Assignement 3

Analysis for Turtle Games

LSE

Data Analytics Career Accelerator

**Introduction**

Turtle Games is a game manufacturer and retailer that sells their products and source and sell products manufactured by other companies. Their product range includes Lego and board games, video games, and toys. They have a global customer base and have a business objective of improving overall sales performance.

**Analysis framework**

Turtle Games wants to understand:

|  |
| --- |
| the price at which they should sell particular products |
| a useful way of segmenting their entire product range, to lay the foundations for further analysis |
| the customer group that will most likely leave a review on the products they’ve purchased |
| the most popular, expensive products purchased by a particular group of customers |
| the general sentiment of customers across various products |
| the sales they can expect to achieve for particular products they offer |

**Data ingestion**

Three datasets have been provided for this study:

game\_reviews.csv

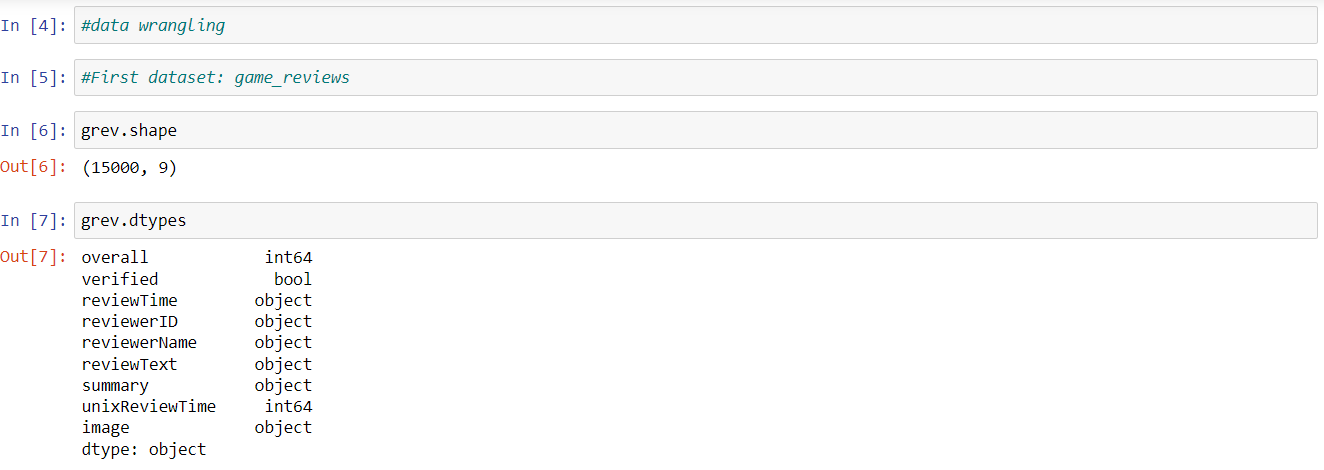
* games\_sales.csv
* lego.csv

They have been uploaded into a dedicated GitHub repository and imported into python.

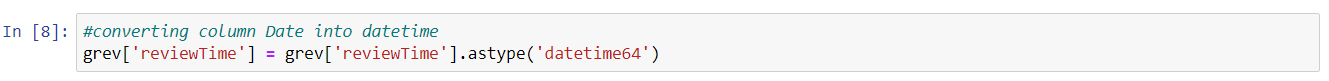
**Data wrangling**

After checking the shape of the first dataset, I proceeded to assess the data types.

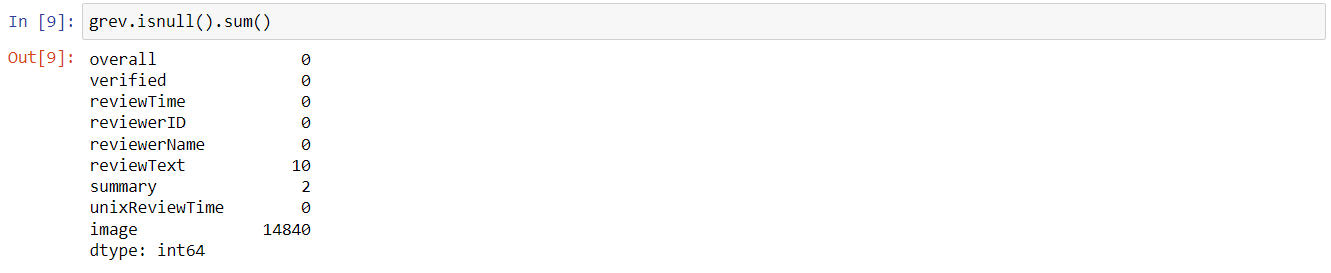
I observed that the column “reviewTime” was imported as an object.



I decided, therefore, to convert it into a DateTime object.



I then checked the presence of NaN values within the Data Frame



The column “image” is almost empty. I, therefore, decided to drop it.



I decided to drop those rows which contained at least one NaN value.

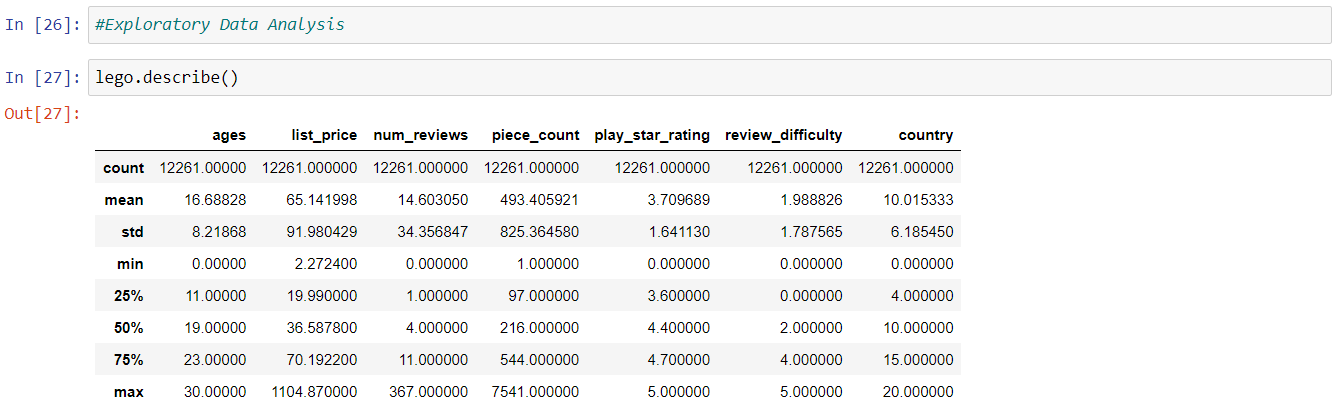


I went through the same process for the other two datasets.

