**Srilasya Garigipaty**

**Web Scraping Class Lab Activity**

**Amazon Bestsellers**

**DATA 1202**

**Student # 100822953**

**Importing necessary libraries**

URL used: https://www.amazon.in/gp/bestsellers/books/ref=zg\_bs\_pg\_



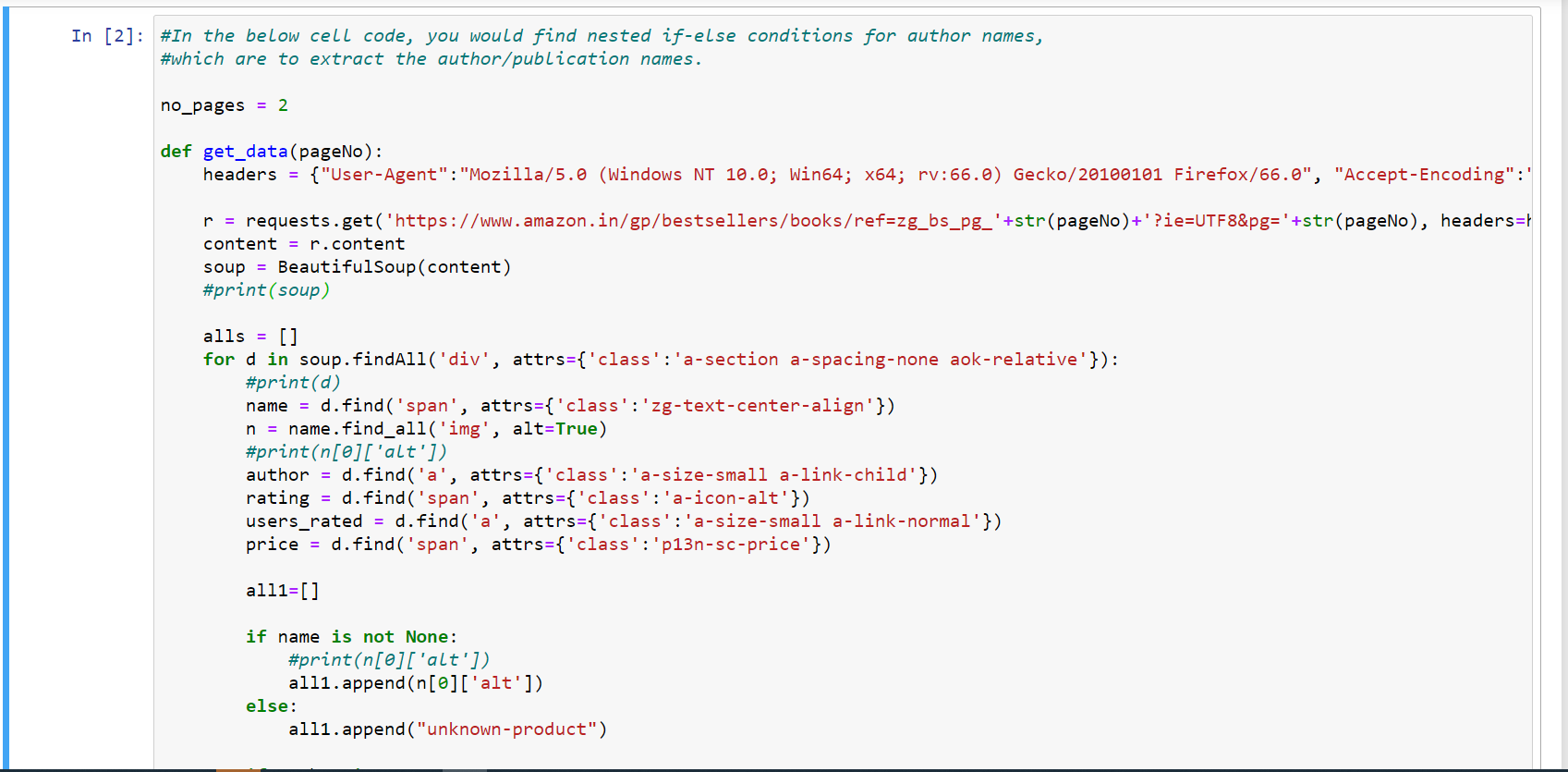
• Define a get data function which will input the page numbers as an argument,

• Define a user-agent which will help in bypassing the detection as a scraper,

• Specify the URL to requests.get and pass the user-agent header as an argument,

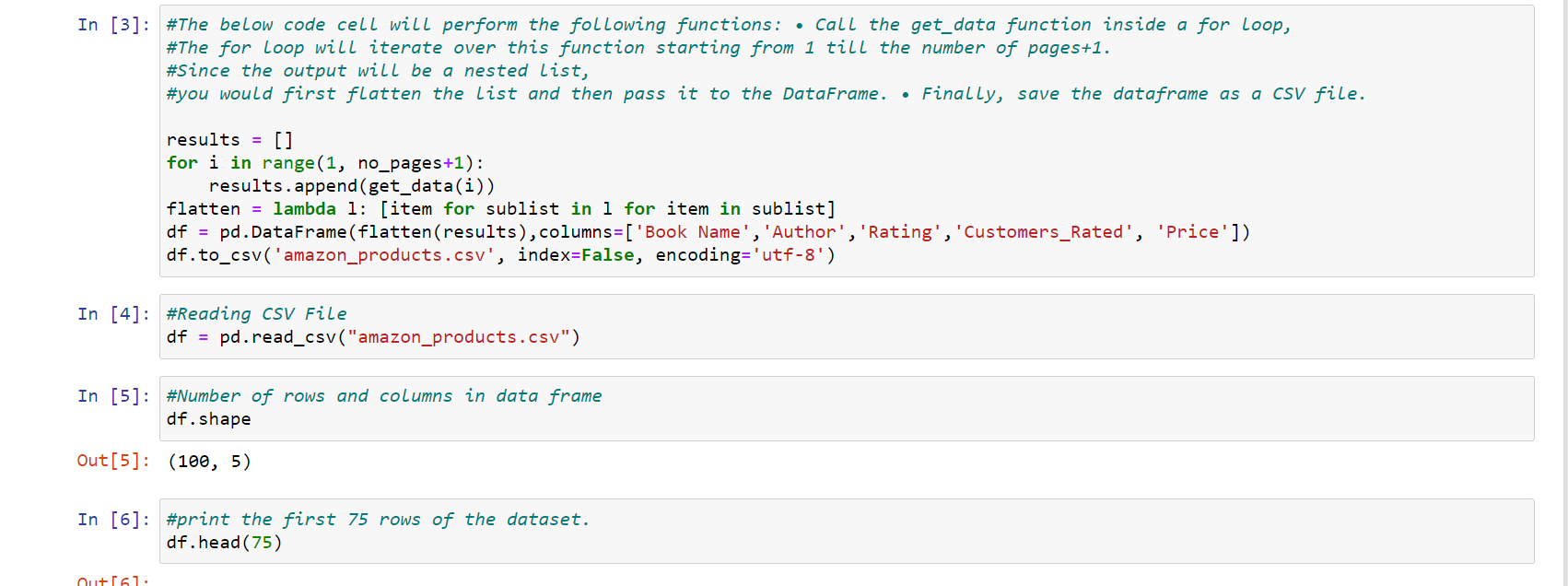
• Extract the content from requests.get,

