E-Commerce Website Analysis

By Aaron Charles Vergara

Scenario

An e-commerce company is selling books of various genres. The company is trying to cut costs and focus on the top 10 genres that would be the most profitable. How will you approach this problem?

[website]

Process

I. Establishing the business problem and asking the right questions

In order to provide a well-defined analysis on this scenario, it is important that the business problem and the goal is well-defined at the start before any preparation and analysis of data is performed. Effective questions are required to establish the foundational details regarding the overall situation, the data and resources to be handled, and the stakeholders involved.

From such, I asked the following questions:

- What is the current problem?
- What are the parameter/s for success in this scenario?
- What is the available data to me? What is the history, condition, and location of this available data?
- What are the gaps or limitations existing in this data? What are the opportunities in this data for analysis?
- What would be the optimal goal or preferred outcome expected in this scenario, from the stakeholders perspective?

With the following questions, I responded with:

- The company is having a problem where the current book repertoire and its profitability does not justify the budget to maintain it, hence the need to cut down costs.

- The parameter of success involves the overall profitability of a genre.
- The available data for analysis involves all book data from the given e-commerce website. Each book consists of descriptive details such as book title, price, and stock availability, etc., while falling under a genre group.
- Limitations in data include unequal distribution of books per genre, the lack of usable values in some descriptions, such as tax, prices including/excluding tax, and number of reviews, and lack of historical data. Opportunities exist in genre, prices, ratings, and availability data.
- The optimal goal is to identify the top 10 most profitable genres to be recommended for the e-commerce's continued service.

With the given scenario, the business problem can be:

- Which book genres are the most profitable and cost-effective to maintain in service?

With the following business problem, the following hypotheses can be made:

- The most expensive genres have the most highly rated reviews.
- Genres with the least expensive books overall have lower stock available.
- The most rated genres have the least amount of book stocks.

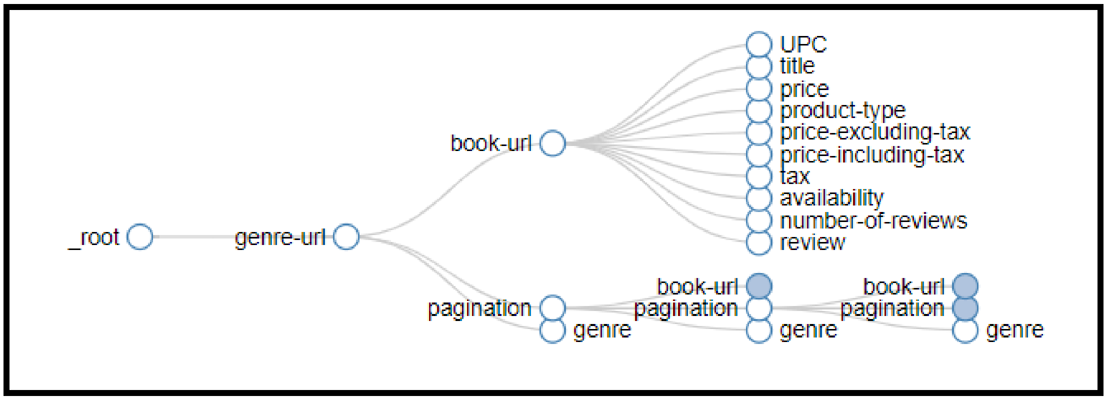
II. Preparing Data for Exploration and Processing

After establishing the business problem, objectives, and hypotheses to guide the data analysis process, it is now time to decide what data is required to answer the situation. Looking at the website, the following data are available:

- Book Title
- Book Genre
- Book Price in pound sterling (£)
- Stock Availability
- Star Rating, from 1 to 5
- Product Description
- UPC
- Product Type
- Price including and excluding tax
- Tax
- Number of Review

From this set of data, we can disregard the book description from the gathering process as the other data are more usable for quantitative analysis, as well as the genre can serve as general categorization of the information related to the description.

With the e-commerce book store website as the data source, the main process to gather the information will be through web scraping. The tool used for web scraping is webscraper.io, which is an easy and intuitive way to gather website information. The sitemap utilized for the web scraping process is as follows:

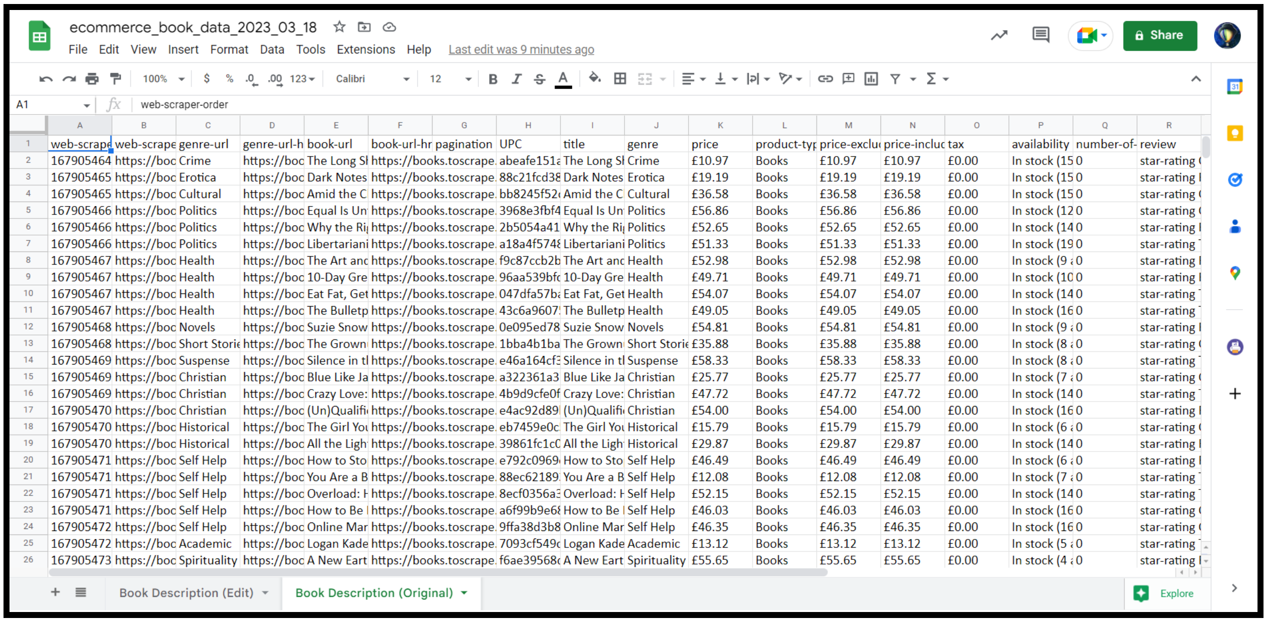


Where *genre-url* selector allows us to visit every genre available, which then consists of a recursive process per each genre section visited. Through this action, the scraper gathers the genre data the books belong to with *genre* selector, visits every book url with *book-url* selector and gathers the needed book information, and then presses the next page button with the *pagination* selector, if available, to repeat the process again. This is then saved as a .xlsx file for data processing and cleaning through a spreadsheet.

III. Data Cleaning Process

Getting into the data cleaning process, I have the following dataset, from left to right:

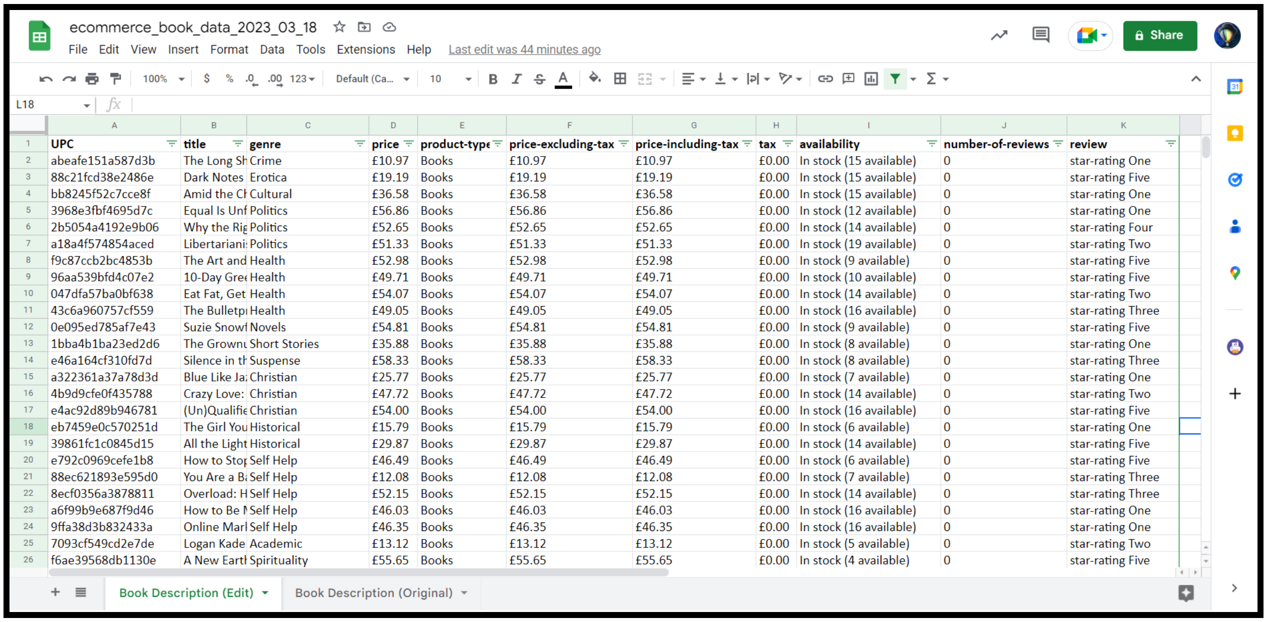
- web-scraper-order
- web-scraper-start-url
- genre-url
- genre-url-href
- book-url
- book-url-href
- pagination
- UPC
- title
- genre
- price
- product-type
- price-excluding-tax
- price-including-tax
- tax
- availability
- number-of-reviews
- review



ecommerce_book_data_2023_03_18
File Edit View Insert Format Data Tools Extensions Help
Last edit was 9 minutes ago
Share
100%
Calibri
12
A1
web-scraper-order
Book Description (Edit)
Book Description (Original)
Explore

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	web-scrape	web-scrape	genre-url	genre-url-h	book-url	book-url-hr	pagination	UPC	title	genre	price	product-typ	price-exclu	price-incluc	tax	availability	number-of-	review
2	167905464	https://boc	Crime	https://boc	The Long Sl	https://books.toscrape.		abeafe151a	The Long Sl	Crime	£10.97	Books	£10.97	£10.97	£0.00	In stock (15	0	star-rating
3	167905465	https://boc	Erotica	https://boc	Dark Notes	https://books.toscrape.		88c21fcd38	Dark Notes	Erotica	£19.19	Books	£19.19	£19.19	£0.00	In stock (15	0	star-rating
4	167905465	https://boc	Cultural	https://boc	Amid the C	https://books.toscrape.		bb8245f52	Amid the C	Cultural	£36.58	Books	£36.58	£36.58	£0.00	In stock (15	0	star-rating
5	167905466	https://boc	Politics	https://boc	Equal Is Un	https://books.toscrape.		3968e3fbf4	Equal Is Un	Politics	£56.86	Books	£56.86	£56.86	£0.00	In stock (12	0	star-rating
6	167905466	https://boc	Politics	https://boc	Why the Ri	https://books.toscrape.		2b5054a41	Why the Ri	Politics	£52.65	Books	£52.65	£52.65	£0.00	In stock (14	0	star-rating
7	167905466	https://boc	Politics	https://boc	Libertariani	https://books.toscrape.		a18a4f574	Libertariani	Politics	£51.33	Books	£51.33	£51.33	£0.00	In stock (19	0	star-rating
8	167905467	https://boc	Health	https://boc	The Art anc	https://books.toscrape.		f9c87ccb2b	The Art anc	Health	£52.98	Books	£52.98	£52.98	£0.00	In stock (9	0	star-rating
9	167905467	https://boc	Health	https://boc	10-Day Gre	https://books.toscrape.		96aa539bfc	10-Day Gre	Health	£49.71	Books	£49.71	£49.71	£0.00	In stock (10	0	star-rating
10	167905467	https://boc	Health	https://boc	Eat Fat, Get	https://books.toscrape.		047dfa57ba	Eat Fat, Get	Health	£54.07	Books	£54.07	£54.07	£0.00	In stock (14	0	star-rating
11	167905467	https://boc	Health	https://boc	The Bulletp	https://books.toscrape.		43c6a9607	The Bulletp	Health	£49.05	Books	£49.05	£49.05	£0.00	In stock (16	0	star-rating
12	167905468	https://boc	Novels	https://boc	Suzie Snow	https://books.toscrape.		0e095ed78	Suzie Snow	Novels	£54.81	Books	£54.81	£54.81	£0.00	In stock (9	0	star-rating
13	167905468	https://boc	Short Storie	https://boc	The Grown	https://books.toscrape.		1bba4b1ba	The Grown	Short Storie	£35.88	Books	£35.88	£35.88	£0.00	In stock (8	0	star-rating
14	167905469	https://boc	Suspense	https://boc	Silence in tl	https://books.toscrape.		e46a164cf3	Silence in tl	Suspense	£58.33	Books	£58.33	£58.33	£0.00	In stock (8	0	star-rating
15	167905469	https://boc	Christian	https://boc	Blue Like Ja	https://books.toscrape.		a322361a3	Blue Like Ja	Christian	£25.77	Books	£25.77	£25.77	£0.00	In stock (7	0	star-rating
16	167905469	https://boc	Christian	https://boc	Crazy Love:	https://books.toscrape.		4b9d9cfe0f	Crazy Love:	Christian	£47.72	Books	£47.72	£47.72	£0.00	In stock (14	0	star-rating
17	167905470	https://boc	Christian	https://boc	(Un)Qualifi	https://books.toscrape.		e4ac92d89	(Un)Qualifi	Christian	£54.00	Books	£54.00	£54.00	£0.00	In stock (16	0	star-rating
18	167905470	https://boc	Historical	https://boc	The Girl You	https://books.toscrape.		eb7459e0c	The Girl You	Historical	£15.79	Books	£15.79	£15.79	£0.00	In stock (6	0	star-rating
19	167905470	https://boc	Historical	https://boc	All the Ligh	https://books.toscrape.		39861fc1c0	All the Ligh	Historical	£29.87	Books	£29.87	£29.87	£0.00	In stock (14	0	star-rating
20	167905471	https://boc	Self Help	https://boc	How to Sto	https://books.toscrape.		e792c0969	How to Sto	Self Help	£46.49	Books	£46.49	£46.49	£0.00	In stock (6	0	star-rating
21	167905471	https://boc	Self Help	https://boc	You Are a B	https://books.toscrape.		88ec62189	You Are a B	Self Help	£12.08	Books	£12.08	£12.08	£0.00	In stock (7	0	star-rating
22	167905471	https://boc	Self Help	https://boc	Overload: H	https://books.toscrape.		8ecf0356a3	Overload: H	Self Help	£52.15	Books	£52.15	£52.15	£0.00	In stock (14	0	star-rating
23	167905471	https://boc	Self Help	https://boc	How to Be	https://books.toscrape.		a6f99b9e6	How to Be	Self Help	£46.03	Books	£46.03	£46.03	£0.00	In stock (16	0	star-rating
24	167905472	https://boc	Self Help	https://boc	Online Mar	https://books.toscrape.		9ffa38d3b8	Online Mar	Self Help	£46.35	Books	£46.35	£46.35	£0.00	In stock (16	0	star-rating
25	167905472	https://boc	Academic	https://boc	Logan Kade	https://books.toscrape.		7093cf549	Logan Kade	Academic	£13.12	Books	£13.12	£13.12	£0.00	In stock (5	0	star-rating
26	167905473	https://boc	Spirituality	https://boc	A New Eart	https://books.toscrape.		f6ae39568	A New Eart	Spirituality	£55.65	Books	£55.65	£55.65	£0.00	In stock (4	0	star-rating

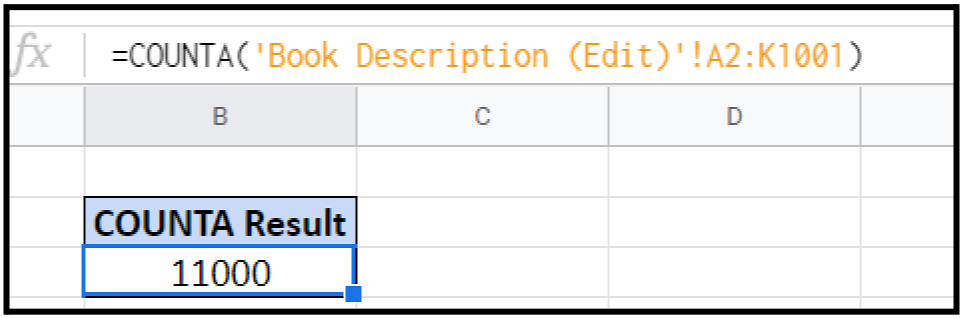
Before starting, it is good practice to duplicate the dataset first before transformation and cleaning, to ensure that I will be capable of undoing the edit if needed, as well as for easier verification purposes after the processing. Alongside this, proper naming convention should be followed with the document name, which I named as “ecommerce_book_data_2023_03_18”.