**Analysis question n. 1:**

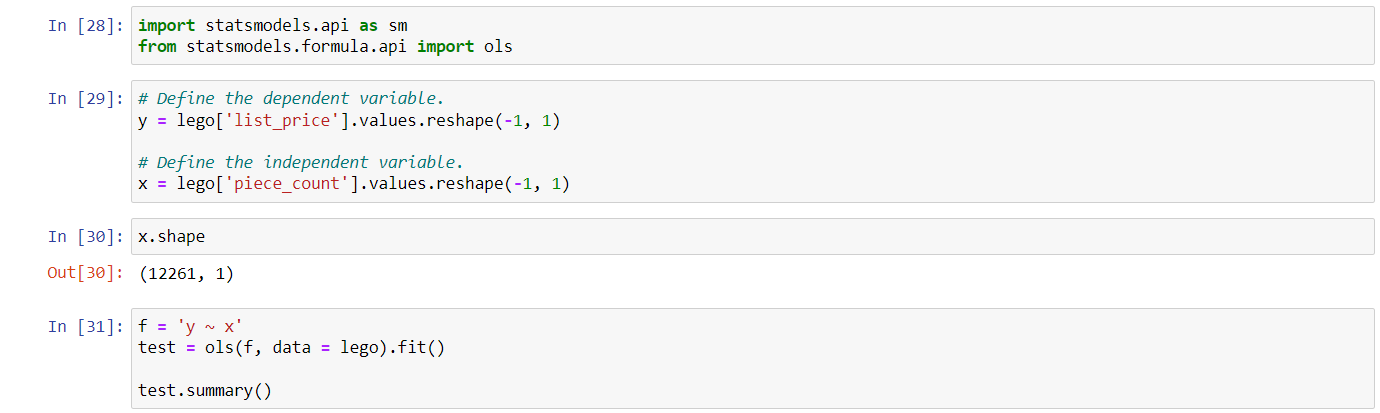
What price should be set for the Lego sets that have the 8000 Lego pieces?

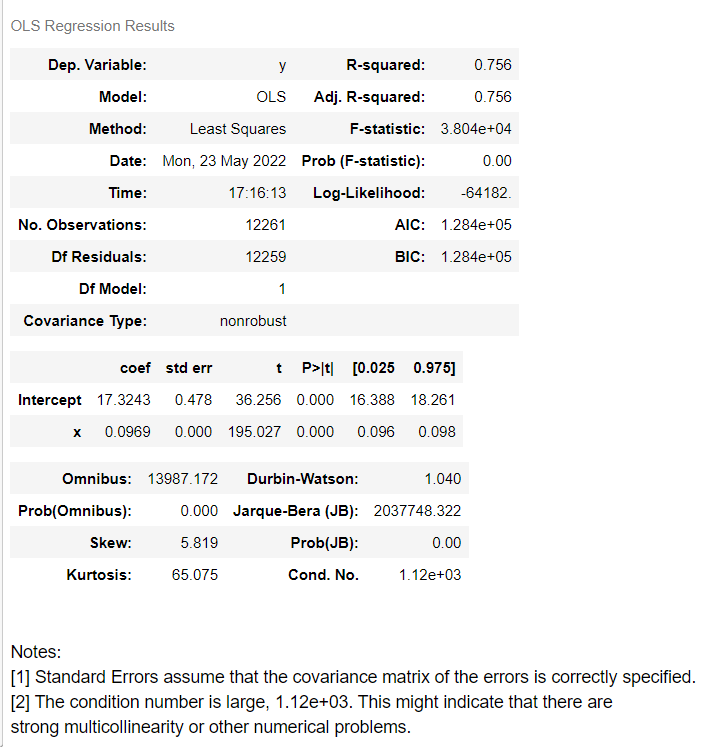
I started with my exploratory analysis first.



I defined the dependent variable (y) and the independent variable (x).

I then created the ordinary least square regression and visualised the results.



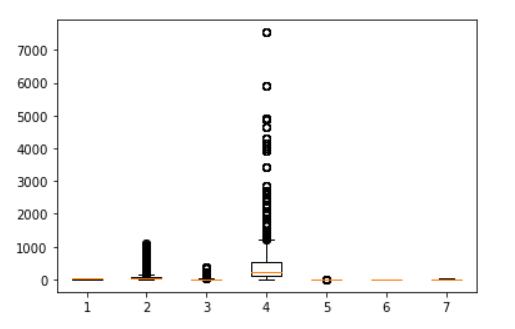


R-squared > 0.6 indicates that the data points are close to the regression line.

P>|t| 0.000 means I can reject the null hypothesis and accept the alternative hypothesis (coef!=0).

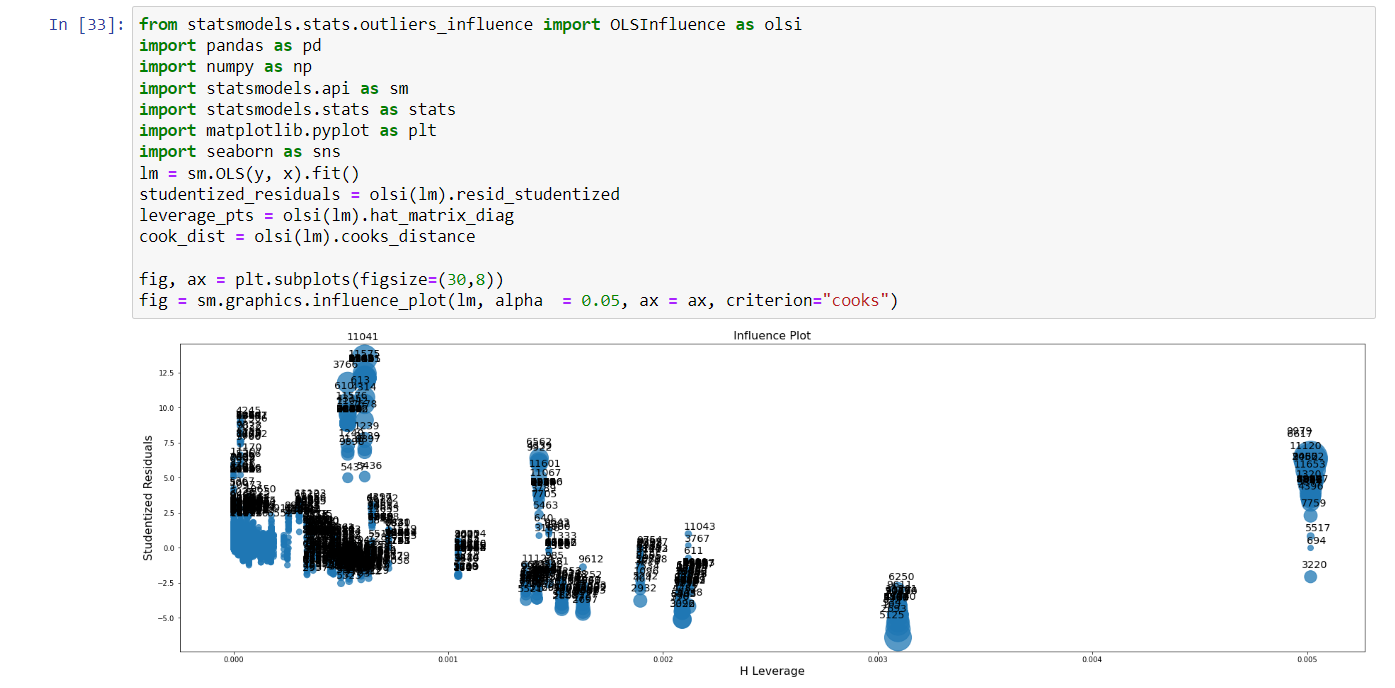
Note [2] indicates that there might be strong multicollinearity (in a linear regression??) or other numerical problems.

