# Comments from amazon: how people reviewed prince Harry's new book



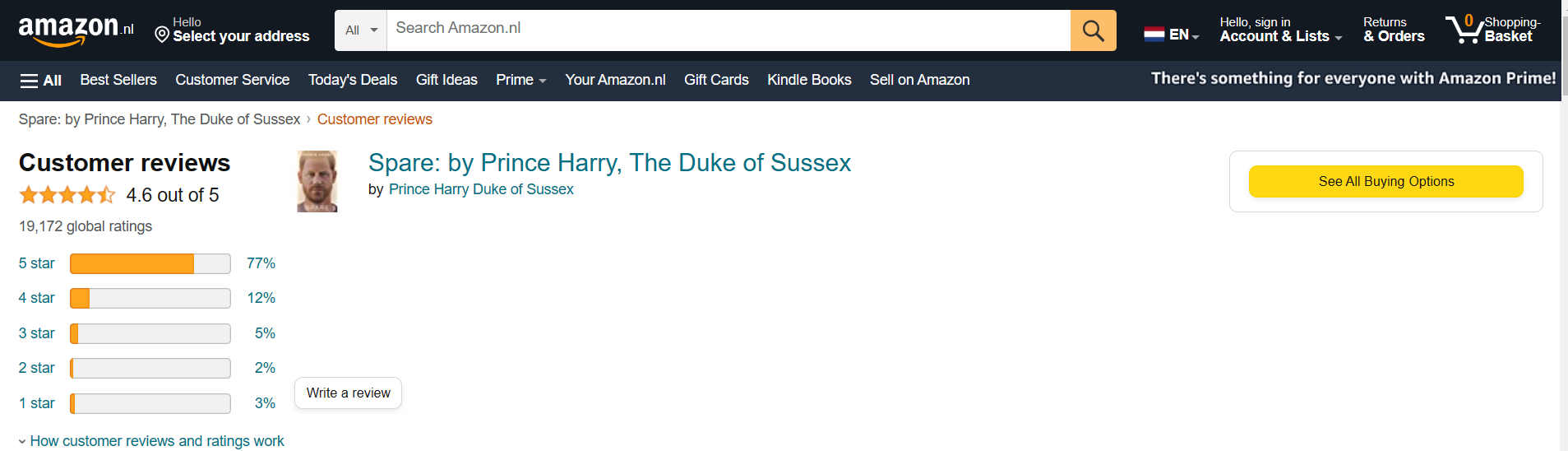
**Introduction**

Prince Harry,The Duke of Sussex and the fifth in line to the British throne, published his memoir "Spare" in October 2022, caused a worldwide storm. The image of the British royal family depicted in Harry's new book - bickering, divisive, turning on each other for their personal image and colluding with the media to slander others - shocked the outside world. The New York Times says several British royal experts say the farce has become the most serious crisis facing the British monarchy since Princess Diana's death. But for general readers, how do they feel about this book? what's the real review of the book really like? To get closer to the truth, I decided to collect comments data by web scraping through beautifulsoup4 and data analysis by pandas. The datasets are on my github.

<https://github.com/delphineliu/Collecting_Data_Individual_project.git>

At first, I choose goodreads.com to collect the comments data, but the the website is set to display only the first 300 comments, which is too small data for the data analysis to get a valid conclusion. So I choose amazon as the target with roughly 6000 informative comments.

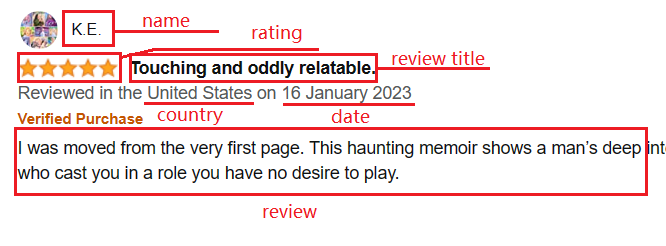
This report contains two parts: the first is web scraping the second is data cleaning and analyzing. Crawling through these reviews and simply analyzing the data will give us an idea of what readers are saying about the book and what issues are being discussed in depth.



