Exploratory Data Analysis and Hypothesis Significance Testing of Kindle eBook dataset

Author: Arnold Richards Alvarez

Introduction

In this document I will explain the process of a Exploratory Data Analysis and a Significance Testing of a Hypothesis deduced from a Kindle eBook dataset that will be referenced in the Dataset Selection section, for detailed information about the process followed with code and formulas to obtain this information please check the Jupyter notebook included with this document

Dataset Selection

For this Exploratory data Analysis, I will be utilizing a kindle eBook dataset obtained from Kaggle.

This dataset contains an array of kindle books with their respective attributes, and we will be focusing on the price, reviews and stars along with other Boolean type information like if it is on kindle unlimited subscription, if it is Bestseller or if it is a GoodReadersChoice.

This dataset was scraped on October 2023, and it was uploaded by user asaniczka and it can be accessed via the following link:

https://www.kaggle.com/datasets/asaniczka/amazon-kindle-books-dataset-2023-130k-books

I chose this dataset because I am an avid reader, and I find it interesting discovering the relations between the different attributes and how that impacts the book reviews, pricing, etc.

In the following cells we will be showing the shape of the dataset along with the information about it including the descriptions of its attributes.

Dataset Description:

Attributes description (These descriptions were obtained directly from the dataset source):

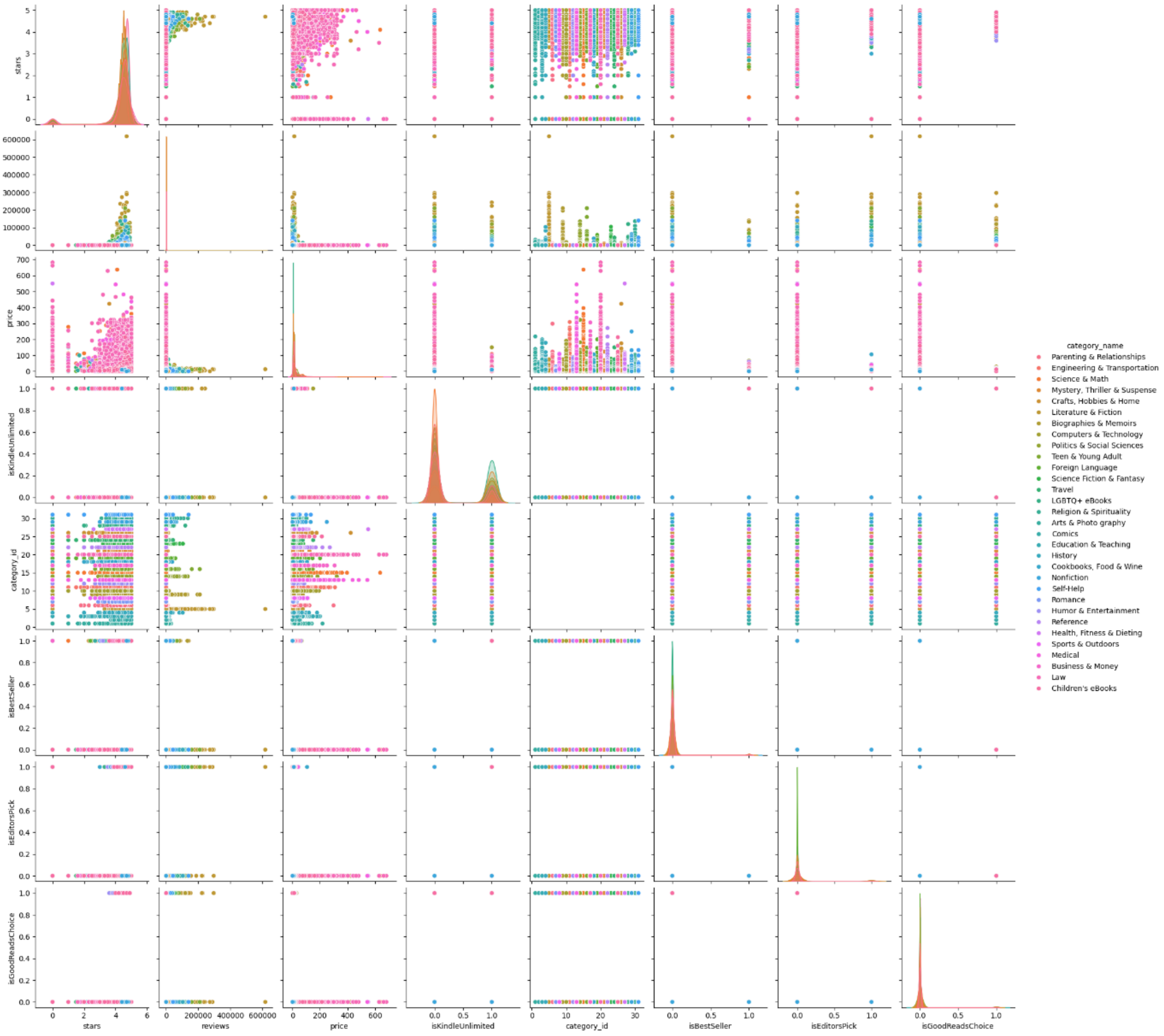
- asin: Product ID from Amazon.
- title: Title of the book.
- author: Author(s) of the book.
- soldBy: Seller(s) of the book.
- imgUrl: URL of the book.
- productURL: URL to the publication on which the eBook is sold.
- stars: Average rating of the book. If 0, no ratings were found.
- reviews: Number of reviews. If 0, no reviews were found.
- price: Price of the book. If 0, the price was unavailable.
- isKindleUnlimited: Whether the book is available through Kindle Unlimited.
- category_id: Serial id assigned to the category this book belongs to.
- isBestSeller: Whether the book had the Amazon Best Seller status or not.
- isEditorsPick: Whether the book had the Editor's Pick status or not.
- isGoodReadsChoice: Whether the book had Goodreads Choice status or not.
- publishedDate: Publication date of the book.
- category_name: Name of the book category.