• Scrape the specified page and assign it to soup variable,





* Call the get\_data function inside a for loop,
* • The for loop will iterate over this function starting from 1 till the number of pages+1.
* • Since the output will be a nested list, you would first flatten the list and then pass it to the DataFrame.
* • Finally, save the DataFrame as a CSV file.





**some preprocessing on the ratings, customers rated, and price column.**

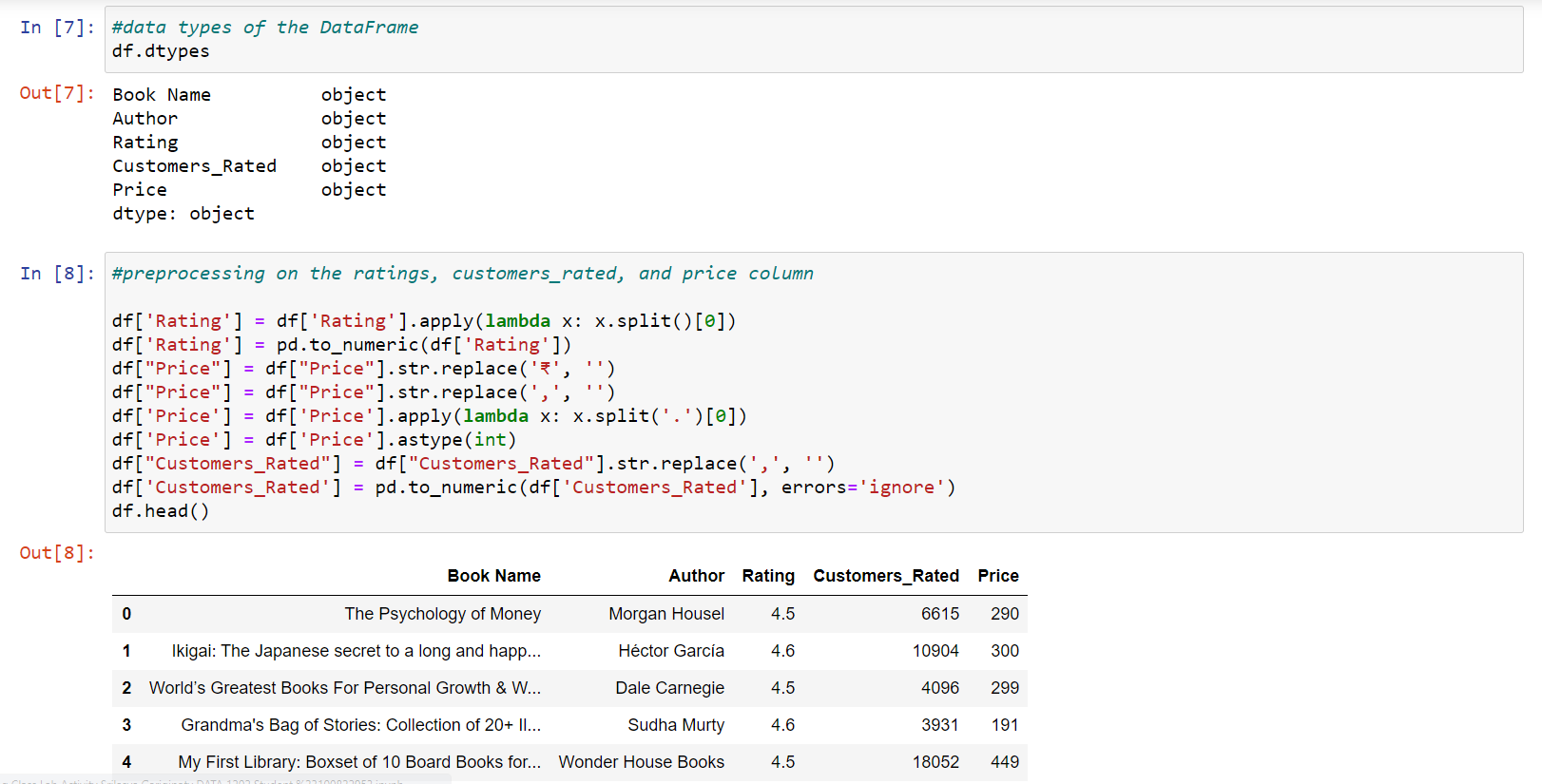
Since the ratings are out of 5, you can keep only the rating and remove the extra part from it.