I then wanted to check the presence of outliers in the given dataset.

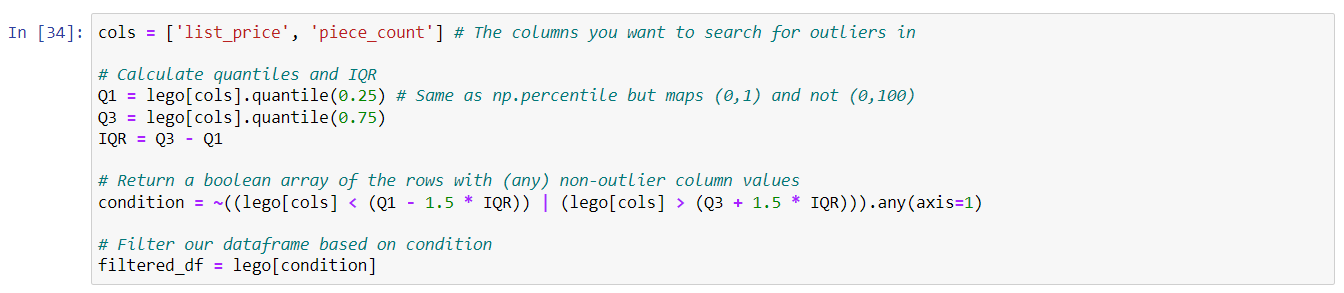


Column 2 () and 4 () contain outliers. Are they influential outliers? Or should I keep them?

I proceeded to determine the standardised residuals and the H leverage. I know that the X-axis outliers might significantly influence my linear regression.

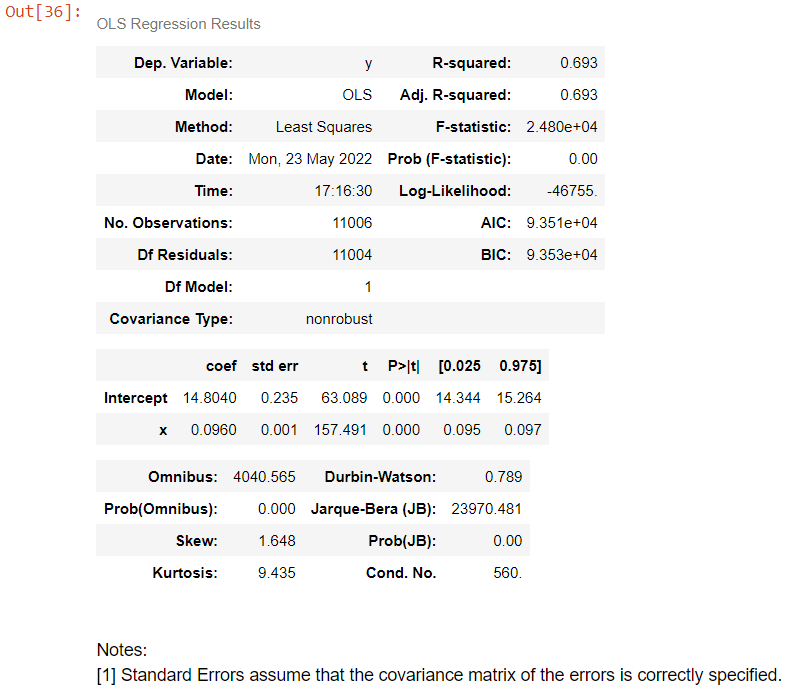


I was still not sure whether to keep the outliers or not. I, therefore, created a second DataFrame without outliers:



I then created OLS regression, and I compared the results with the two DataFrames (with outliers vs without outliers).