Initial plan for data exploration

For initial exploration I chose to visualize the amount of data I was working with and the different columns in the dataset to start deciding which ones I would be focusing on the analysis and exploration and for further hypothesis declaration.

I also looked at the dataset info and with the help of features information provided by the author I was able to look at the real amount of missing values.

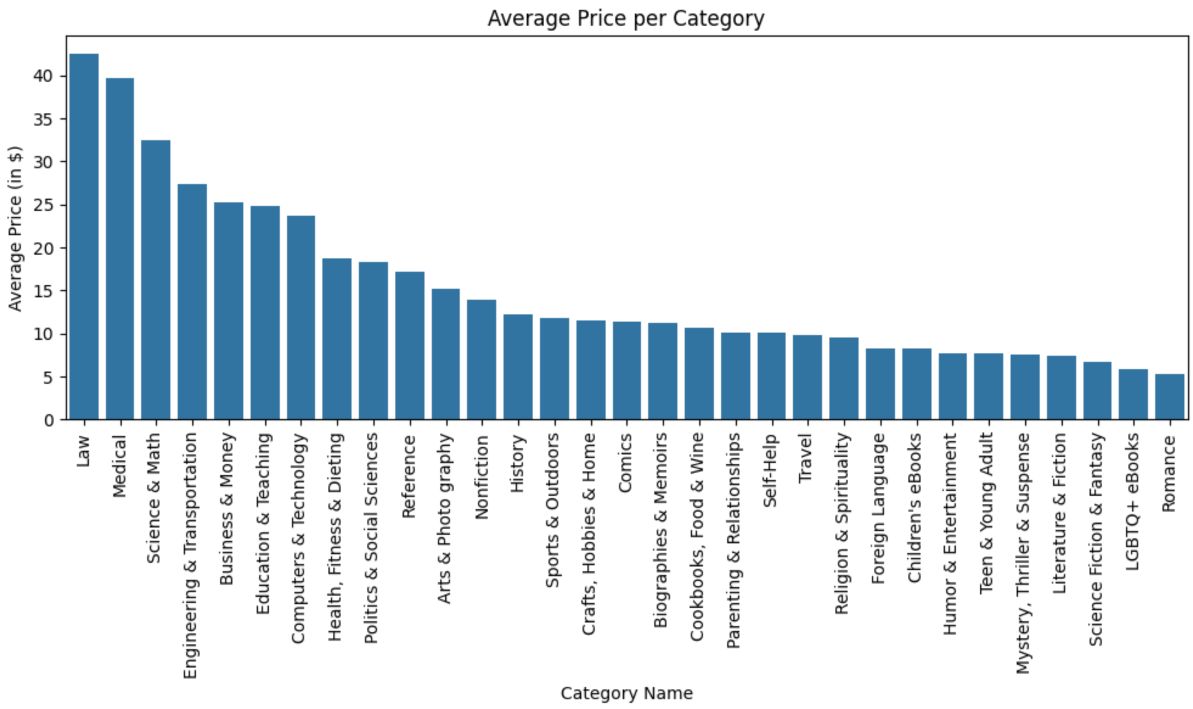
After that I created a pair plot based on the categories to have a more accurate idea of how the dataset is distributed.