ecommerce_book_data_2023_03_18

	A	B	C	D	E	F	G	H	I	J	K
1	UPC	title	genre	price	product-type	price-excluding-tax	price-including-tax	tax	availability	number-of-reviews	review
2	abeafe151a587d3b	The Long Sh	Crime	£10.97	Books	£10.97	£10.97	£0.00	In stock (15 available)	0	star-rating One
3	88c21fcd38e2486e	Dark Notes	Erotica	£19.19	Books	£19.19	£19.19	£0.00	In stock (15 available)	0	star-rating Five
4	bb8245f52c7cce8f	Amid the Cł	Cultural	£36.58	Books	£36.58	£36.58	£0.00	In stock (15 available)	0	star-rating One
5	3968e3fbf4695d7c	Equal Is Unf	Politics	£56.86	Books	£56.86	£56.86	£0.00	In stock (12 available)	0	star-rating One
6	2b5054a4192e9b06	Why the Rig	Politics	£52.65	Books	£52.65	£52.65	£0.00	In stock (14 available)	0	star-rating Four
7	a18a4f574854aced	Libertariani	Politics	£51.33	Books	£51.33	£51.33	£0.00	In stock (19 available)	0	star-rating Two
8	f9c87ccb2bc4853b	The Art and	Health	£52.98	Books	£52.98	£52.98	£0.00	In stock (9 available)	0	star-rating Five
9	96aa539bfd4c07e2	10-Day Gree	Health	£49.71	Books	£49.71	£49.71	£0.00	In stock (10 available)	0	star-rating Five
10	047dfa57ba0bf638	Eat Fat, Get	Health	£54.07	Books	£54.07	£54.07	£0.00	In stock (14 available)	0	star-rating Two
11	43c6a960757cf559	The Bulletpı	Health	£49.05	Books	£49.05	£49.05	£0.00	In stock (16 available)	0	star-rating Three
12	0e095ed785af7e43	Suzie Snowf	Novels	£54.81	Books	£54.81	£54.81	£0.00	In stock (9 available)	0	star-rating Five
13	1bba4b1ba23ed2d6	The Grownu	Short Stories	£35.88	Books	£35.88	£35.88	£0.00	In stock (8 available)	0	star-rating One
14	e46a164cf310fd7d	Silence in th	Suspense	£58.33	Books	£58.33	£58.33	£0.00	In stock (8 available)	0	star-rating Three
15	a322361a37a78d3d	Blue Like Ja:	Christian	£25.77	Books	£25.77	£25.77	£0.00	In stock (7 available)	0	star-rating One
16	4b9d9cfe0f435788	Crazy Love:	Christian	£47.72	Books	£47.72	£47.72	£0.00	In stock (14 available)	0	star-rating Two
17	e4ac92d89b946781	(Un)Qualifie	Christian	£54.00	Books	£54.00	£54.00	£0.00	In stock (16 available)	0	star-rating Five
18	eb7459e0c570251d	The Girl You	Historical	£15.79	Books	£15.79	£15.79	£0.00	In stock (6 available)	0	star-rating One
19	39861fc1c0845d15	All the Light	Historical	£29.87	Books	£29.87	£29.87	£0.00	In stock (14 available)	0	star-rating Five
20	e792c0969cefe1b8	How to Stoլ	Self Help	£46.49	Books	£46.49	£46.49	£0.00	In stock (6 available)	0	star-rating Five
21	88ec621893e595d0	You Are a B:	Self Help	£12.08	Books	£12.08	£12.08	£0.00	In stock (7 available)	0	star-rating Three
22	8ecf0356a3878811	Overload: H	Self Help	£52.15	Books	£52.15	£52.15	£0.00	In stock (14 available)	0	star-rating Three
23	a6f99b9e687f9d46	How to Be N	Self Help	£46.03	Books	£46.03	£46.03	£0.00	In stock (16 available)	0	star-rating One
24	9ffa38d3b832433a	Online Marl	Self Help	£46.35	Books	£46.35	£46.35	£0.00	In stock (16 available)	0	star-rating One
25	7093cf549cd2e7de	Logan Kade	Academic	£13.12	Books	£13.12	£13.12	£0.00	In stock (5 available)	0	star-rating Two
26	f6ae39568db1130e	A New Eartł	Spirituality	£55.65	Books	£55.65	£55.65	£0.00	In stock (4 available)	0	star-rating Five

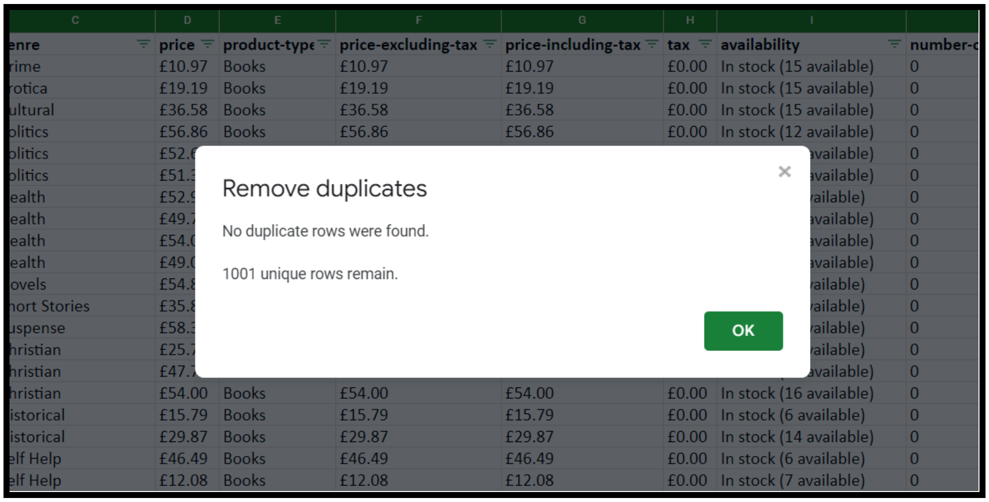
Looking at this data set, there is no need for the link values, pagination, and web scraping information for the analysis, which makes it easy to be removed from the sheet. Looking back at the e-commerce website, there are a total of 1000 records or books overall. From such, we can use the COUNTA function to check whether the 11 columns of data left are all filled with 1000 records each with no null values, resulting in an overall count of 11000.



```
=COUNTA('Book Description (Edit)'!A2:K1001)
```

	B	C	D
	COUNTA Result		
	11000		

We can then check for duplicates with the remove duplicates tool. Result shows that all rows are unique with no duplicates. Alongside this, we can also trim possible leading, trailing, and repeated spaces in data with the trim whitespace tool. Again, results show that there are no whitespaces in the data.

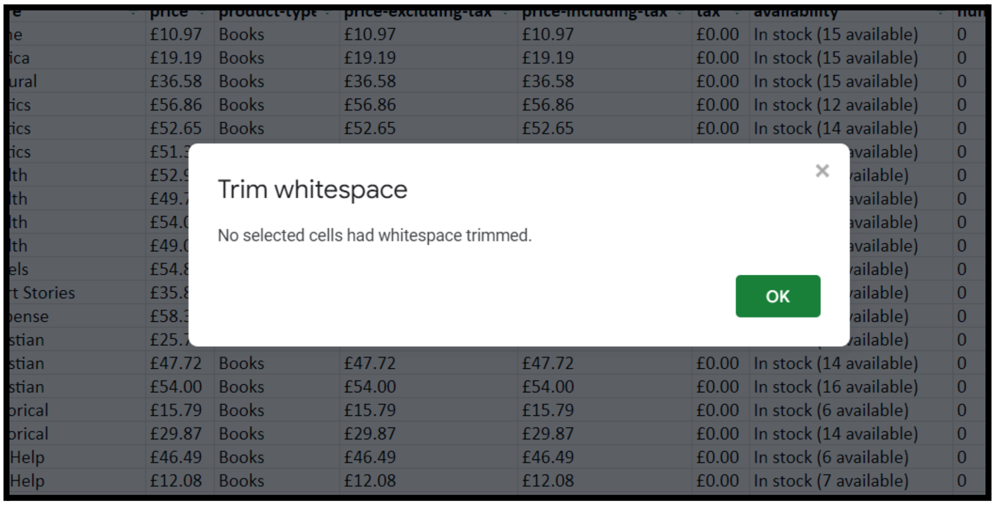


Remove duplicates

No duplicate rows were found.