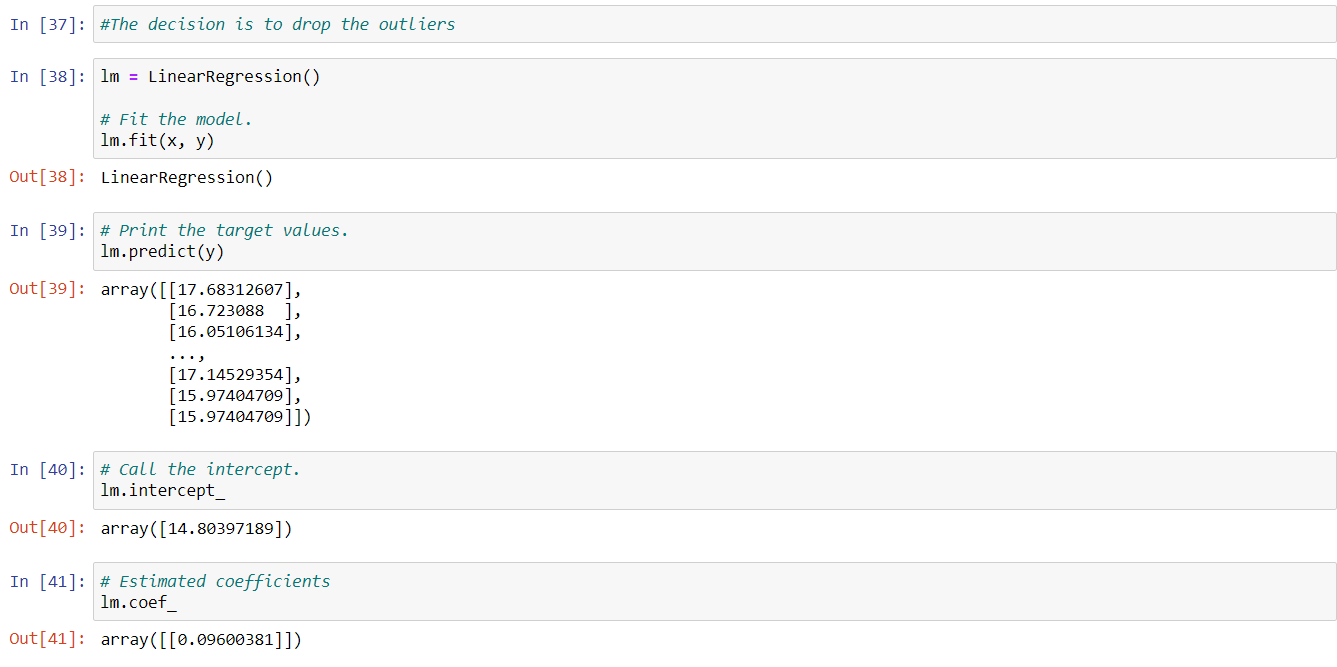


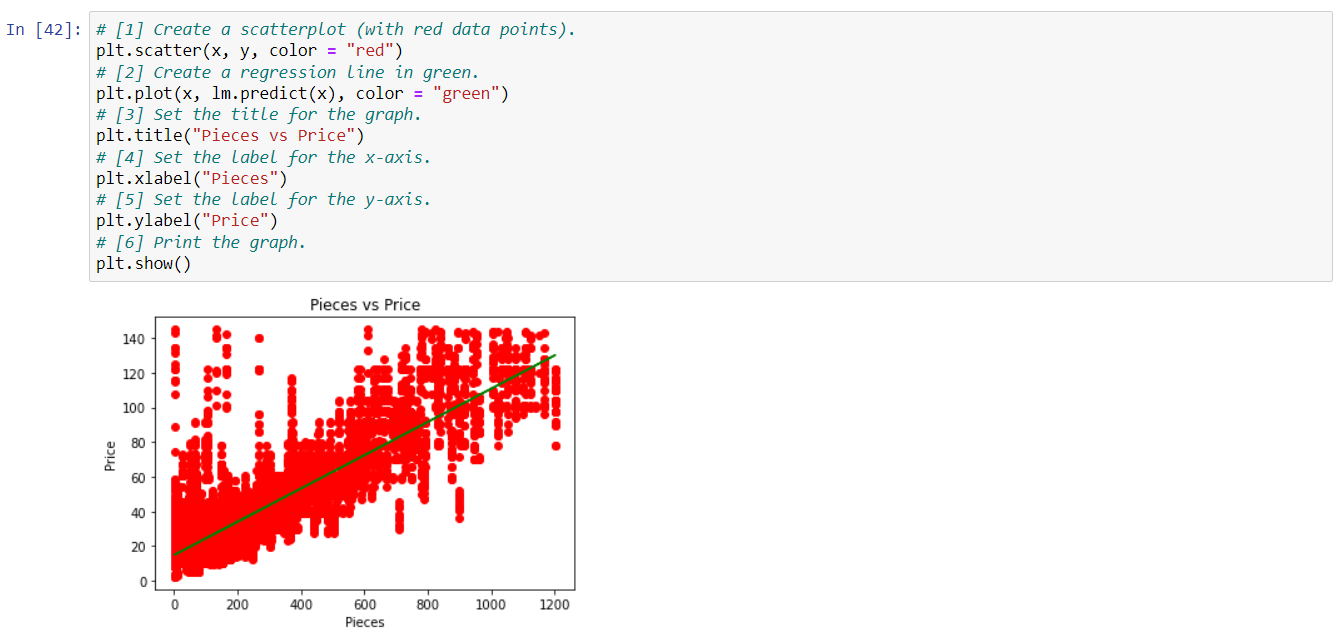


R-squared is smaller compared to before, but Note [2] disappeared.

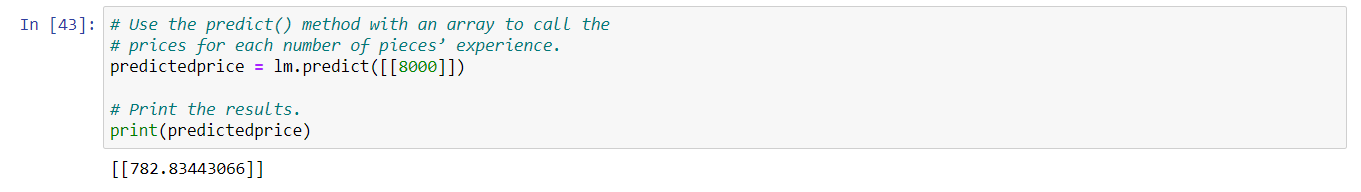
I decided to drop the outliers.

I then proceeded to calculate the linear regression.





**Response to Analysis question n. 1:**

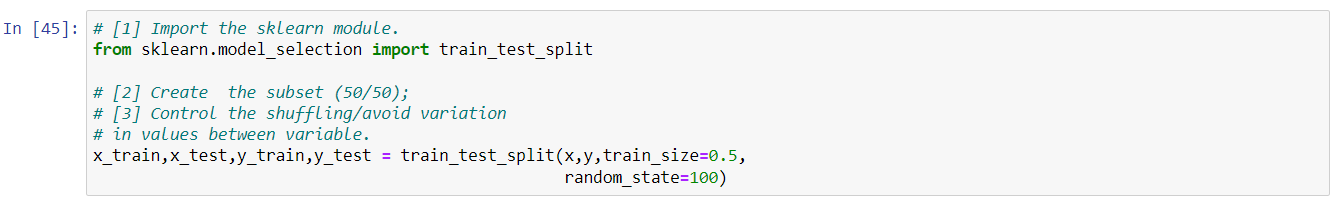


Lego sets that have 8000 Lego pieces should be sold at $783.

**Testing the model:**

I imported the sklearn module.

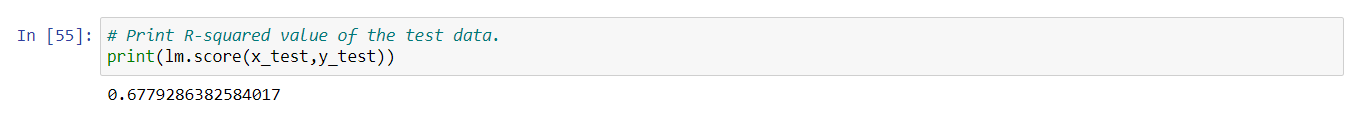
I split my dataset into two equal parts (train + test).



In a few lines of codes, I got the following result:



The R-squared value of the test data is close to the ones of the trained data.



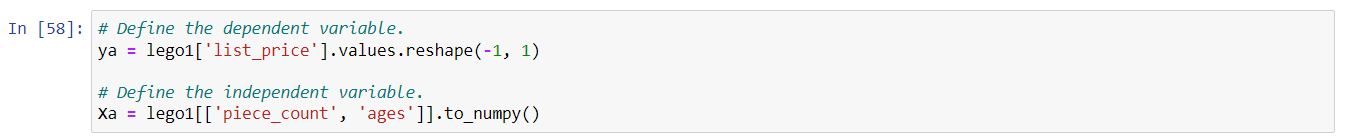
The intercept value and Coefficient value are also close to the trained data.



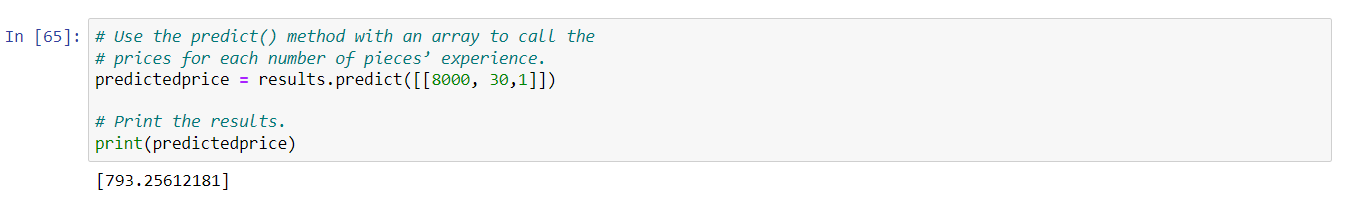
I can consider my model trustworthy.