1. web scraping

In 2019, the Ninth Circuit Court of Appeals ruled that scraping publicly accessible websites likely does not violate federal anti-hacking laws().This data collection uses Python's requests and bs4 library。

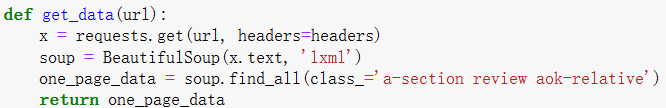
The valid data that can be scraped in Amazon reviews are reader name, rating, review content, country, time, comment, etc.



From the URL of Amazon reviews, we can see that we can crawl all the reviews by using the modification of the review page number. The crawling process learns that there are 10 reviews per page.



Firstly I defined the function get\_data() that crawls the comments section of a page.



Then, based on the analysis of the page, beautifulsoup was used to find the text part of the information to be scraped in the tags, and added to the all\_reviews list one by one. The process takes into account the anti-crawling mechanism of Amazon.com, it may disguise user-agent and other headers information. Despite this, in the process of scraping may also encounter the situation of not crawling data, then can make it rest for 5 seconds and scrap again, until the data is collected 

Finally, the DataFrame of pandas is used to save the crawled review data. The data are reader name, time and country information , date, rating, title, and review.

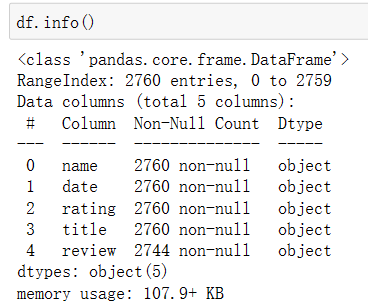


1. **Data analysis**

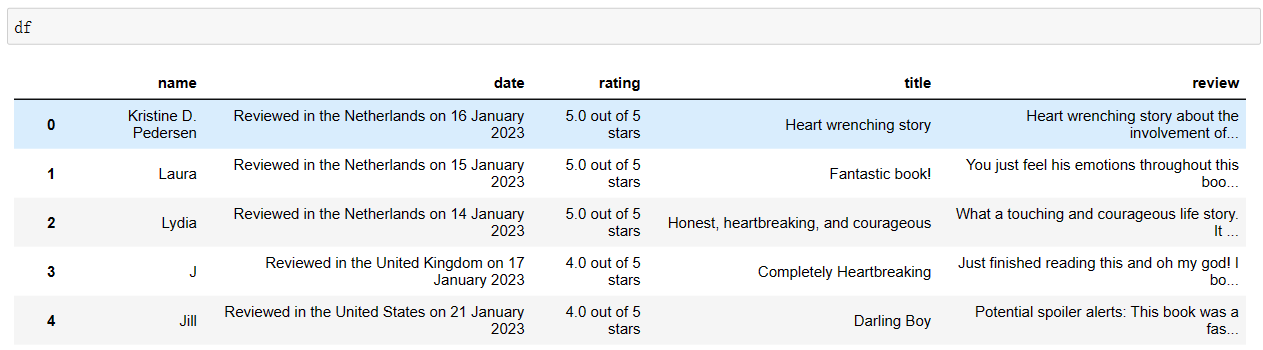
Firstly, read the saved reviews.csv file for data processing and data analysis.



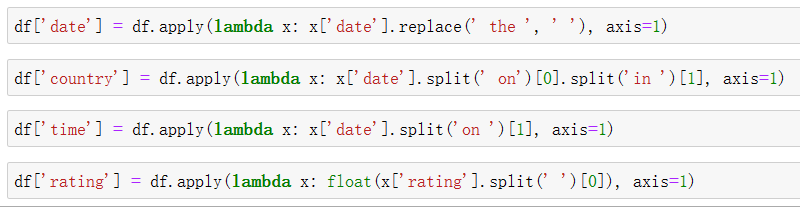
* 1. data process

Look at the data information we can see that there are missing values in the review column. so I delete the missing values.