1001 unique rows remain.

enre	price	product-type	price-excluding-tax	price-including-tax	tax	availability	number-c
rime	£10.97	Books	£10.97	£10.97	£0.00	In stock (15 available)	0
rotica	£19.19	Books	£19.19	£19.19	£0.00	In stock (15 available)	0
ultural	£36.58	Books	£36.58	£36.58	£0.00	In stock (15 available)	0
olitics	£56.86	Books	£56.86	£56.86	£0.00	In stock (12 available)	0
olitics	£52.6					vailable)	0
olitics	£51.3					vailable)	0
ealth	£52.9					ailable)	0
ealth	£49.7					vailable)	0
ealth	£54.0					vailable)	0
ealth	£49.0					vailable)	0
ovels	£54.8					ailable)	0
hort Stories	£35.8					ailable)	0
uspense	£58.3					ailable)	0
hristian	£25.7					ailable)	0
hristian	£47.7					vailable)	0
hristian	£54.00	Books	£54.00	£54.00	£0.00	In stock (16 available)	0
istorical	£15.79	Books	£15.79	£15.79	£0.00	In stock (6 available)	0
istorical	£29.87	Books	£29.87	£29.87	£0.00	In stock (14 available)	0
elf Help	£46.49	Books	£46.49	£46.49	£0.00	In stock (6 available)	0
elf Help	£12.08	Books	£12.08	£12.08	£0.00	In stock (7 available)	0



Trim whitespace

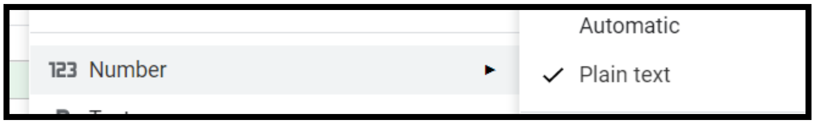
No selected cells had whitespace trimmed.

e	price	product-type	price-excluding-tax	price-including-tax	tax	availability	nun
e	£10.97	Books	£10.97	£10.97	£0.00	In stock (15 available)	0
ica	£19.19	Books	£19.19	£19.19	£0.00	In stock (15 available)	0
ural	£36.58	Books	£36.58	£36.58	£0.00	In stock (15 available)	0
ics	£56.86	Books	£56.86	£56.86	£0.00	In stock (12 available)	0
ics	£52.65	Books	£52.65	£52.65	£0.00	In stock (14 available)	0
ics	£51.3					vailable)	0
th	£52.9					ailable)	0
th	£49.7					vailable)	0
th	£54.0					vailable)	0
th	£49.0					vailable)	0
els	£54.8					ailable)	0
t Stories	£35.8					ailable)	0
ense	£58.3					ailable)	0
stian	£25.7					ailable)	0
stian	£47.72	Books	£47.72	£47.72	£0.00	In stock (14 available)	0
stian	£54.00	Books	£54.00	£54.00	£0.00	In stock (16 available)	0
orical	£15.79	Books	£15.79	£15.79	£0.00	In stock (6 available)	0
orical	£29.87	Books	£29.87	£29.87	£0.00	In stock (14 available)	0
Help	£46.49	Books	£46.49	£46.49	£0.00	In stock (6 available)	0
Help	£12.08	Books	£12.08	£12.08	£0.00	In stock (7 available)	0

From here, I can now check for each column and manage or transform them to be usable for analysis. The process I took are as follows:

- The UPC and Book Title values are maintained as they can be used as identification of each book available in the website.
- Genre is also kept as it is already in a usable and descriptive enough format.
- For price, which is in pound sterling (£), was detected to be in text format, upon looking at the Format -> Number tab. From such, I tried to change it to currency format but it would show the value 0 when I tried checking through the SUM function. Therefore, I considered that the existing pound sign might be influencing the output, which made me perform a SPLIT function with the £ as the

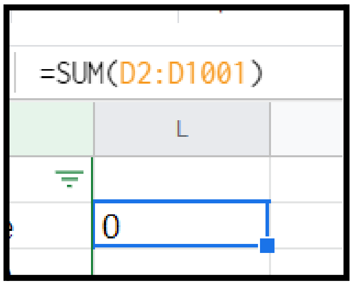
delimiter to get a new set of prices column without the preexisting sign. Afterwards, I removed the old values with the new results converted to the appropriate currency.



123 Number ▸

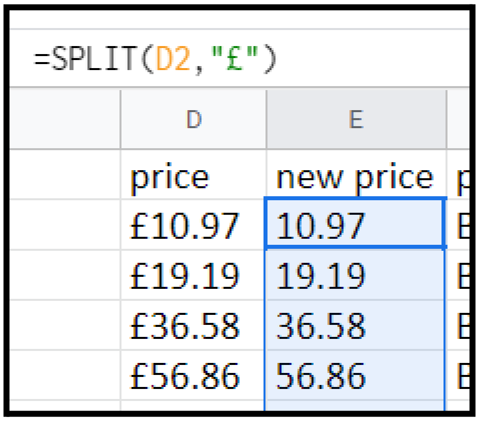
Automatic

✓ Plain text



```
=SUM(D2:D1001)
```

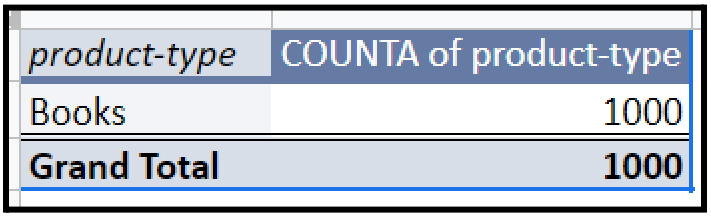
	L
	0



```
=SPLIT(D2,"£")
```

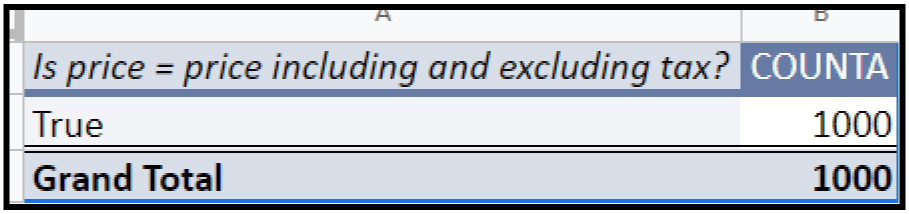
	D	E
	price	new price
	£10.97	10.97
	£19.19	19.19
	£36.58	36.58
	£56.86	56.86

- With product type, there is only one type which is “Books”, as counted through a Pivot Table. The column can be removed as it would not introduce any differing insight in the analysis.



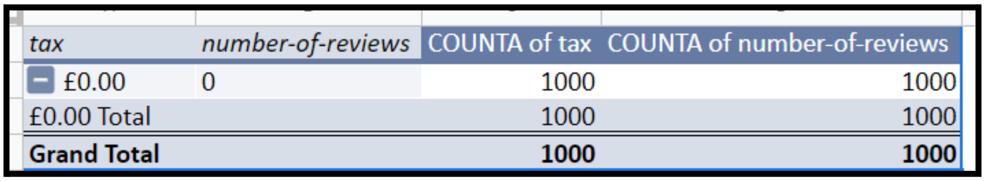
product-type	COUNTA of product-type
Books	1000
Grand Total	**1000**

- For both price including and excluding tax, I performed the same process I did with the original price column. However, one noticeable aspect of all price-related columns is that the prices look similar, which might indicate that all three columns have the same values. From such, I utilized the IF function to check the similarity of values per row, and then counted all true values and checked if it totals to 1000 with a pivot table. Achieving a total of 1000 concluded that all three columns are the same, which allows me to remove the price-tax related columns.



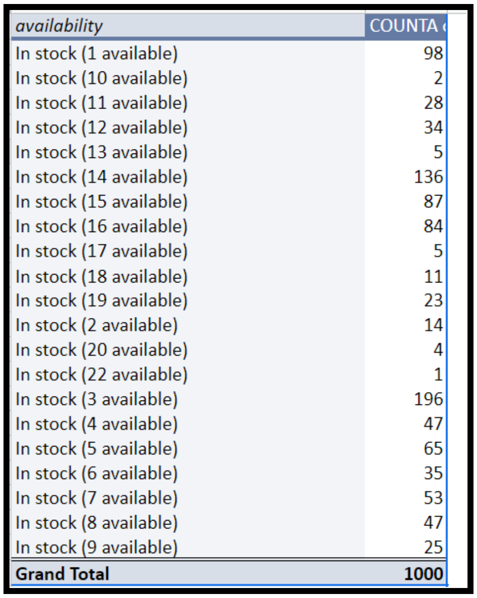
Is price = price including and excluding tax?	COUNTA
True	1000
Grand Total	**1000**

- Tax and number of reviews column also follows the unchanging value situation like product type, allowing me to remove both of the columns without issue.