**Analysis question n. 2:**

What price should be set for all the Lego sets that have 8000 Lego pieces and are most likely to be purchased by customers who are 30 years old?







**Response to Analysis question n. 2:**

For all the Lego sets that have 8000 Lego pieces and are most likely to be purchased by customers who are 30 years old, a price of $793 should be set.

**Analysis question n. 3:**

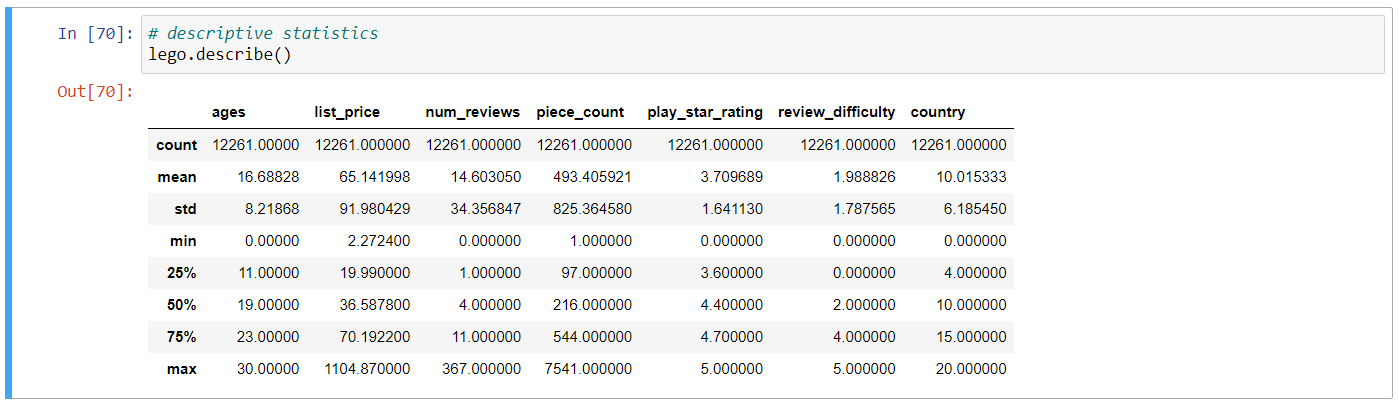
How many product segments should be utilised in sales and marketing strategies?