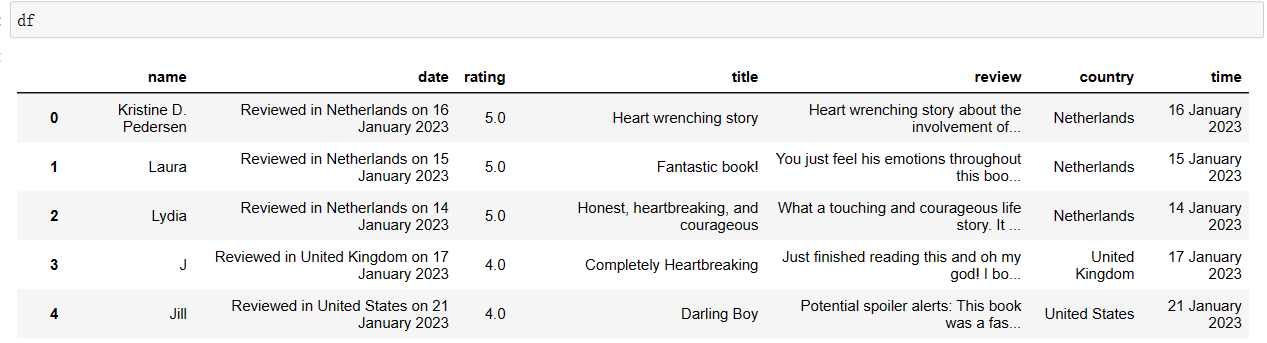


Then I found the two columns: date, rating need to be processed.

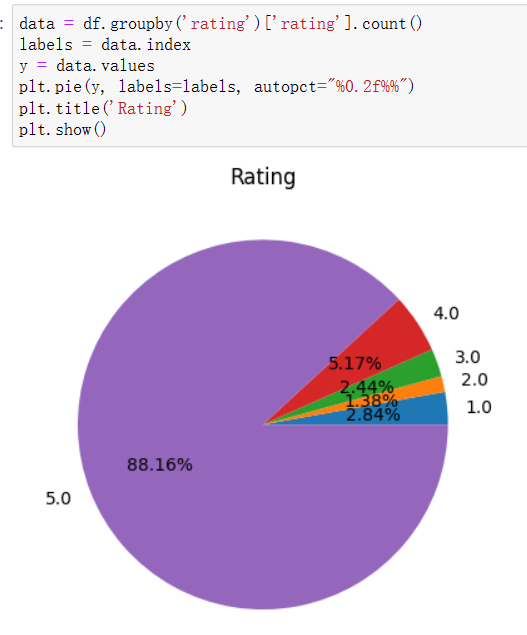
I Extracted country and date information from the date column and rating column from scores。



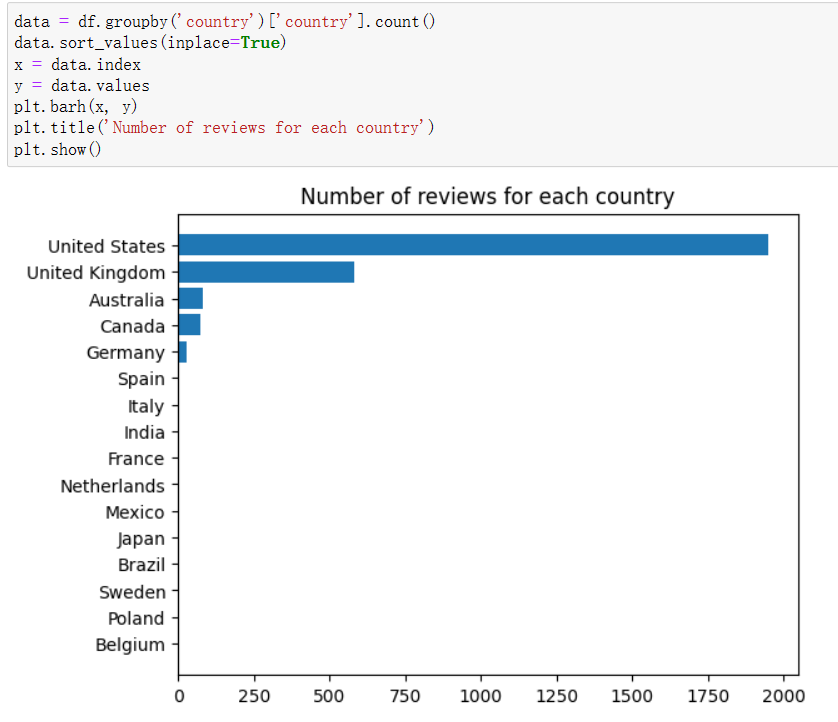
The results are as below.