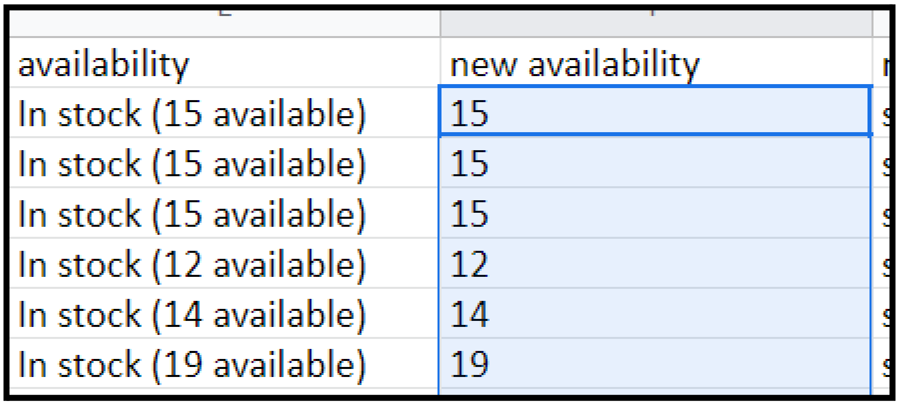


tax	*number-of-reviews*	COUNTA of tax	COUNTA of number-of-reviews
⊟ £0.00	0	1000	1000
£0.00 Total		1000	1000
Grand Total		**1000**	**1000**

- For data under the availability column, I could first check whether there is a “no stock” value through a pivot table. Seeing that only values present are the ones in stock as well as having a consistent first half of the values, I can use the MID function to extract all the stock number values consistently.

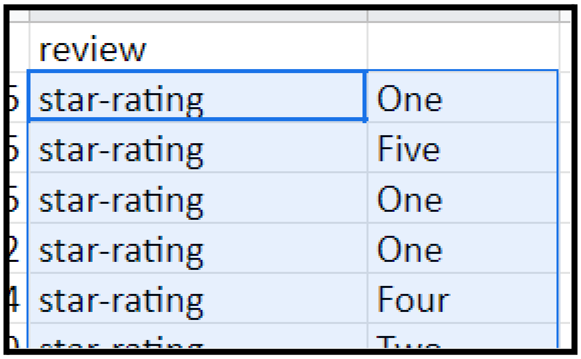


availability	COUNTA
In stock (1 available)	98
In stock (10 available)	2
In stock (11 available)	28
In stock (12 available)	34
In stock (13 available)	5
In stock (14 available)	136
In stock (15 available)	87
In stock (16 available)	84
In stock (17 available)	5
In stock (18 available)	11
In stock (19 available)	23
In stock (2 available)	14
In stock (20 available)	4
In stock (22 available)	1
In stock (3 available)	196
In stock (4 available)	47
In stock (5 available)	65
In stock (6 available)	35
In stock (7 available)	53
In stock (8 available)	47
In stock (9 available)	25
Grand Total	1000



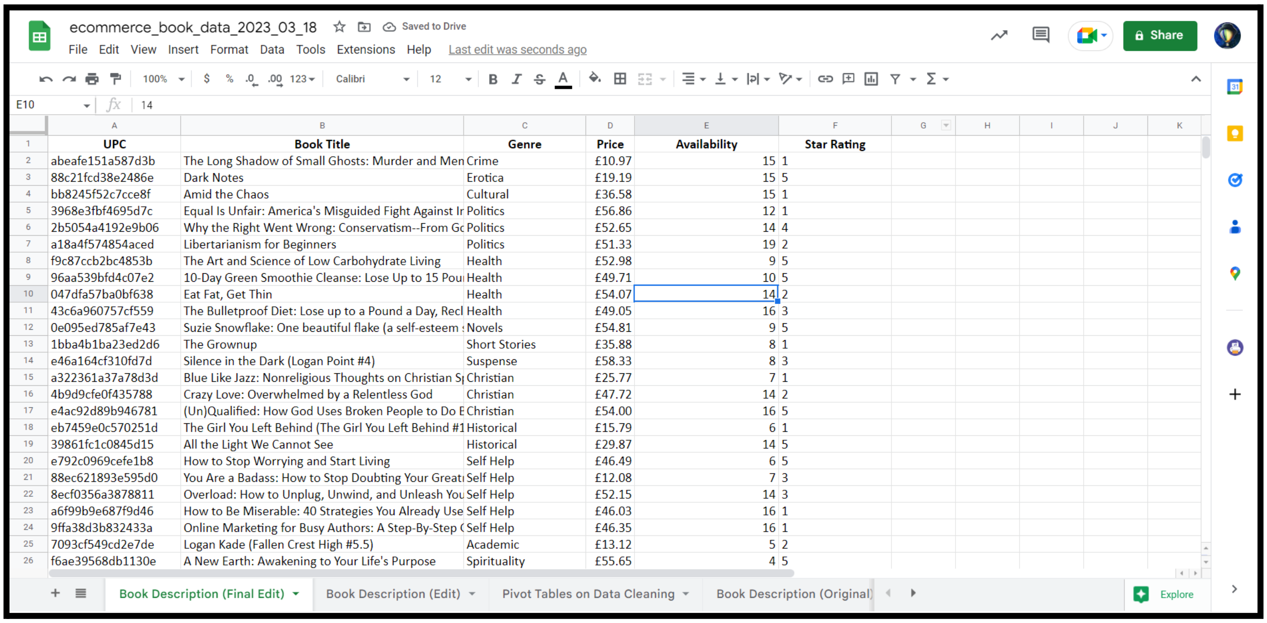
availability	new availability
In stock (15 available)	15
In stock (15 available)	15
In stock (15 available)	15
In stock (12 available)	12
In stock (14 available)	14
In stock (19 available)	19

- Lastly, the review column contains the element attribute scrapped from the website. The star-rating text and the number value are separated by the split text to columns tool. The numbers are then replaced with numerical equivalents with the find and replace tool.



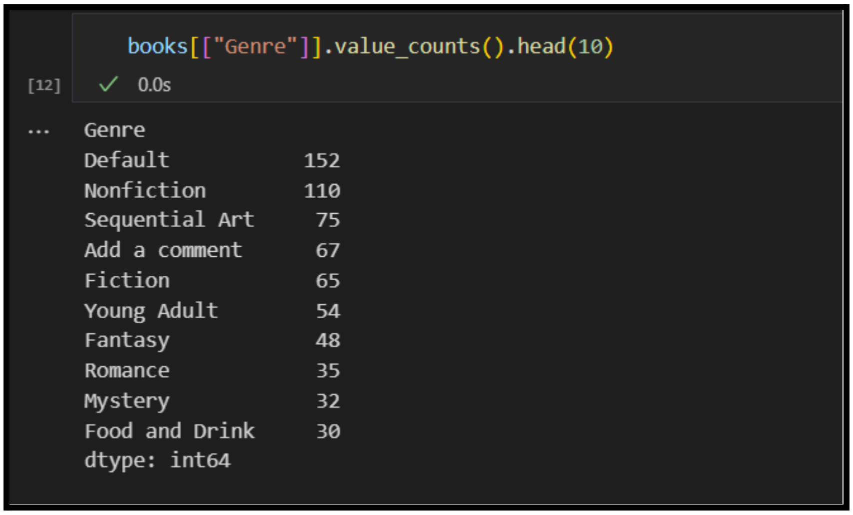
review	
star-rating	One
star-rating	Five
star-rating	One
star-rating	One
star-rating	Four
star-rating	Two

After the cleaning, the data is left with UPC, Book Title, Genre, Price, Availability, and Star Rating columns, which are cleaned for analysis.


ecommerce_book_data_2023_03_18
Saved to Drive
File Edit View Insert Format Data Tools Extensions Help
Last edit was seconds ago
Share
100%
Calibri
12
E10
14
Book Description (Final Edit)
Book Description (Edit)
Pivot Tables on Data Cleaning
Book Description (Original)
Explore

