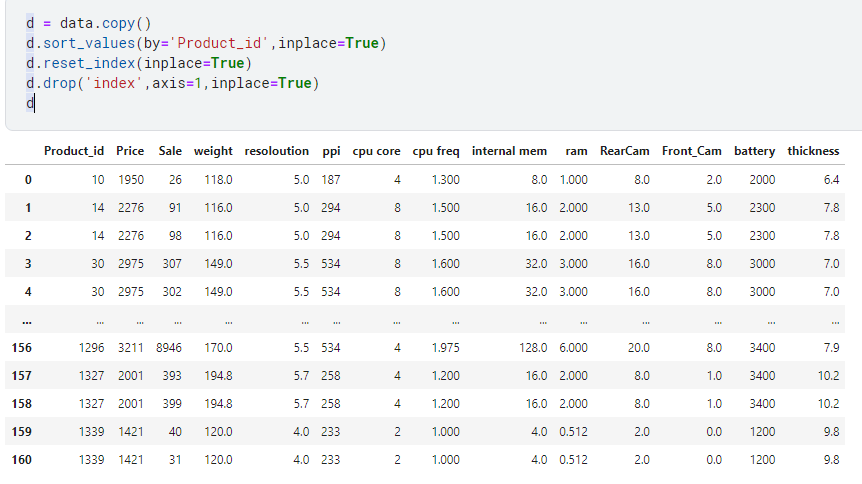
# Initial plan

The plan would go as follows:

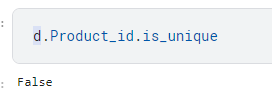
* Check for duplicates and deal with any
* Check for missing values and deal with any
* Calculate correlation values
* Check for skewness of data
* Visualize through boxplots to check for outliers
* Apply feature engineering to formulate possible useful features
* Use seaborn pair plots to see underlying patterns
* Construct hypothesis about data set

# Data cleaning & Feature engineering

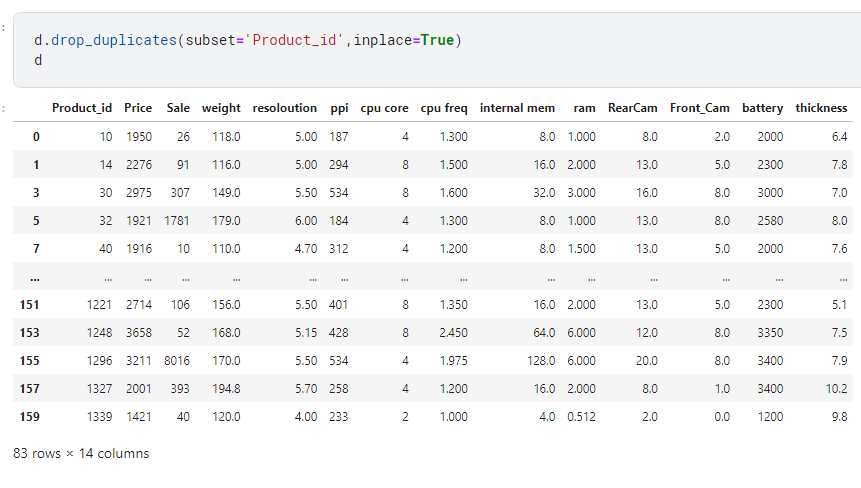
We will first sort by Product\_id to make data more readable



## Check for any duplicates



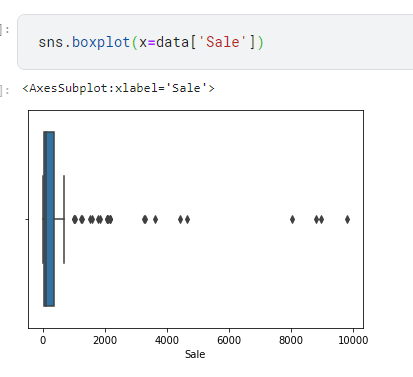
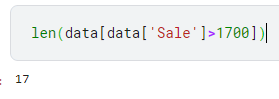
We will drop rows with duplicate ids