Determining the null values

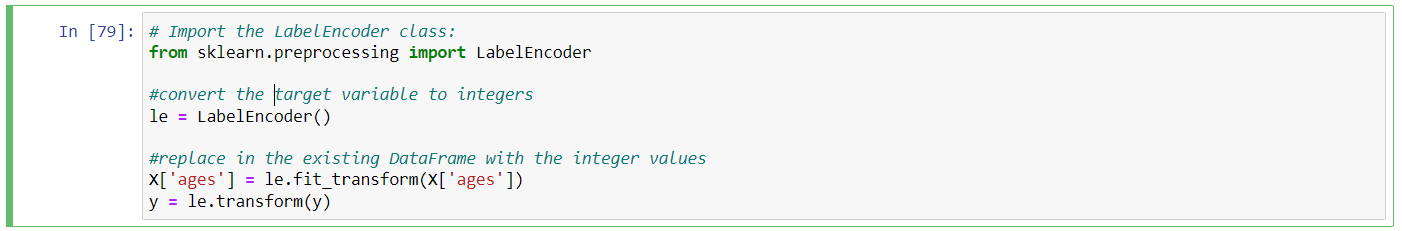


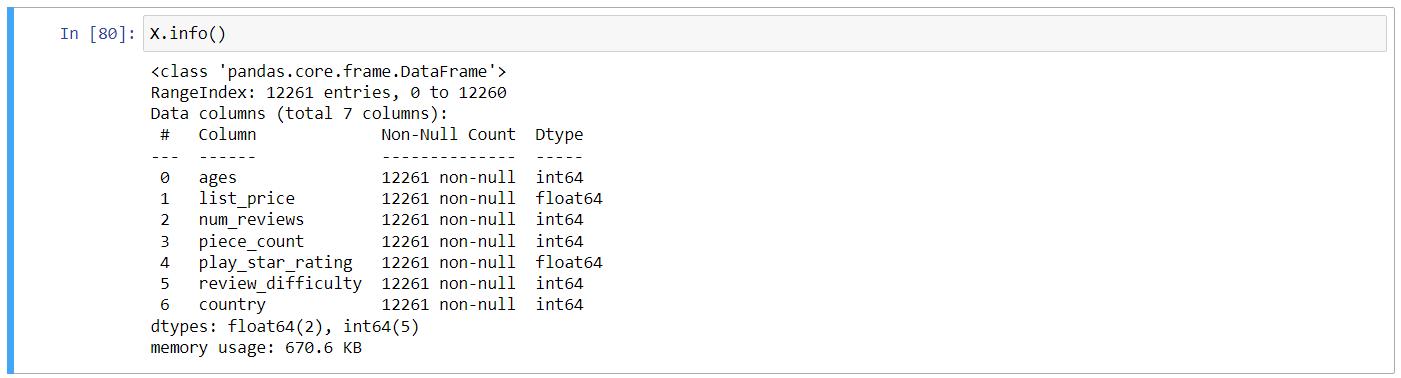
Performing descriptive statistics

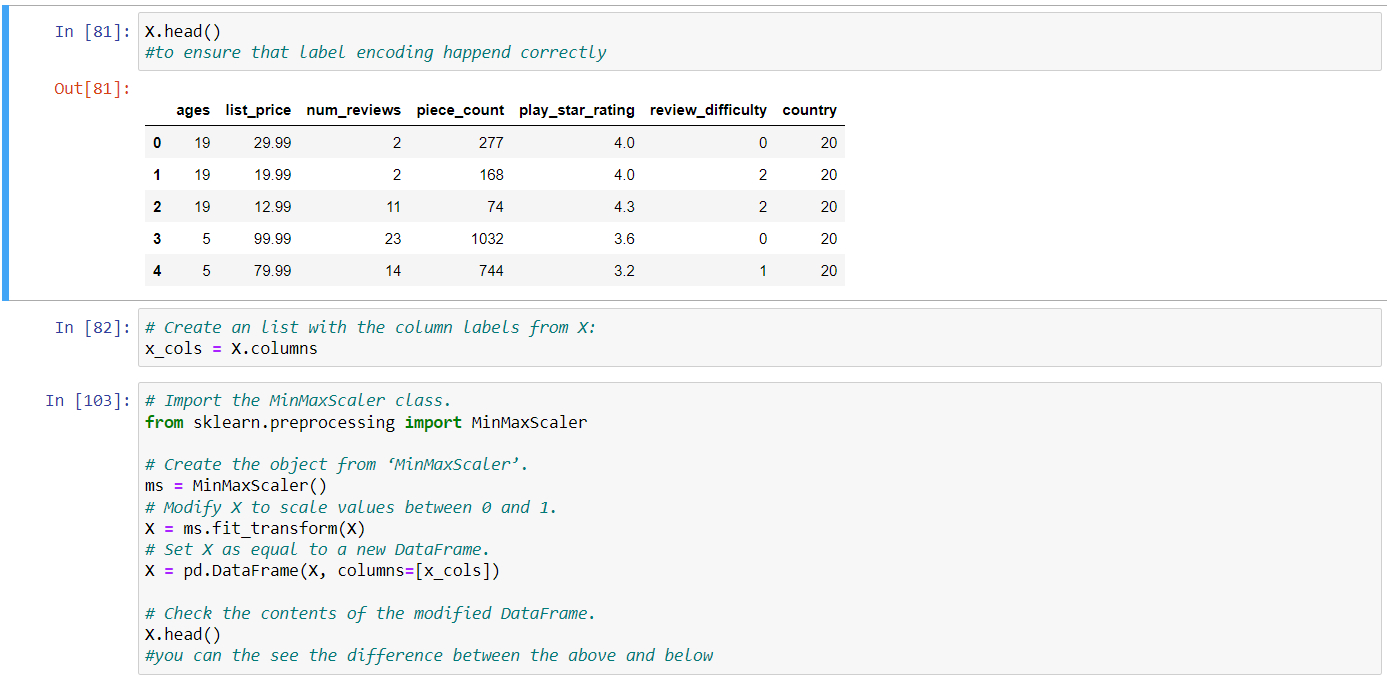


Defining the target variable

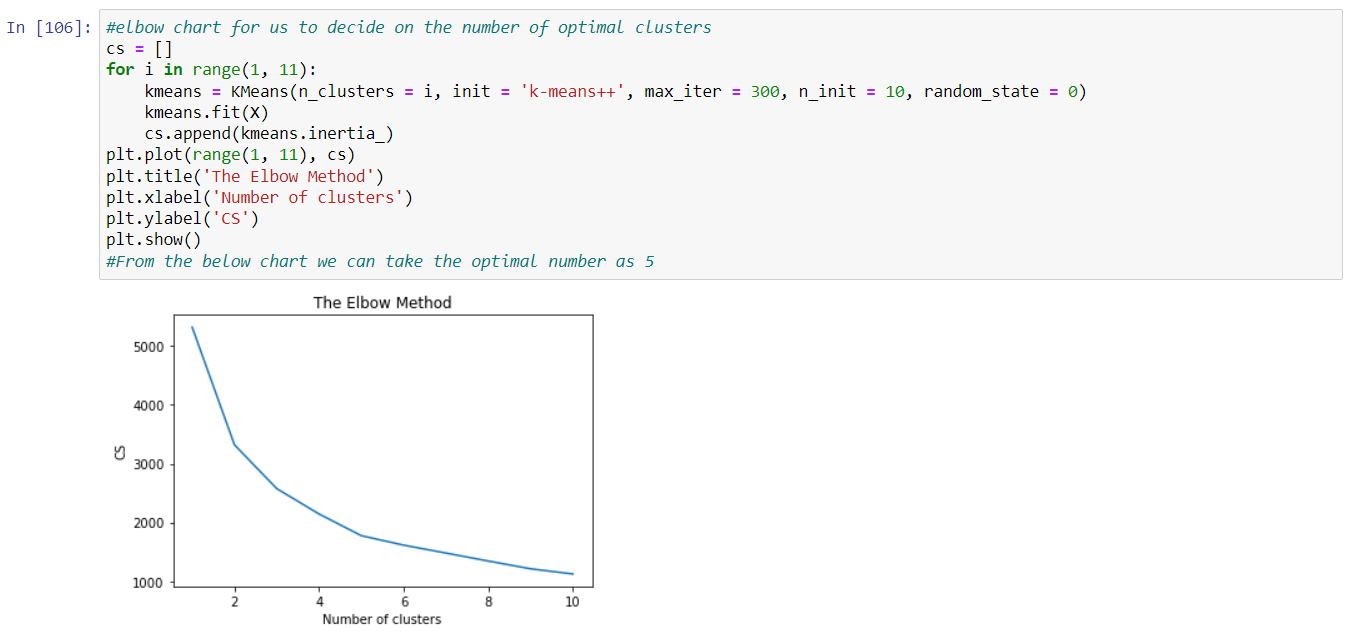








Using the elbow method to plot the variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.

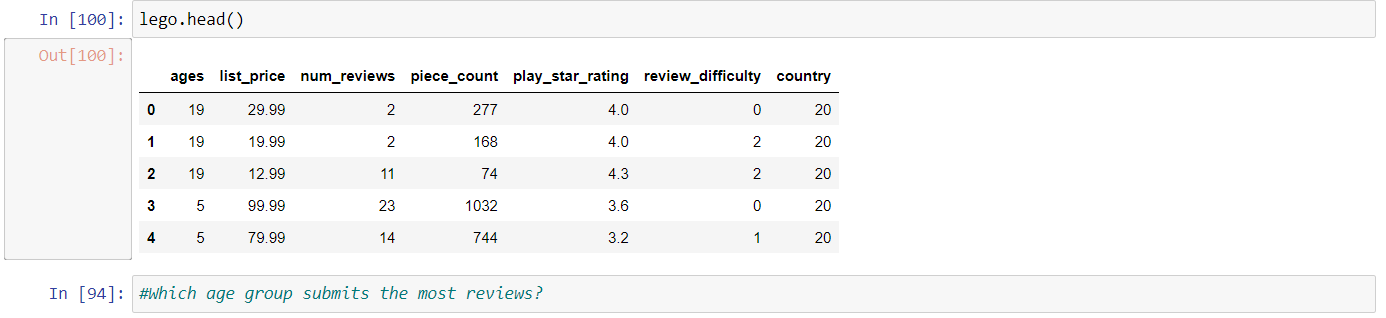


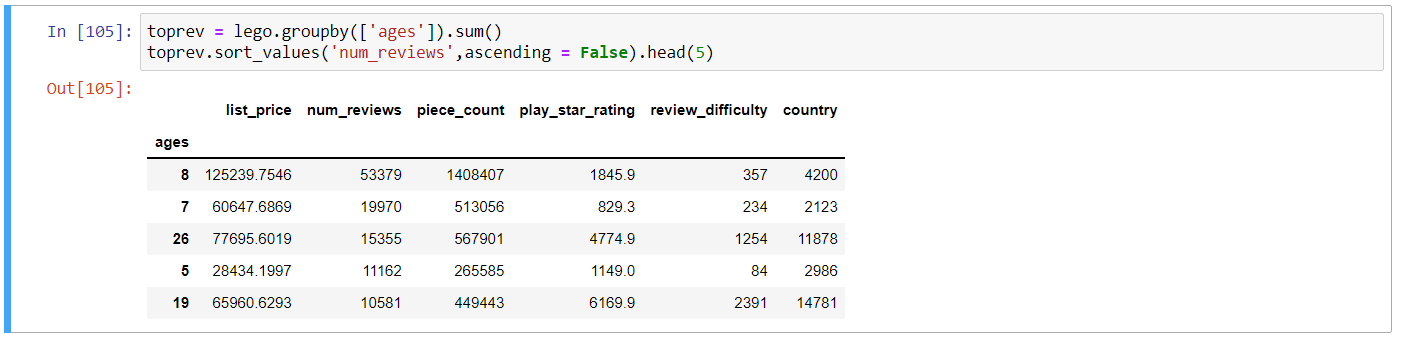
**Response to Analysis question n. 3:**

5 product segments should be utilised in sales and marketing strategies.