Img. 1

In Img. 1 you can see how the category may affect some of the other values such as price and maybe even.

Then I plotted the average price per category to dig deeper into this relationship to decide how to proceed with the data cleaning.



Img. 2

Based on the latter I deduce that there is a relationship between the category and the price and that there could be other similar relationships with price and that they are affected too by the missing values.

Therefore, given that the missing values in price are not a substantial amount of the dataset I decided to ignore them, and I will be eliminating them from a copy of the dataset to work with.

With this I only must decide what to do with the other missing values in reviews and stars.

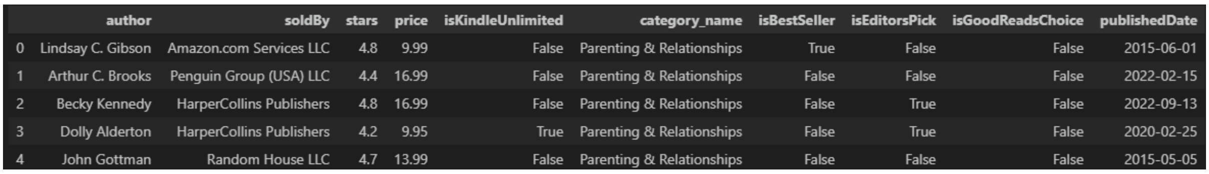
As for stars, since there is not a substantial number of missing values and most rated books are more likely to not be very poorly reviewed, I decided to impute them with the media.

Finaly for reviews, since there is a lot of missing data, and the relationships seem minimal with the features other than stars I will ignore the whole column for the rest of this analysis.

Data cleaning and Feature engineering

My first step into data cleaning and feature engineering was to create a smaller copy of the dataset excluding some columns that wouldn’t be useful for my hypothesis testing and data analysis along with eliminating the before mentioned column of reviews.

This was the result:



	author	soldBy	stars	price	isKindleUnlimited	category_name	isBestSeller	isEditorsPick	isGoodReadsChoice	publishedDate
0	Lindsay C. Gibson	Amazon.com Services LLC	4.8	9.99	False	Parenting & Relationships	True	False	False	2015-06-01
1	Arthur C. Brooks	Penguin Group (USA) LLC	4.4	16.99	False	Parenting & Relationships	False	False	False	2022-02-15
2	Becky Kennedy	HarperCollins Publishers	4.8	16.99	False	Parenting & Relationships	False	True	False	2022-09-13
3	Dolly Alderton	HarperCollins Publishers	4.2	9.95	True	Parenting & Relationships	False	True	False	2020-02-25
4	John Gottman	Random House LLC	4.7	13.99	False	Parenting & Relationships	False	False	False	2015-05-05

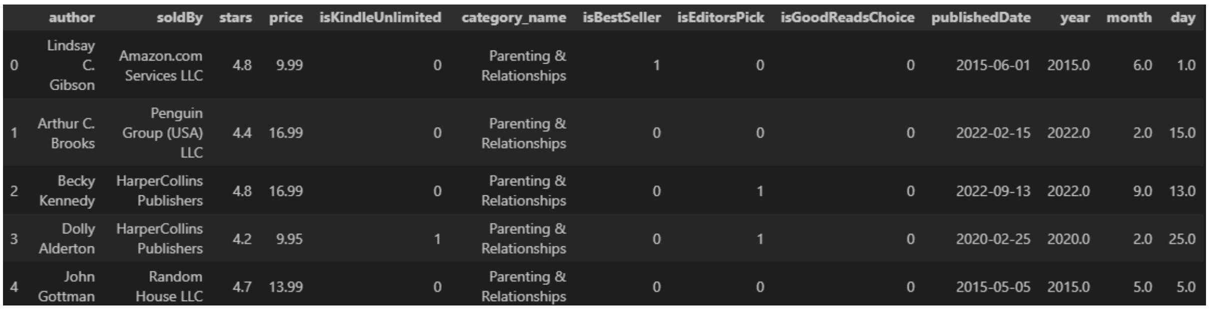
Img. 3

To continue with the data cleaning, I replaced all the missing values in price and stars (ceros) for NA from the pandas library.

After that I eliminated all the rows with NA values in price from the dataset.

Then I replaced all NA values in stars with the median to be able to work with the full dataset.

Finally, I followed the example in the lab to separate the date into three fields (year, month and date) and transformed all the Boolean values of the data into ceros and ones arriving to the final work dataset.