## Check for missing values

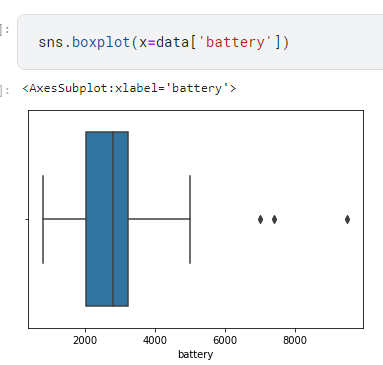
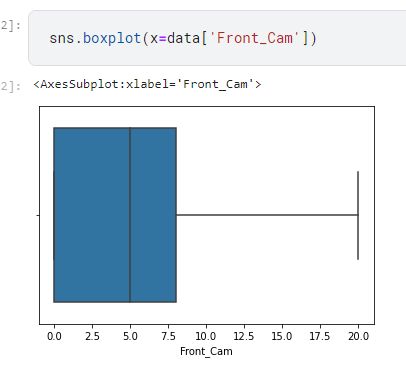
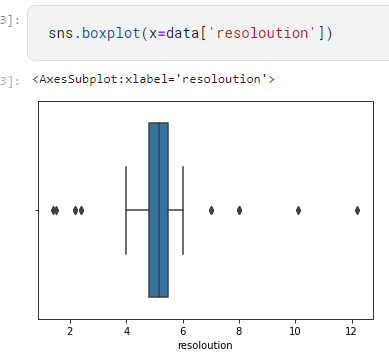
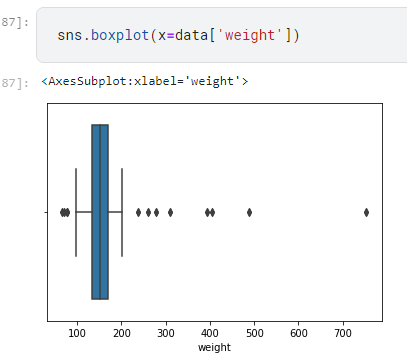
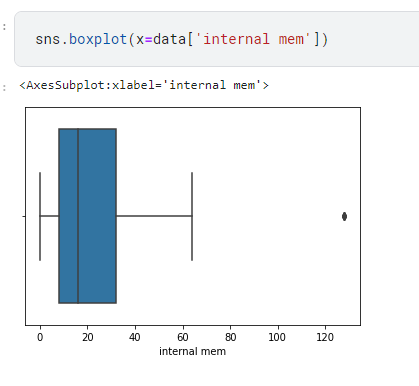
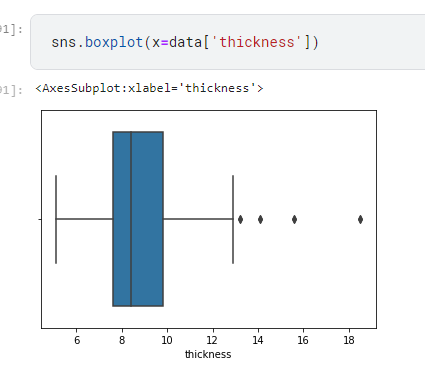
Since number of entries is 83 and all columns contain 83 non-null entries then we do not have any missing values.

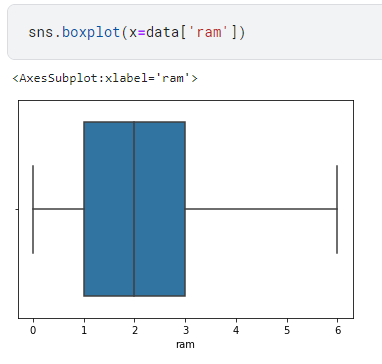
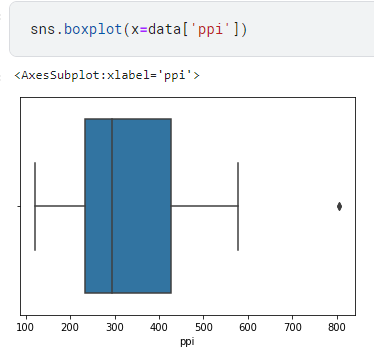
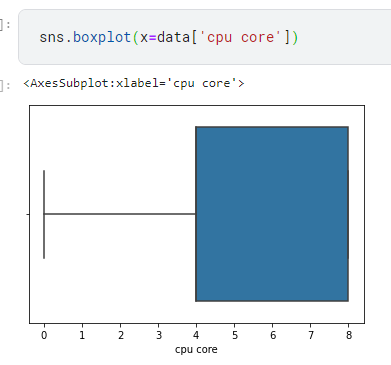
## Check for outliers

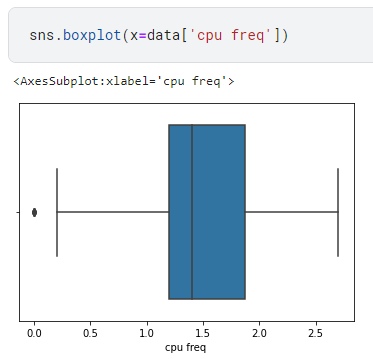
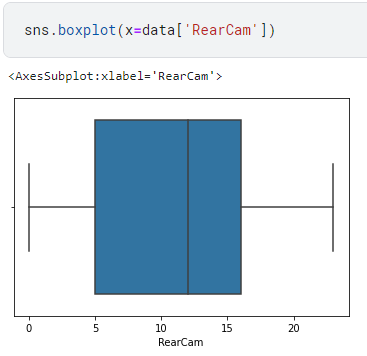


Will drop outliers in Sale column. On the next page, there will be box plots for all features.