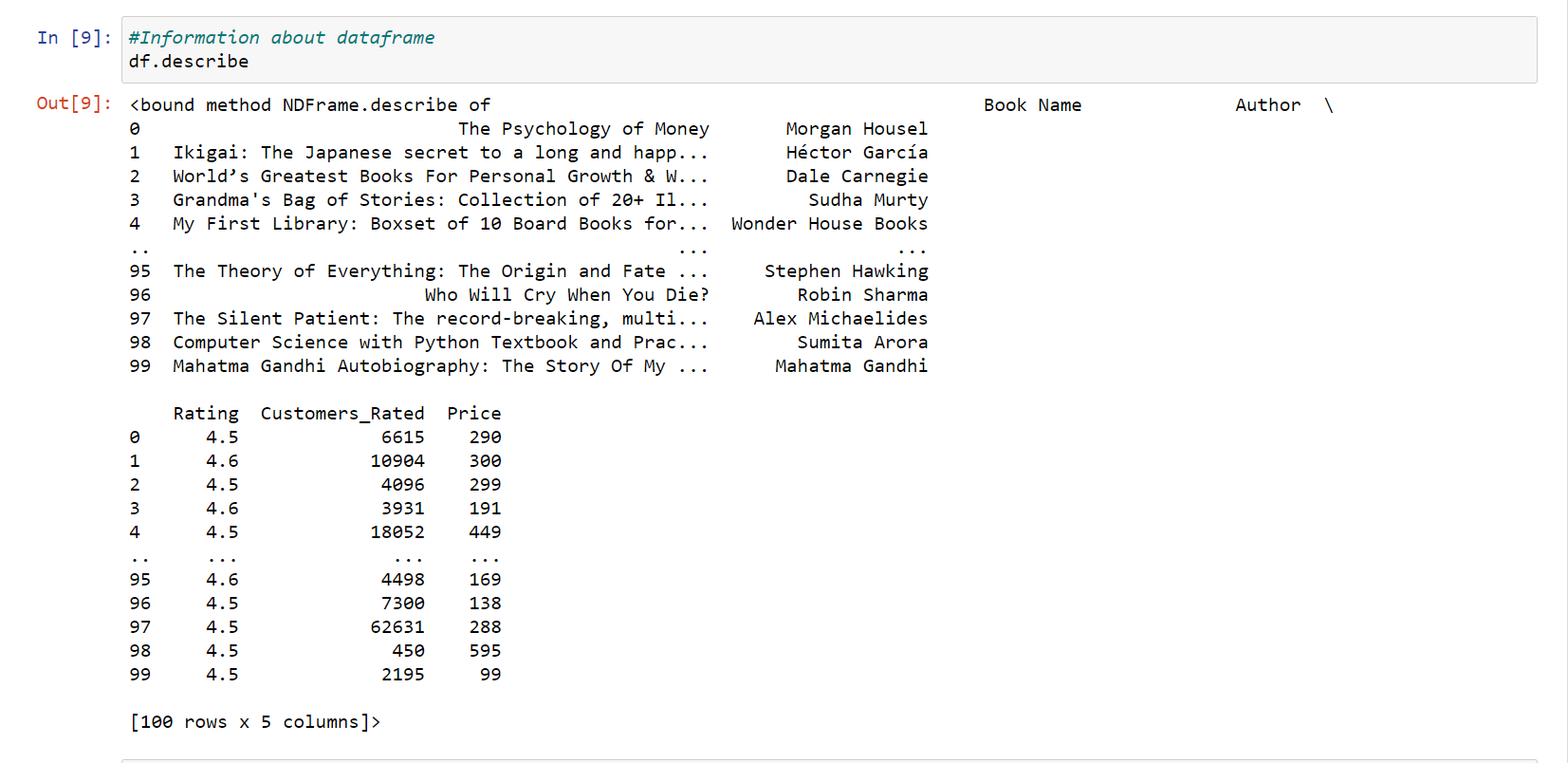
• From the customers rated column, remove the comma.

• From the price column, remove the rupees symbol, comma, and split it by dot.

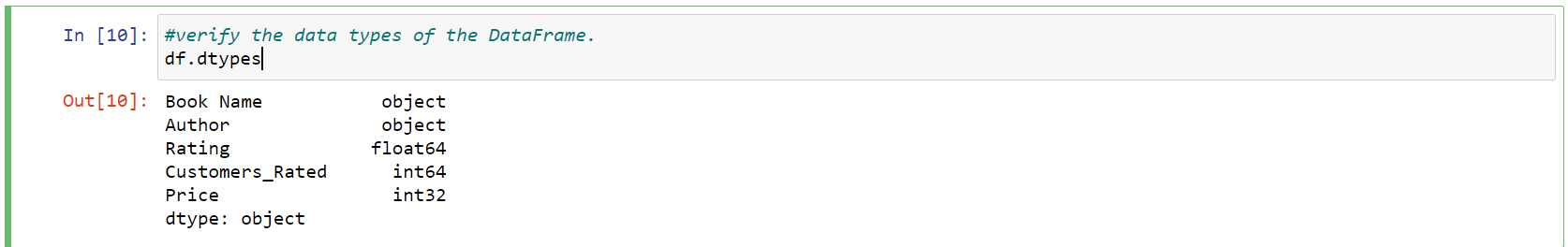
• Finally, convert all the three columns into integer or float.

* verify the data types of the Data Frame.