**Analysis question n. 4:**

Which age group submits the most reviews?



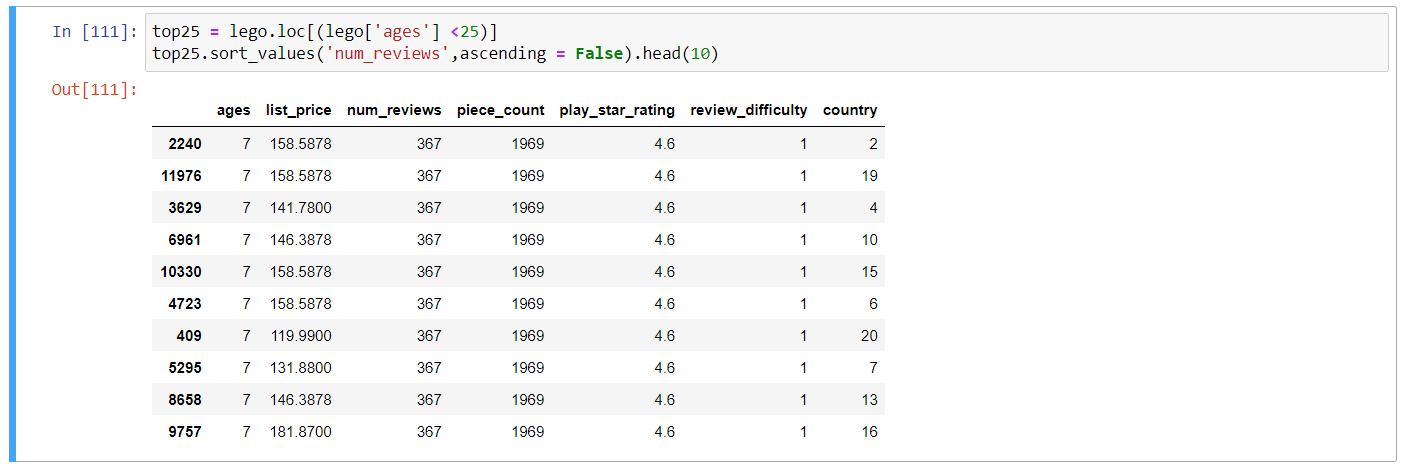


**Response to Analysis question n.4:**

The top 5 age groups for the number of reviews are 8, 7, 26, 5 and 19.

**Analysis question n.5:**

Which are the most popular (i.e. with the most number of reviews) Lego sets purchased by customers who are at the most 25 years old (<25 years)?



**Response to Analysis question n.5:**

The most popular Lego sets purchased by customers who are at most 25 years old are the 1696 pieces Lego sets.

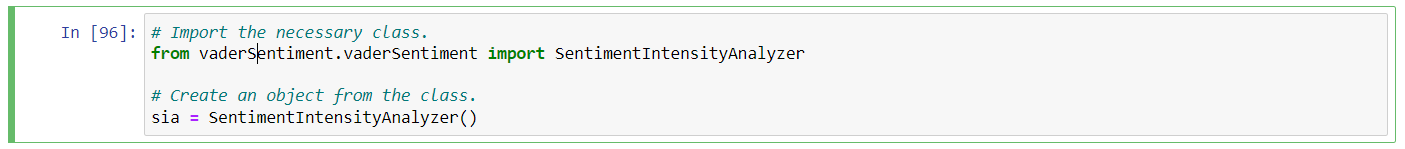
**Analysis question n.6:**

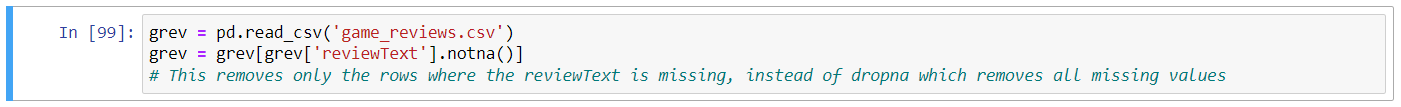
Based on the polarity of the sentiment, what are the top 20 positive and top 20 negative reviews?

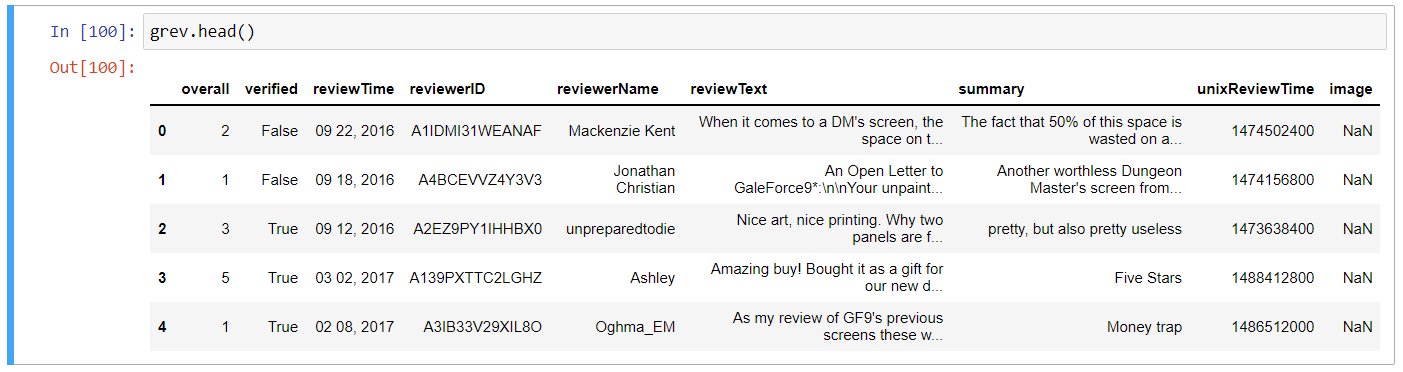
Let’s now proceed with the sentiment analysis.

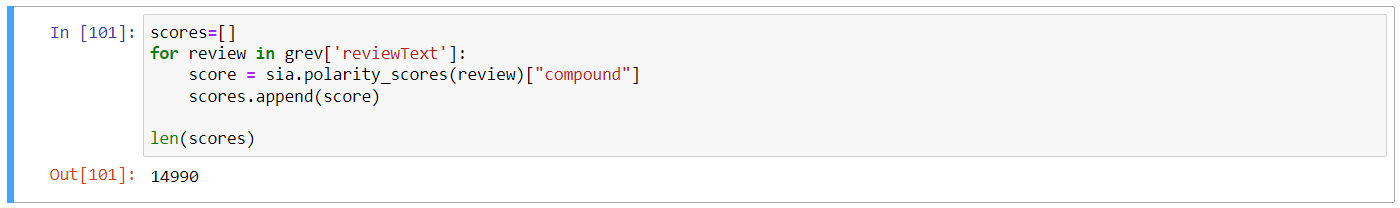
“Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations.” (towards data science).