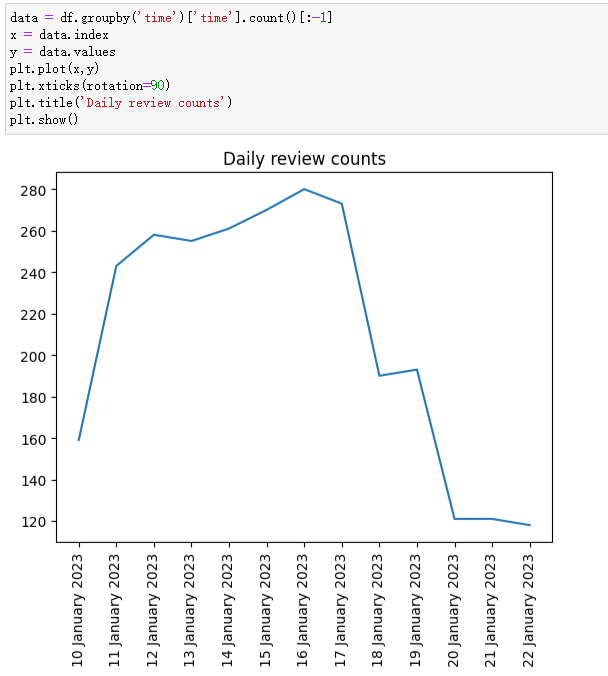
2.2 data analysis

In order to explore the datasets, data visualization tools are needed to better analyze the data. Data visualization is more intuitive and effective in presenting information about the data. There are usually many visualization tools, Seaborn, Matplotlib, and Plotly, et,. This time I used Matplotlib library for data visualization. 

The pie chart of the distribution of ratings shows that 88.16% of the readers gave 5.0, 5.17% gave 4.0, 2.44% gave 3.0, 1.38% gave 2.0, and 2.83% gave 1.0.

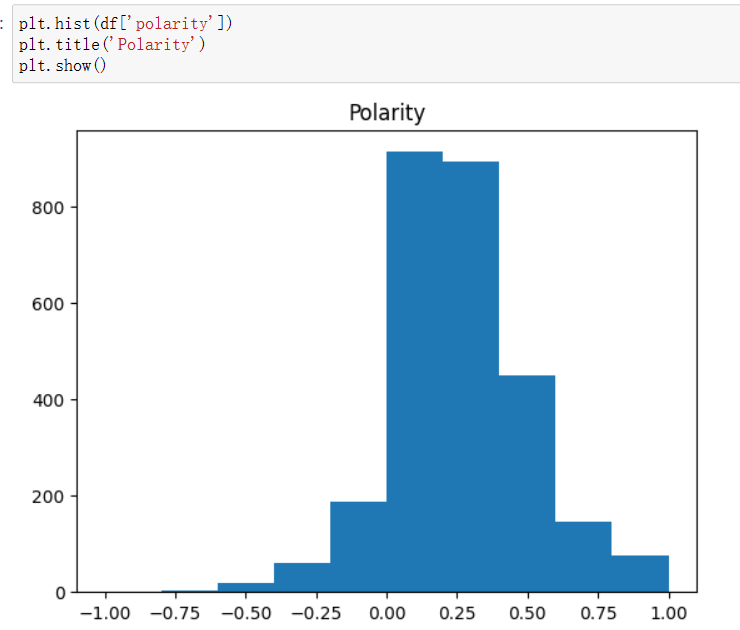


As we can see from the graph above, the largest number of readers come from the United States and the United Kingdom. Here we found an interesting result, the readers from America are much more interested in this book about British loyal stories rather than the native readers.



The analysis of the daily number of comments shows that the number of comments starts to get smaller and smaller as the date changes.

Finally, I use the textblob library to perform sentiment polarity analysis on each comment.



As we can see, the the comments are basically inclined to be positive.

**Conclusion**

By collect and analysis the reviews of Spare, we can draw a conclusion that readers are mostly give this book high marks.

**Reference**

Walsh, M. (n.d.). Cultural Analytics Open Science Guide - 25  Users’ Data: Legal & Ethical Considerations. <https://groningendh.github.io/Cultural-Analytics-Open-Science-Guide/code/09-User-Ethics-Legal-Concerns.html>