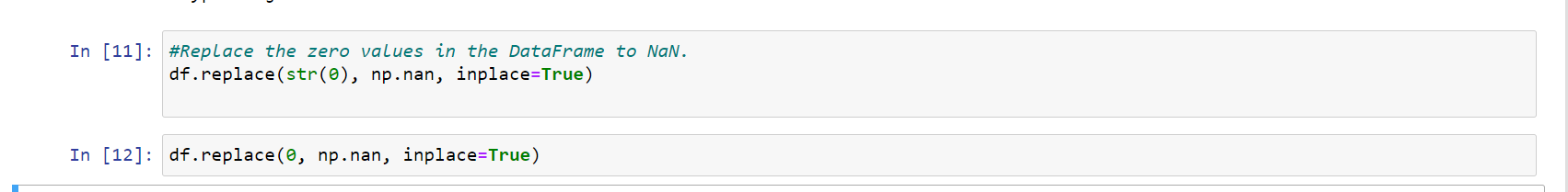




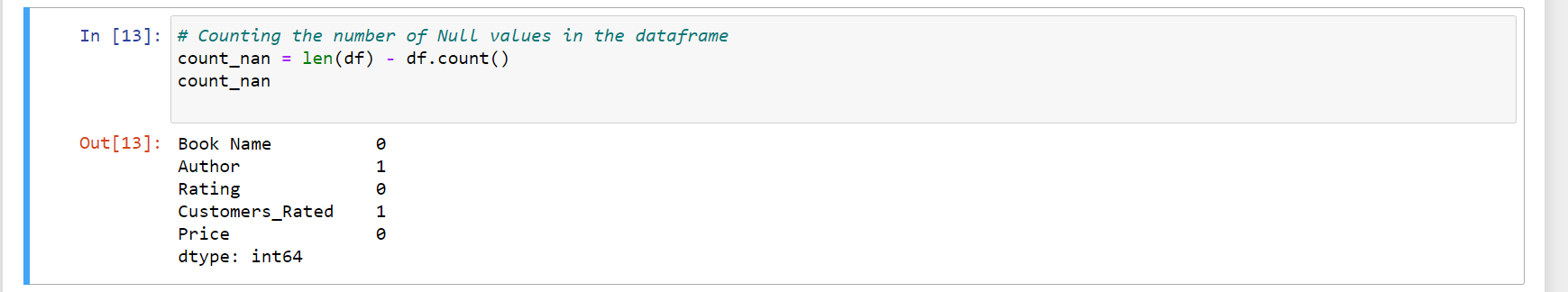
* Verify the data types of the Data Frame.



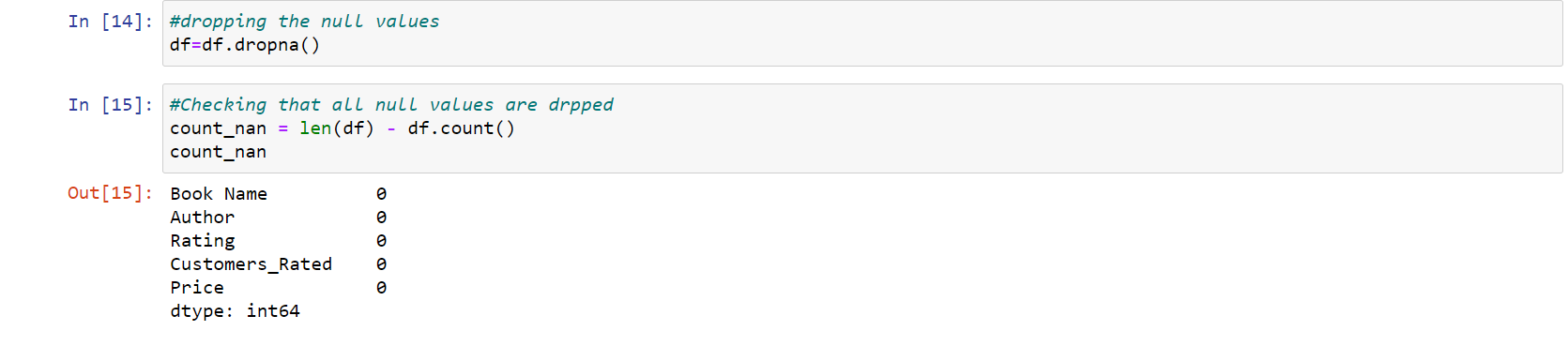
* Replace the zero values in the DataFrame to NaN.