	A	B	C	D	E	F
1	UPC	Book Title	Genre	Price	Availability	Star Rating
2	abeafe151a587d3b	The Long Shadow of Small Ghosts: Murder and Men	Crime	£10.97	15	1
3	88c21fcd38e2486e	Dark Notes	Erotica	£19.19	15	5
4	bb8245f52c7cce8f	Amid the Chaos	Cultural	£36.58	15	1
5	3968e3fbf4695d7c	Equal Is Unfair: America's Misguided Fight Against Ir	Politics	£56.86	12	1
6	2b5054a4192e9b06	Why the Right Went Wrong: Conservatism--From Gc	Politics	£52.65	14	4
7	a18a4f574854aced	Libertarianism for Beginners	Politics	£51.33	19	2
8	f9c87ccb2bc4853b	The Art and Science of Low Carbohydrate Living	Health	£52.98	9	5
9	96aa539bfd4c07e2	10-Day Green Smoothie Cleanse: Lose Up to 15 Pou	Health	£49.71	10	5
10	047dfa57ba0bf638	Eat Fat, Get Thin	Health	£54.07	14	2
11	43c6a960757cf559	The Bulletproof Diet: Lose up to a Pound a Day, Recl	Health	£49.05	16	3
12	0e095ed785af7e43	Suzie Snowflake: One beautiful flake (a self-esteem :	Novels	£54.81	9	5
13	1bba4b1ba23ed2d6	The Grownup	Short Stories	£35.88	8	1
14	e46a164cf310fd7d	Silence in the Dark (Logan Point #4)	Suspense	£58.33	8	3
15	a322361a37a78d3d	Blue Like Jazz: Nonreligious Thoughts on Christian S	Christian	£25.77	7	1
16	4b9d9cfe0f435788	Crazy Love: Overwhelmed by a Relentless God	Christian	£47.72	14	2
17	e4ac92d89b946781	(Un)Qualified: How God Uses Broken People to Do E	Christian	£54.00	16	5
18	eb7459e0c570251d	The Girl You Left Behind (The Girl You Left Behind #1	Historical	£15.79	6	1
19	39861fc1c0845d15	All the Light We Cannot See	Historical	£29.87	14	5
20	e792c0969cefe1b8	How to Stop Worrying and Start Living	Self Help	£46.49	6	5
21	88ec621893e595d0	You Are a Badass: How to Stop Doubting Your Great	Self Help	£12.08	7	3
22	8ecf0356a3878811	Overload: How to Unplug, Unwind, and Unleash You	Self Help	£52.15	14	3
23	a6f99b9e687f9d46	How to Be Miserable: 40 Strategies You Already Use	Self Help	£46.03	16	1
24	9ffa38d3b832433a	Online Marketing for Busy Authors: A Step-By-Step (	Self Help	£46.35	16	1
25	7093cf549cd2e7de	Logan Kade (Fallen Crest High #5.5)	Academic	£13.12	5	2
26	f6ae39568db1130e	A New Earth: Awakening to Your Life's Purpose	Spirituality	£55.65	4	5

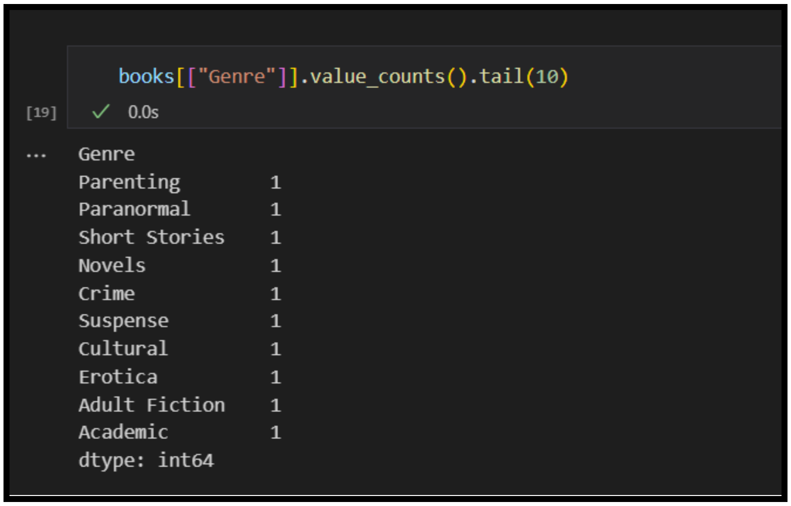
For the final processing, I needed to handle the imbalanced dataset in terms of genre. For this, I used Python with Jupyter Notebook to analyze the book distribution by genre. The analysis displayed a big gap between the maximum and minimum book count among genres.



```
books[["Genre"]].value_counts().head(10)
```

[12] ✓ 0.0s

```
Genre
Default            152
Nonfiction         110
Sequential Art      75
Add a comment       67
Fiction             65
Young Adult         54
Fantasy             48
Romance             35
Mystery             32
Food and Drink      30
dtype: int64
```

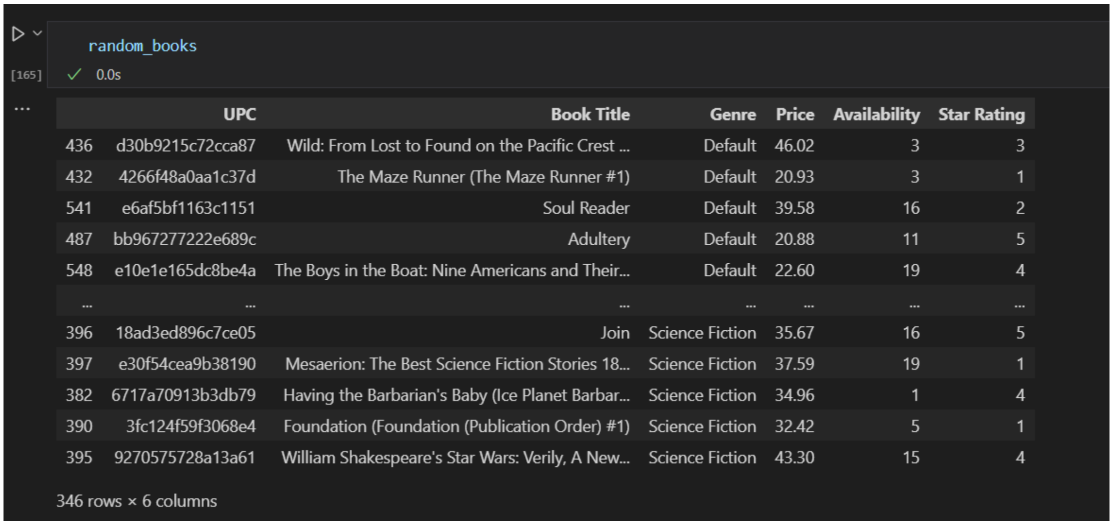


```
books[["Genre"]].value_counts().tail(10)
[19]  ✓ 0.0s
```

```
Genre
Parenting        1
Paranormal       1
Short Stories    1
Novels           1
Crime            1
Suspense         1
Cultural         1
Erotica          1
Adult Fiction    1
Academic         1
dtype: int64
```

This result shows that performing analysis with our current dataset may introduce disproportionate influence or bias leading to misrepresentation, as the data is heavily skewed to genres with more books. From such, I considered undersampling the high book count genres to equalize with the less represented genres, as well as removing some of the genres with too little data for analysis. This would introduce limitations to our analysis as few genres will not be analyzed, but it would make the analysis more balanced.

To do such, I extracted the mean value of all genre's book counts, which resulted in 20 books available per genre on average. From then, I extracted a randomized sample of 20 for every genre with book count greater than 20, and also took all genres with 15-20 books, as the book count can be satisfactory enough to make an analysis. Results are saved as a .csv file for spreadsheet analysis and visualization.



```
random_books
[165]  ✓ 0.0s
```

	UPC	Book Title	Genre	Price	Availability	Star Rating
436	d30b9215c72cca87	Wild: From Lost to Found on the Pacific Crest ...	Default	46.02	3	3
432	4266f48a0aa1c37d	The Maze Runner (The Maze Runner #1)	Default	20.93	3	1
541	e6af5bf1163c1151	Soul Reader	Default	39.58	16	2
487	bb967277222e689c	Adultery	Default	20.88	11	5
548	e10e1e165dc8be4a	The Boys in the Boat: Nine Americans and Their...	Default	22.60	19	4
...	...	...	...	...	...	...
396	18ad3ed896c7ce05	Join	Science Fiction	35.67	16	5
397	e30f54cea9b38190	Mesaerion: The Best Science Fiction Stories 18...	Science Fiction	37.59	19	1
382	6717a70913b3db79	Having the Barbarian's Baby (Ice Planet Barbar...	Science Fiction	34.96	1	4
390	3fc124f59f3068e4	Foundation (Foundation (Publication Order) #1)	Science Fiction	32.42	5	1
395	9270575728a13a61	William Shakespeare's Star Wars: Verily, A New...	Science Fiction	43.30	15	4