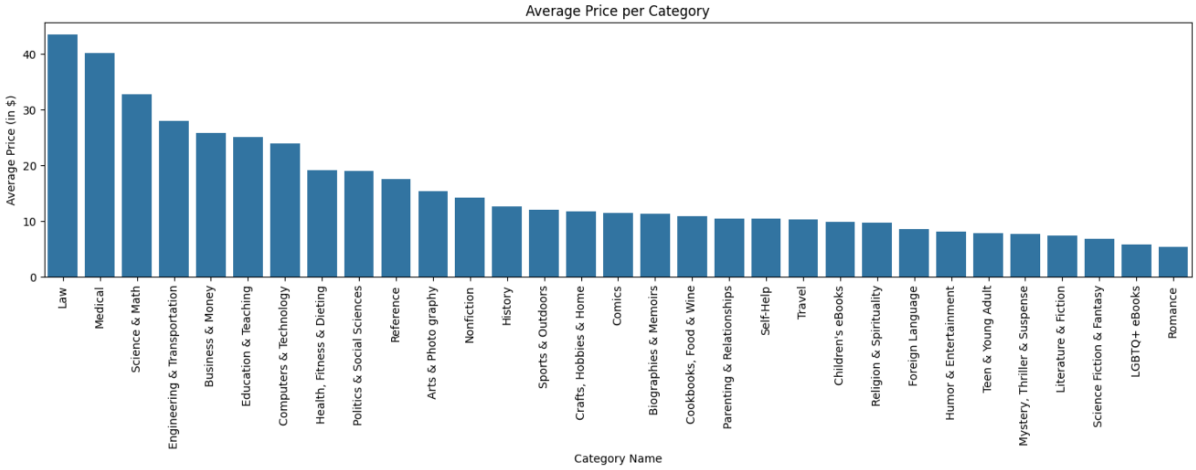
	author	soldBy	stars	price	isKindleUnlimited	category_name	isBestSeller	isEditorsPick	isGoodReadsChoice	publishedDate	year	month	day
0	Lindsay C. Gibson	Amazon.com Services LLC	4.8	9.99	0	Parenting & Relationships	1	0	0	2015-06-01	2015.0	6.0	1.0
1	Arthur C. Brooks	Penguin Group (USA) LLC	4.4	16.99	0	Parenting & Relationships	0	0	0	2022-02-15	2022.0	2.0	15.0
2	Becky Kennedy	HarperCollins Publishers	4.8	16.99	0	Parenting & Relationships	0	1	0	2022-09-13	2022.0	9.0	13.0
3	Dolly Alderton	HarperCollins Publishers	4.2	9.95	1	Parenting & Relationships	0	1	0	2020-02-25	2020.0	2.0	25.0
4	John Gottman	Random House LLC	4.7	13.99	0	Parenting & Relationships	0	0	0	2015-05-05	2015.0	5.0	5.0

Img. 4

Key Findings and Insights

After the cleaning of the data I started to look for correlations or patterns in my data by logical thinking and several visualization techniques that led me to the following conclusions.