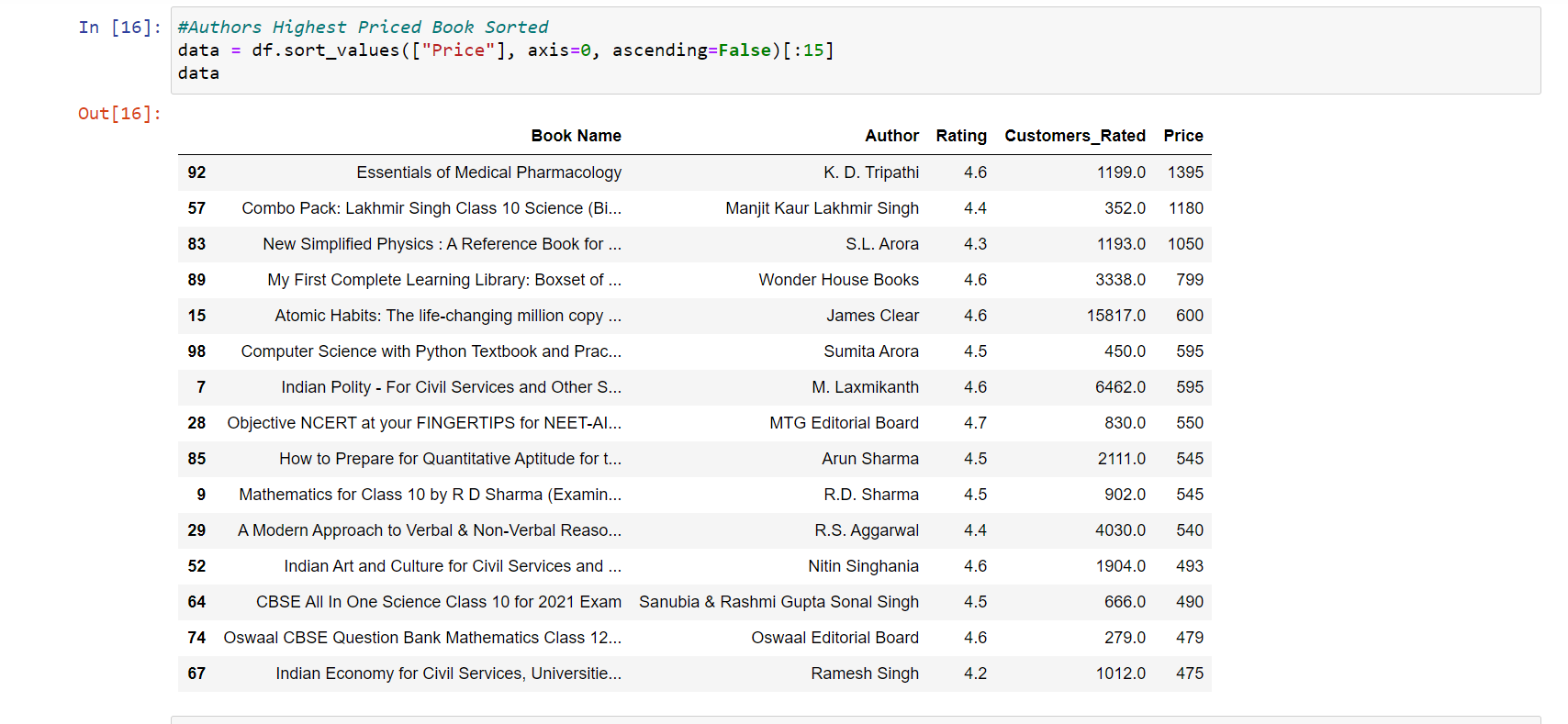


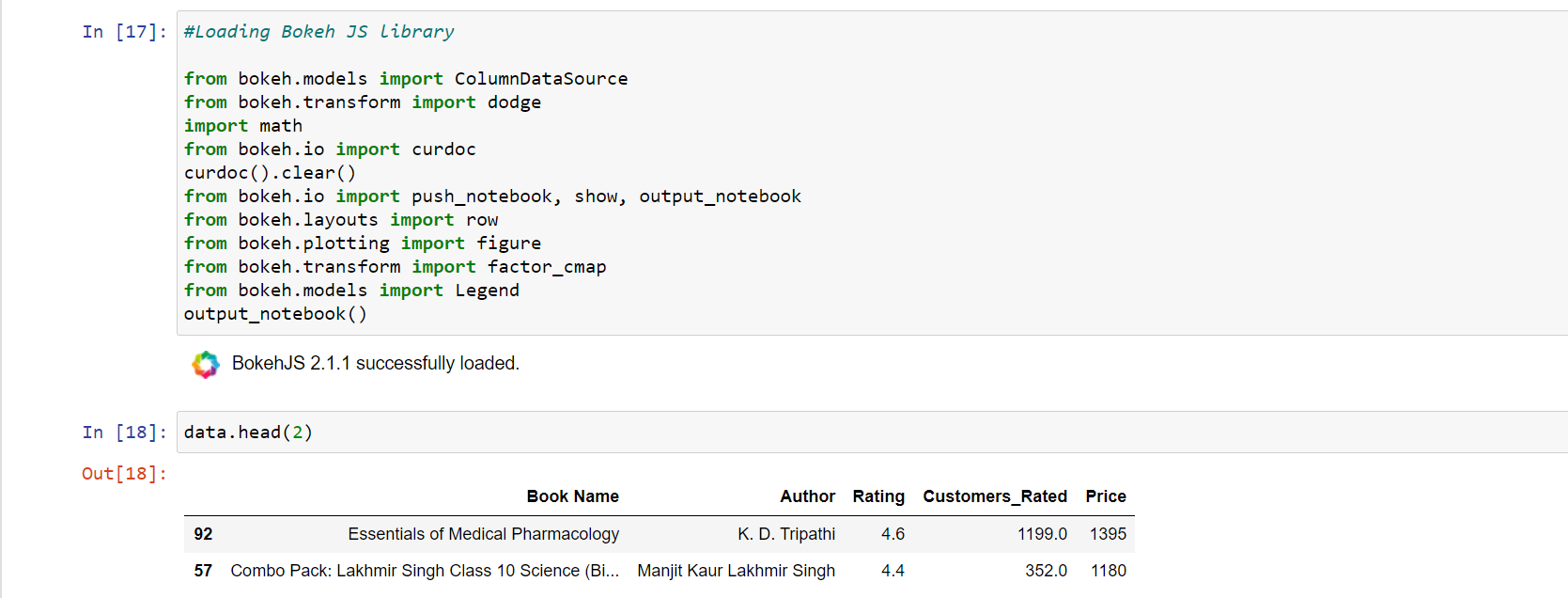
* Counting the Number of NaNs in the Data Frame

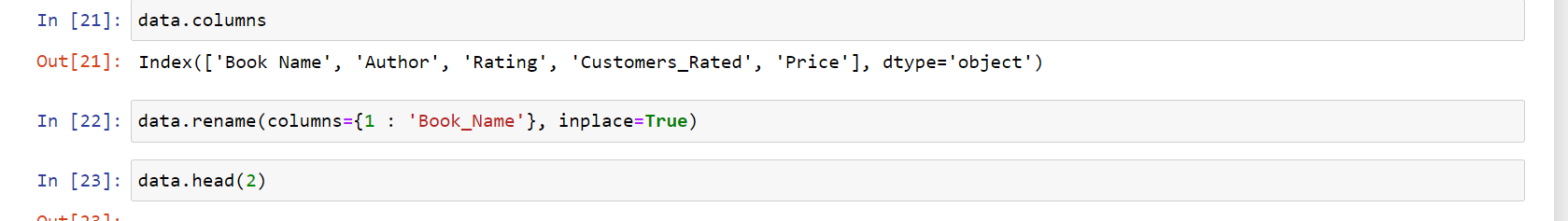


* Drop these NaNs.