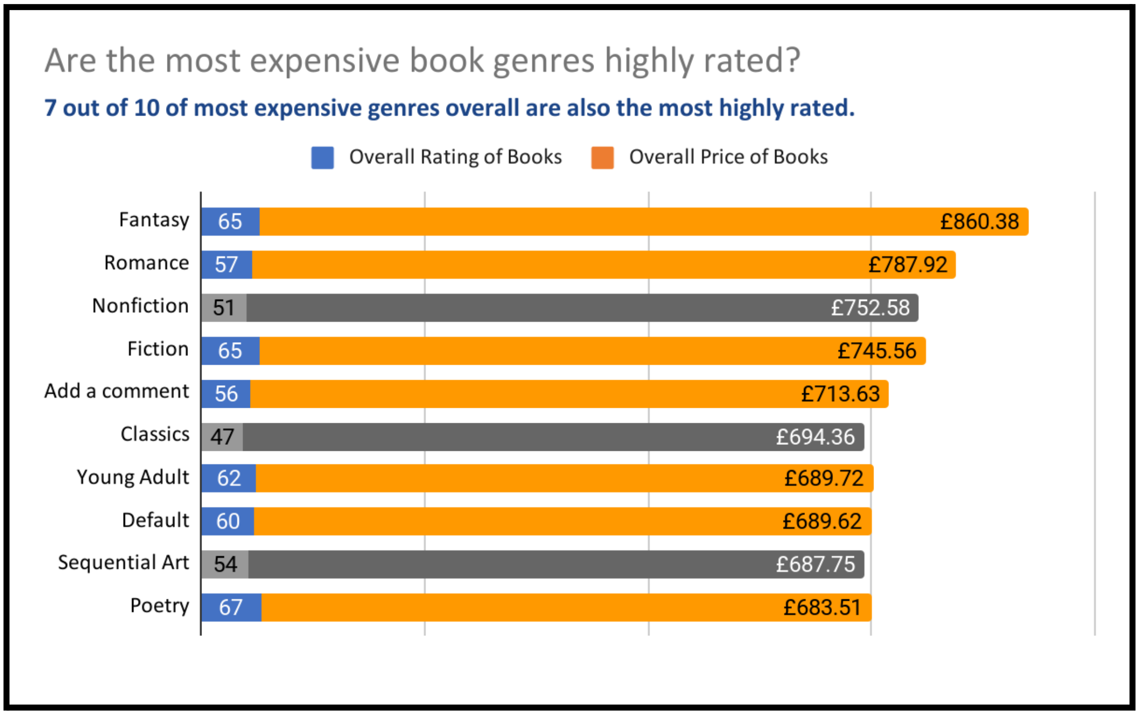
346 rows × 6 columns

Overall, there are 346 records of books extracted. The randomized books per genre is then saved as a .csv file to be analyzed and visualized in the spreadsheet.

IV. Data Analysis and Findings

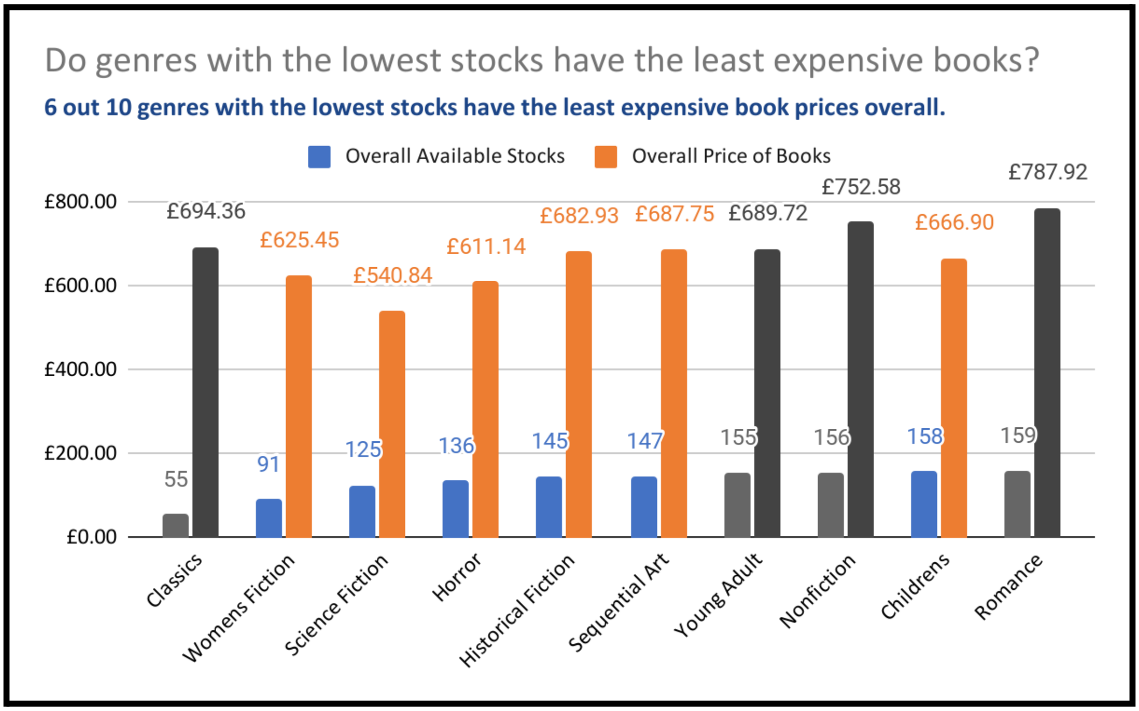
The dataset contains 346 rows of book inputs, with 6 columns of book descriptions, which includes: (1) the UPC code, (2) the title of the book, (3) the book genre, (4) the price of the book (£), (5) the stock available for sale, and (6) the review rating of the book, from 1 to 5 stars. With this, we can attempt to answer our initial hypotheses.

- The most expensive genres have the most highly rated reviews.



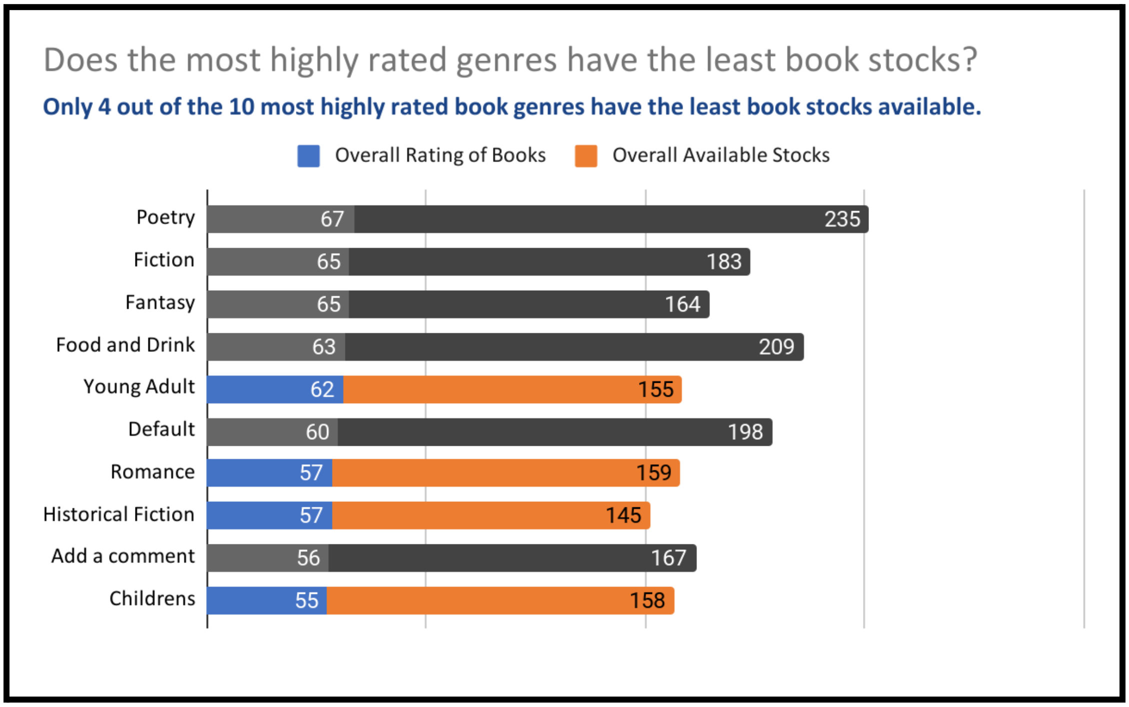
With the following graph, it can be seen that more expensive books are more likely to be highly rated. This could imply that these genres have higher quality books that justifies the high price customers pay for, which can indicate customer satisfaction. Meanwhile, the genres of nonfiction, classics, and sequential art are expensive overall but rated poorly in comparison. This may indicate customer dissatisfaction due to spending highly for a poorer quality book. From such, we can consider expensive genres as long as it is rated highly.