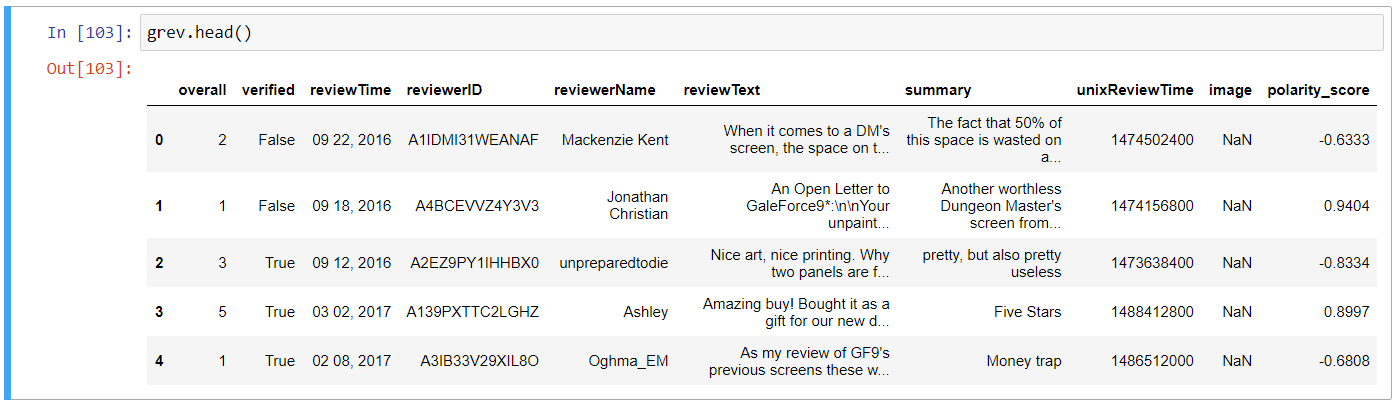




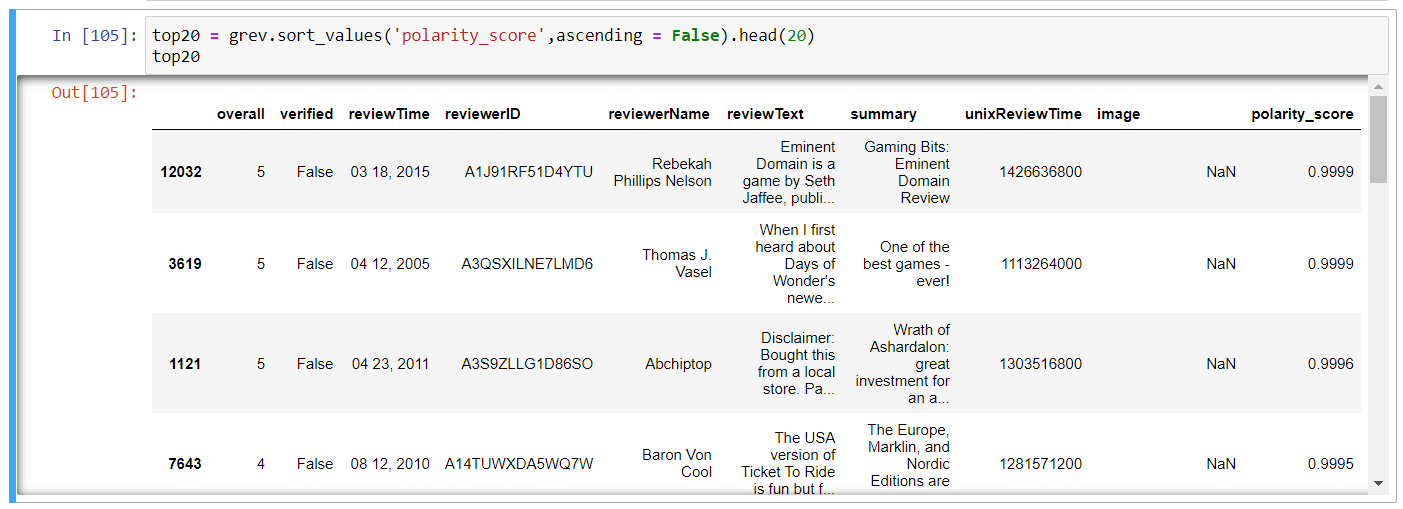


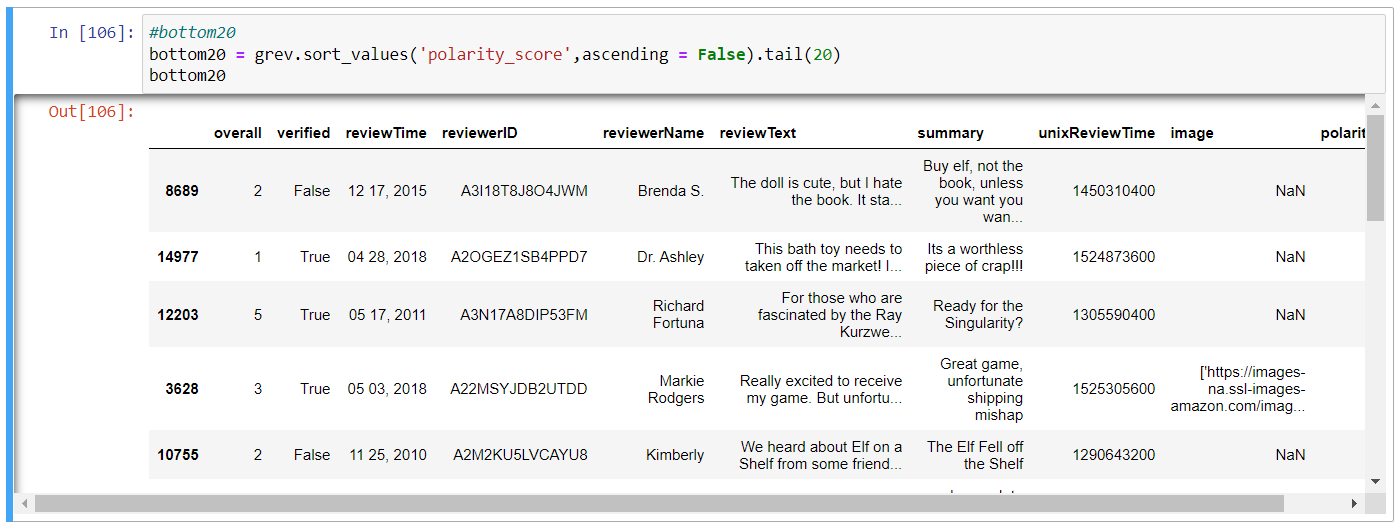






**Response to Analysis question n.6:**





See the <https://github.com/Pota1987/Data-Analytics-LSE> to visualize the complete lists.