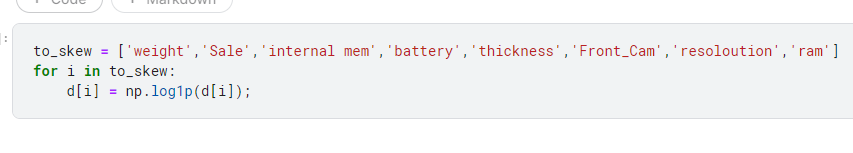


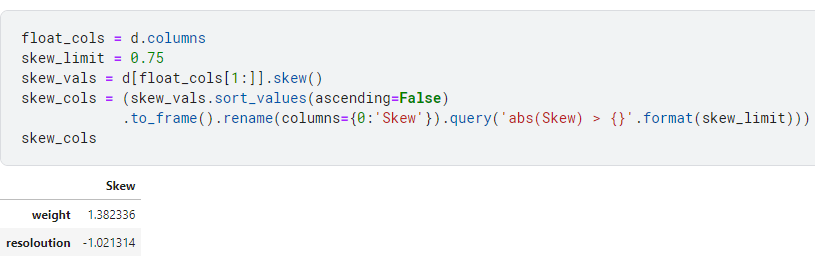




Calculate each skew value

We will apply **log transformation** to all skewed columns.





After applying log transformation to our columns, skewness values are mostly corrected. We end up with only 2 columns with skewed values instead of the initial 8 columns.

## Feature Engineering

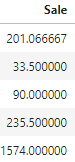
Since thickness is more likely to have an inverse relation to price (would be shown in next section), then we can add a 1/thickness feature.

There are some other notable feature interactions that should be mentioned:

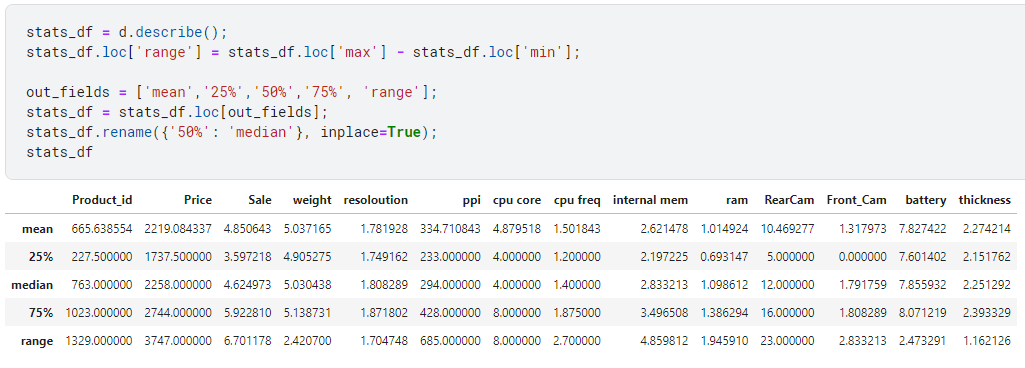
* Weight\_/thickness
* Internal mem\_\*\_ram
* RearCam\_\*\_Front\_cam

# Key Findings and Insights

## Calculate statistics (before log transformation)



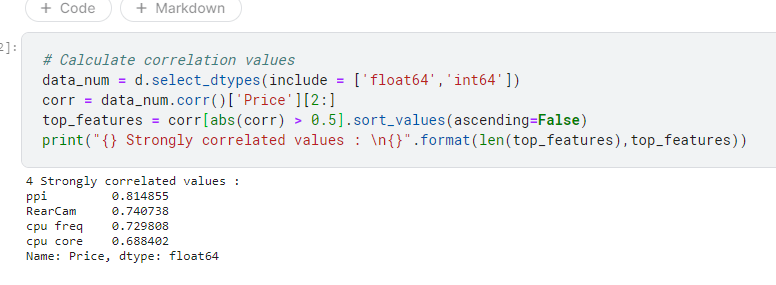
## Calculate statistics (after log transformation)





## Insights

We can tell from the pair plots there are many features how have a positive correlation with price of phone. Let’s calculate the correlation values.



# Hypothesis

We can hypothesize about the data set in several ways. Here are some of the hypotheses we can have about our data set.

1. Ho: A weight of range 140 to 160 represent 50% of examples

Ha: weight range 140 to 160 does not represent 50% of examples

1. Ho: 90% of phones with 8 cores have same range of 4 core phones

Ha: they do not have the same range

1. Ho: all phones around the world have thickness with mu = 10.98 (sample mu)

Ha: mu !=10.98

We will be conducting a formal significance test for the third hypothesis. Since our sample mu = 10.98 we will calculate t\_value, z\_value and having a significance level of alpha=0.05. This would give us a t\_value of 2.00 and z\_value of 0.4798.

# Suggestions

Of course, the analysis we did on the data set is merely scraping the surface of all possible analysis methods we can apply to this data set. We can try to formulate more features by applying feature engineering. In addition to that, we can visualize correlation values using heatmaps. Feature scaling could be one of the methods we would use if we are intending on using the data set in models that are prone to not scaled features. Calculation of z-score could be one of the ways to determine more statistics about our data set.

# Summary

Predicting the price of phones could help companies compete by just choosing the phone’s specifications and estimating how much the phone would sell for and calculate their budget.

In conclusion, I believe that there is much potential in this data set. Although further EDA could be done on this data set and fine-tune it better, but we managed to stick to the initial plan.