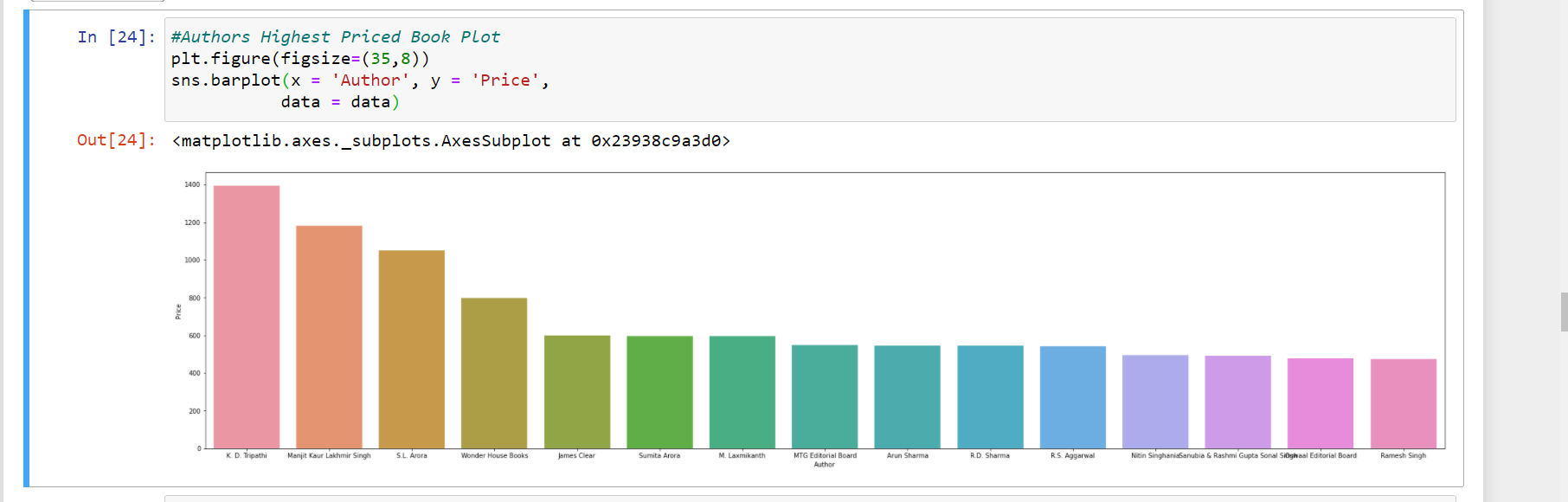


* Find out which all authors had the highest-priced book. Will visualize the results for such the top 15 authors. 
* Loading Bokeh Library