- Genres with the least expensive books overall have lower stock available.



6 of the 10 least expensive genres also have the least amount of available stock left on their books. This may imply that lesser expensive genres cause more purchasing decisions from customers due to affordability. From such, a good balance between affordability and price can be a factor that decides whether more people would buy books, leading to higher profitability.

- The most rated genres have the least amount of book stocks.



The following graph displays that most highly rated genres often have the most stock remaining for purchase. This indicates that despite being higher quality, less books are being purchased by customers, which may be due to their price. Despite this, 4 out of 10 of the most highly rated books still get lower stock count, which may indicate popularity among these genres.

V. Conclusion and Recommendations

The results of the hypothesis displays the following conclusions:

- More expensive genres are more likely to be highly rated.
- Less expensive genres are more likely to have less stock available.
- Highly rated book genres often have more stocks available.

The results also showed that customers would buy expensive books if they are highly rated in quality, as well as if the book is under an affordable price. To achieve the best customer satisfaction, I can provide genre recommendations that fulfill the following criteria:

CRITERIA 1: High price, low in stock, and highly rated

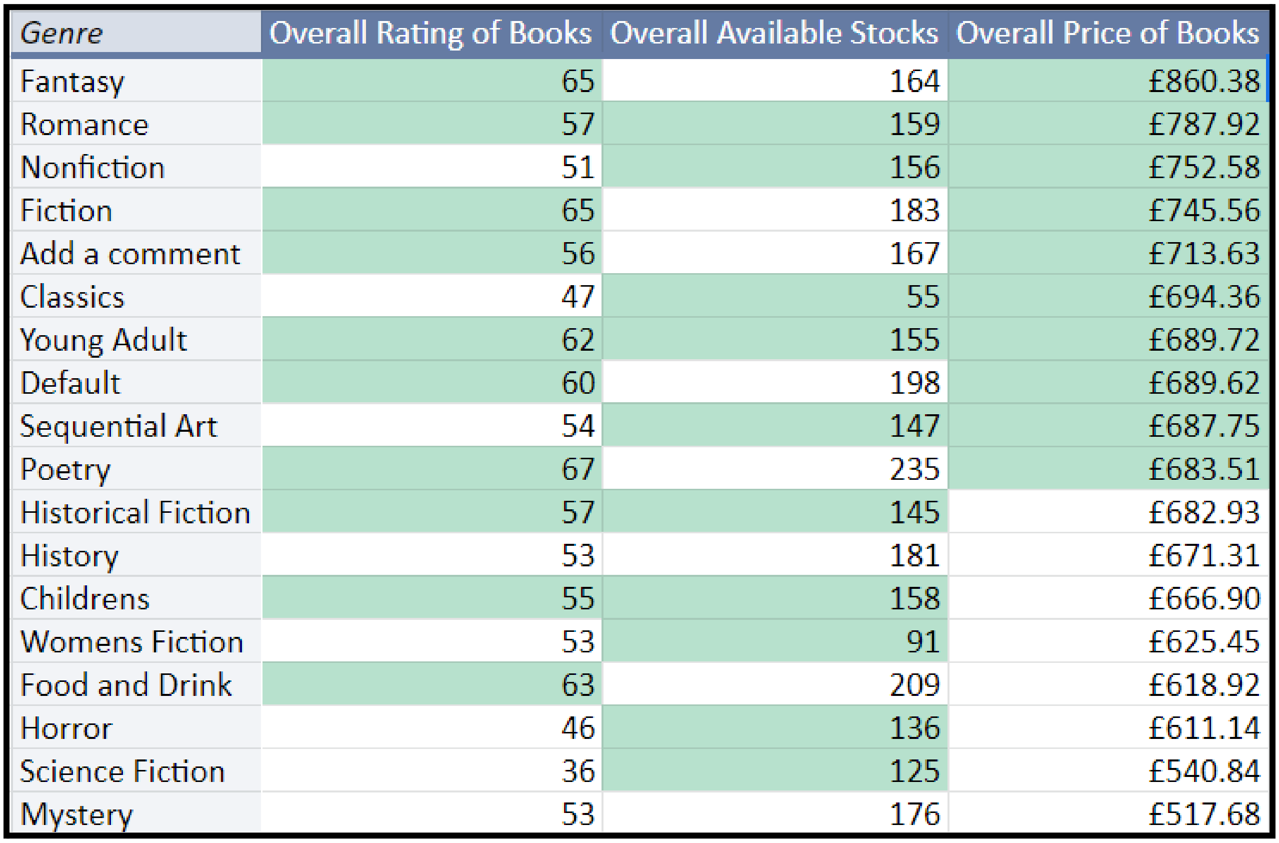
CRITERIA 2: Affordable, low in stock, and highly rated

CRITERIA 3: High price, low in stock, but lowly rated

CRITERIA 4: Affordable and low of stock.

CRITERIA 5: High price and highly rated.

CRITERIA 6: Affordable and highly rated.



Genre	Overall Rating of Books	Overall Available Stocks	Overall Price of Books
Fantasy	65	164	£860.38
Romance	57	159	£787.92
Nonfiction	51	156	£752.58
Fiction	65	183	£745.56
Add a comment	56	167	£713.63
Classics	47	55	£694.36
Young Adult	62	155	£689.72
Default	60	198	£689.62
Sequential Art	54	147	£687.75
Poetry	67	235	£683.51
Historical Fiction	57	145	£682.93
History	53	181	£671.31
Childrens	55	158	£666.90
Womens Fiction	53	91	£625.45
Food and Drink	63	209	£618.92
Horror	46	136	£611.14
Science Fiction	36	125	£540.84
Mystery	53	176	£517.68

**Green Highlights*

1. *Highest Rated Book Genre Overall*
2. *Lowest Available Stocks in a Genre Overall*
3. *Highest Book Prices in a Genre Overall*

With the following table, the genres that fulfill Criteria 1 are **Romance** and **Young Adult** genres, having high ratings, low stocks remaining, and high price. This indicates that both of these genres have expensive books that are actively popular and high in quality, contributing to high profits.

For Criteria 2, genres such as **Historical Fiction** and **Childrens** can be considered, as they have high ratings with low stocks remaining, despite not being in the high price range. This shows their popularity and quality, which is valuable in keeping customer satisfaction leading to more buying decisions.

Nonfiction, **Classics**, and **Sequential Art** belong to Criteria 3, as despite the low rating, the available stock numbers are among the lowest in comparison, particularly with the classics. These genres are popular enough to get bought by customers, which may be due to reasons outside of overall ratings.

With Criteria 4, **Womens Fiction**, **Horror**, and **Science Fiction** can be considered due to their low stock and having an affordable price. From such, it indicates that these genres are popular to be bought even at a lower price and not considering its overall rating.

For Criteria 5, **Fantasy**, **Fiction**, **Add a comment**, **Default**, and **Poetry** are considered, as these are highly rated genres despite their high price. Despite the stock not being lower, the quality of the books under these genres ensure that there would be consistent consumers for these types of content. This means a consistent stream of profits.

Lastly, we have the **Food and Drink** genre for Criteria 6, as it is highly rated and affordable. However, it also has a large amount of stock remaining, which may indicate that there are not many buyers of books under this genre.

History and **Mystery** genres will not be considered for the recommendation, as both are low in rating and price, as well as maintaining many unsold stocks.

With the following, I would recommend the following genres:

- **Romance**
- **Young Adult**
- **Historical Fiction**
- **Childrens**
- **Nonfiction**
- **Classics**
- **Sequential Art**
- **Womens Fiction**
- **Horror**
- **Science Fiction**

` For the following reasons:

- High rating and expensive books that are low in stock
- Affordable, but low on stock and highly rated
- Lowly rated and expensive, but low on stock
- Affordable and low on stock

VI. Recommendations on Further Tasks and Considered Limitations

To further improve the insights of this analysis, the following are recommended to extend past current limitations:

- Gather data regarding the sales of the books, as the only indication of possible sales is the variability of available stock per book.
- Contextualize the target audience of the business by obtaining customer data, if available
- Determine trends based on historical data, may it be daily, weekly, monthly, etc.