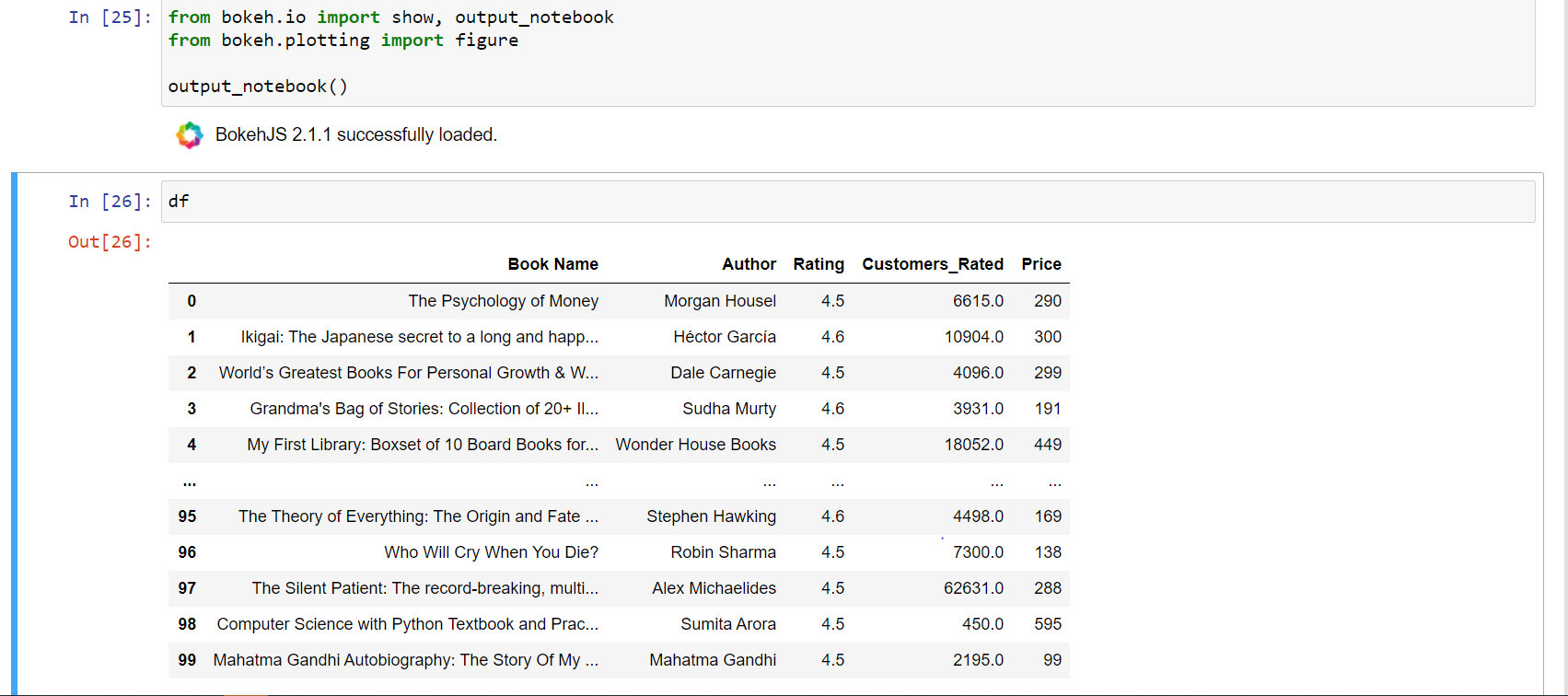




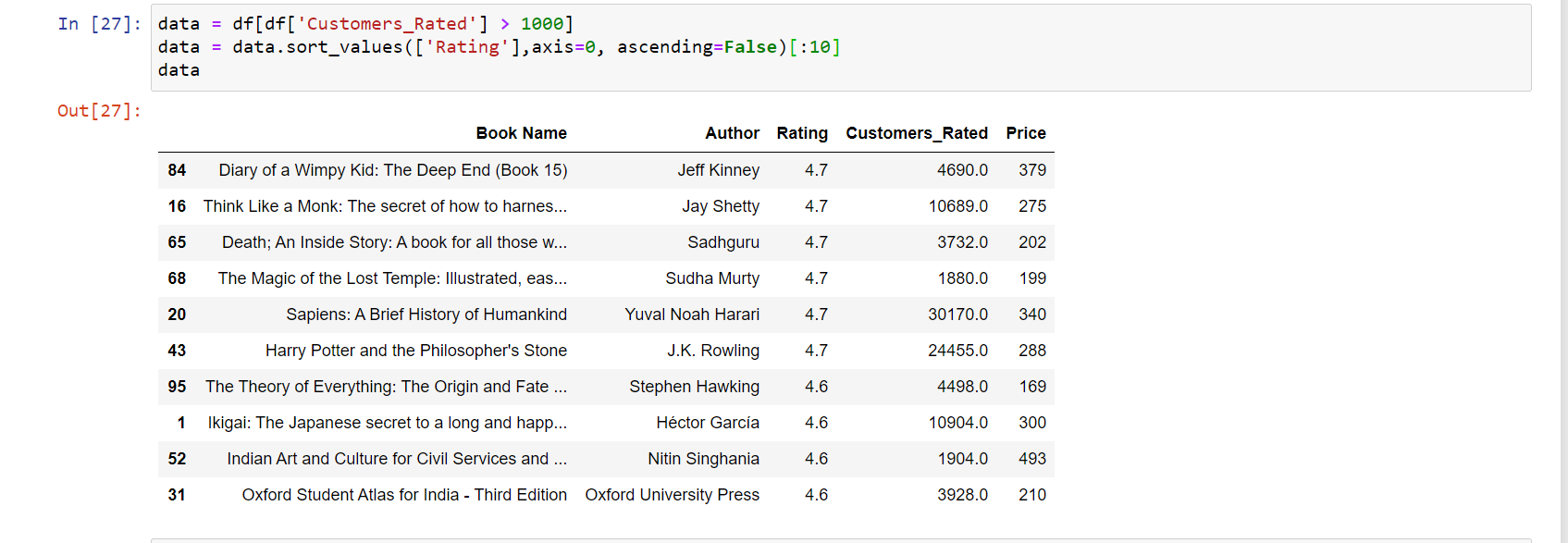
* Authors Highest Priced Book Plot



**From the above graph, you can observe that the top two highest-priced books are by the author K.D Triparthi and Manjit Kaur Lakshmi Singh and lowest price book is by Ramesh Singh.**



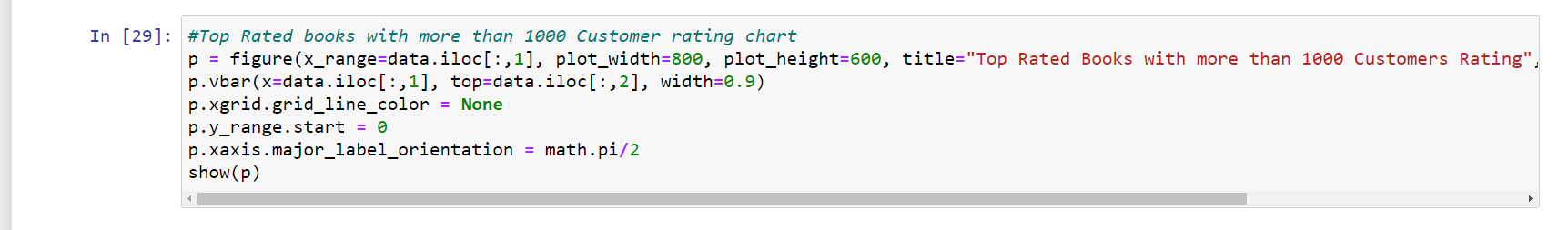
* Creating a new Data Frame with Books with more than 1000 Customer Ratings

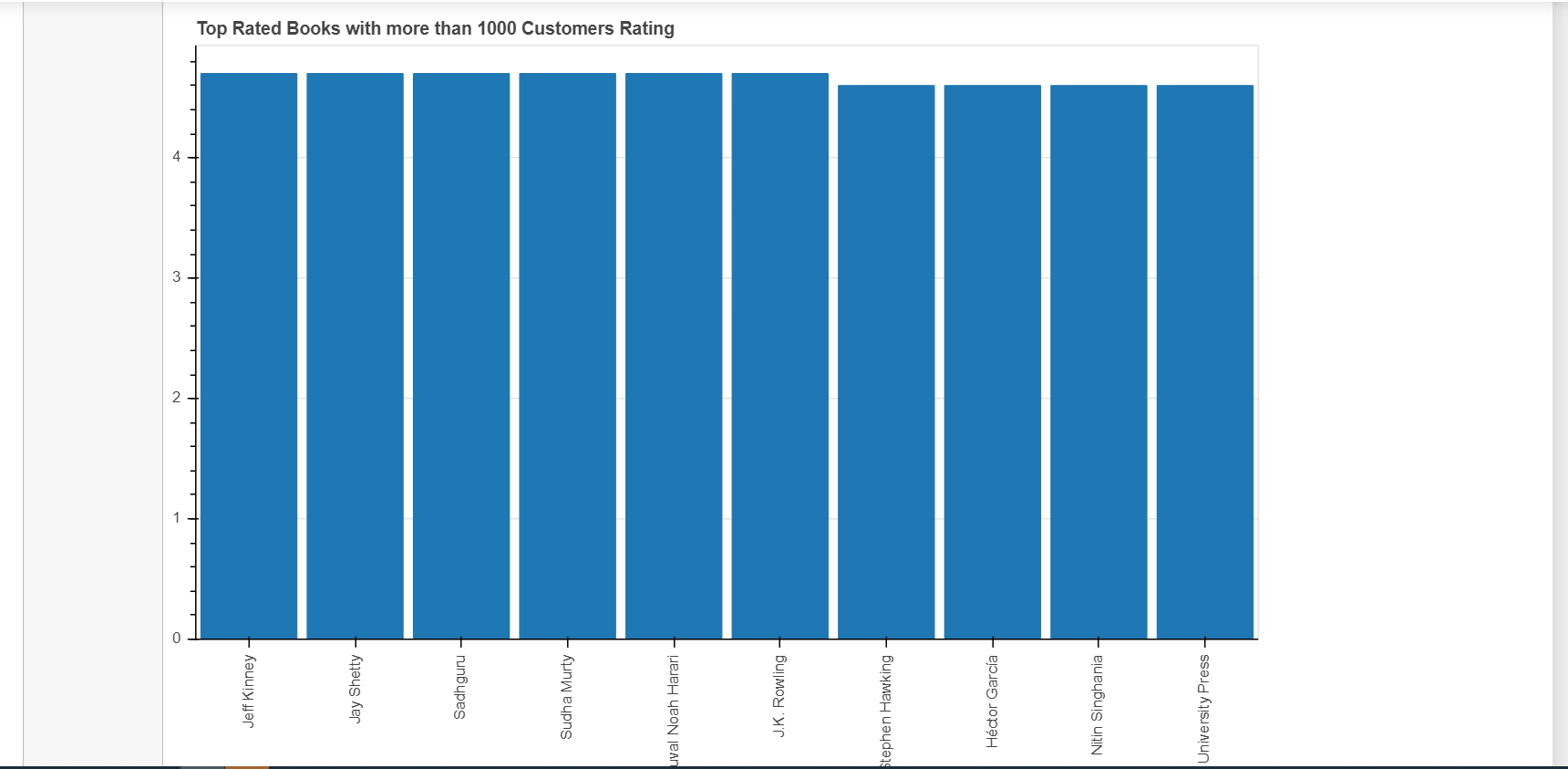


* Creating a chart with the highest priced books



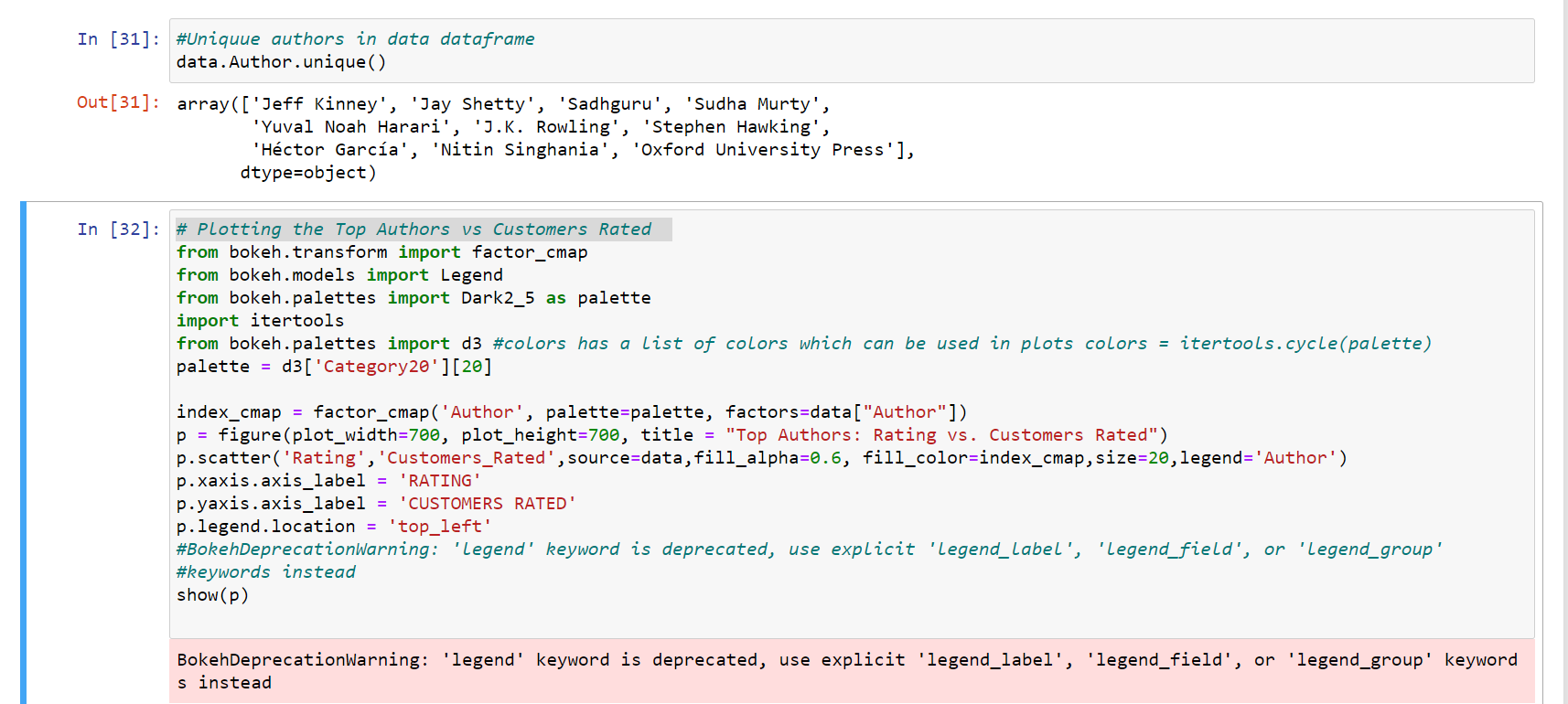
* Top Rated Books with more than 1000 Customer rating Chart





**The above graph shows the top 10 authors in descending order who have the highest rated books with more than 1000 customer ratings, which are Jeff Kinney, Jay Shetty and Sadhguru.**

* Plotting the top-rated Authors vs Customer Rated



­­­­ 

* **The above graph is a scatter plot of Authors who bagged customer rating vs. actual rating. The following conclusions can be made after looking at the above plot.**
* **Jeff Kinney book Diary of a Wimpy Kid: The Deep End (Book 15) is the best-selling book since the rating, and the number of customers rated are both in sync.**
* **Nitin Singhania, Stephen Hawking, Oxford University Press, and Hector Garcia all had a similar Customer Rating of 4.6. Yuval Noah Harari, Jk Rowling, Jay Shetty, Sadh guru, Sudha Murty, and Jeff Kinney all had the highest customer ratings of 4.7. However, the author Jeff Kinney’s The Diary of a Wimpy Kid had the most customers rated followed by J.K Rowling’s Harry Potter. The lowest number of customers rated Nitin Singhania’s book Indian Art and Culture for Civil Services.**

**Conclusion**

This Lab utilized web scraping with beautiful soup to determine how to make sense out of the information extracted from the web by visualizing it using the bokeh plotting library.