Category and Price



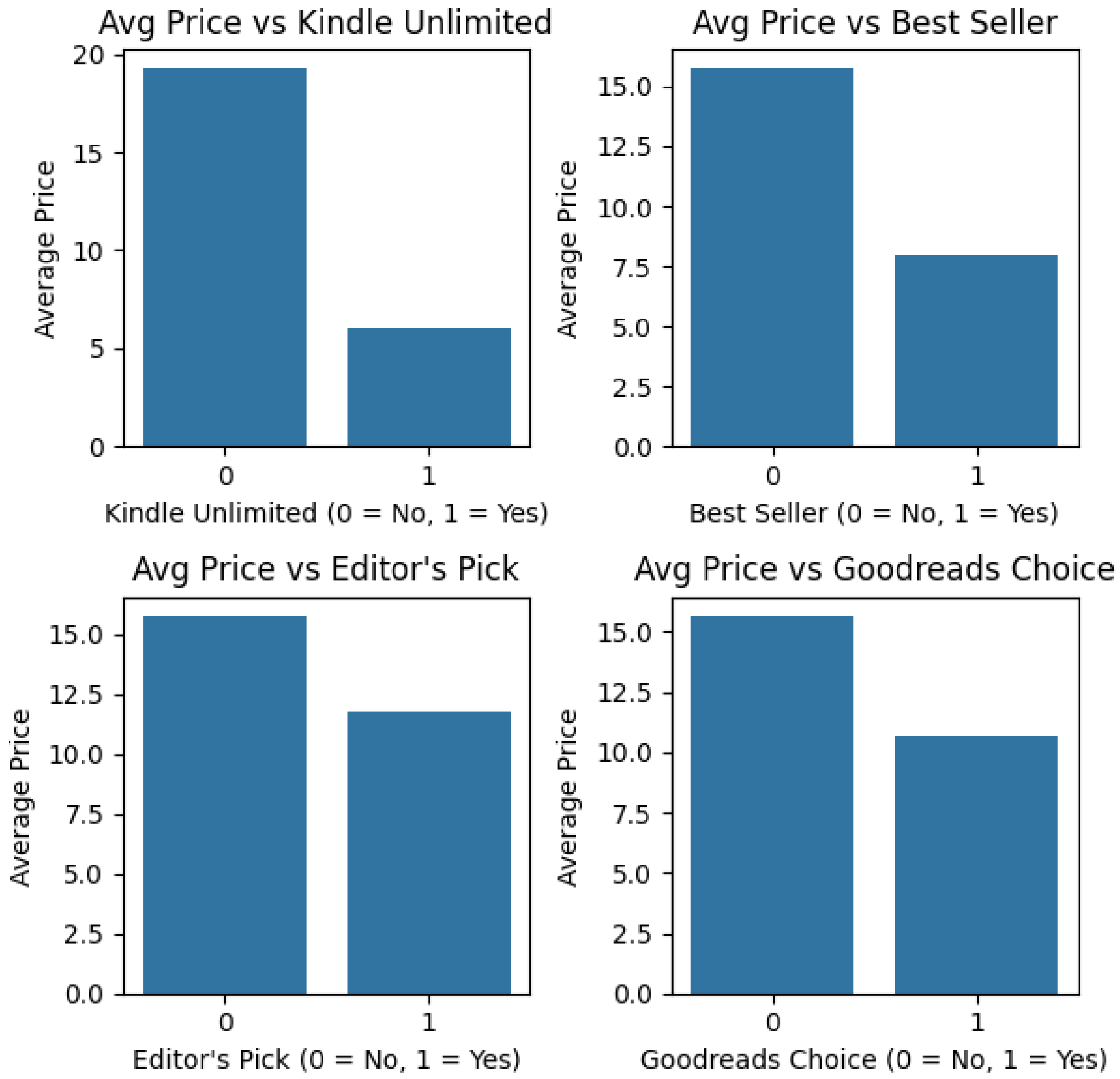
Img. 5

With this chart I realized that depending on the category of the book the prices can vary wildly here we can easily see that the top five categories with higher prices are:

- Law
- Medical
- Science and Math
- Engineering and Transportation
- Business and Money

Therefor if someone would like to get the highest value (price wise) of a new book the data points to these 5 categories as the best choice in average.

Boolean values



Img. 6

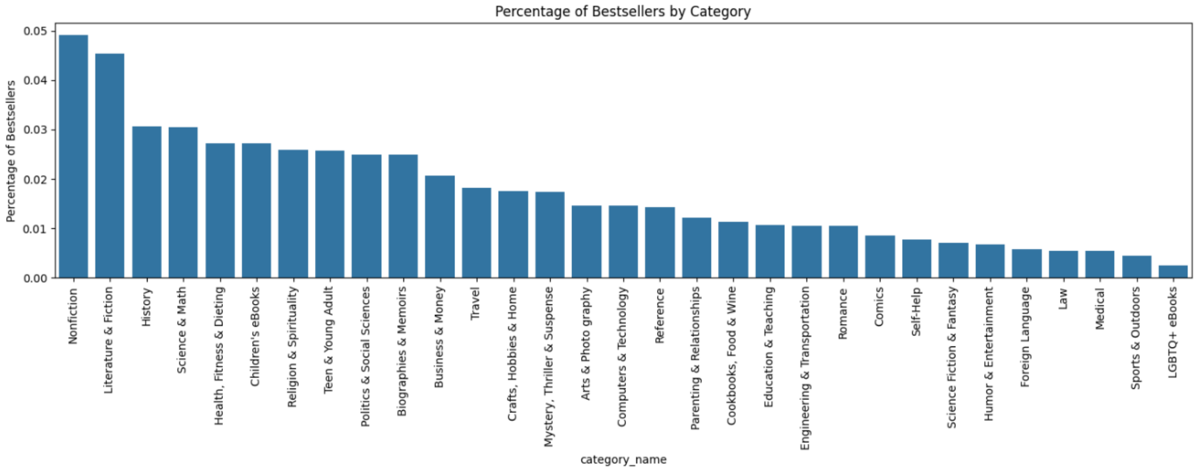
In this chart after analyzing the different average prices between the Boolean values of the dataset, we can see that the kindle unlimited books have the lowest average price.

This can be because they are included in a subscription-based collection of books.

We also can see that the books that are EditorsPick, GoodReadsChoice and BestSeller also have a lower average price than the ones that are not.

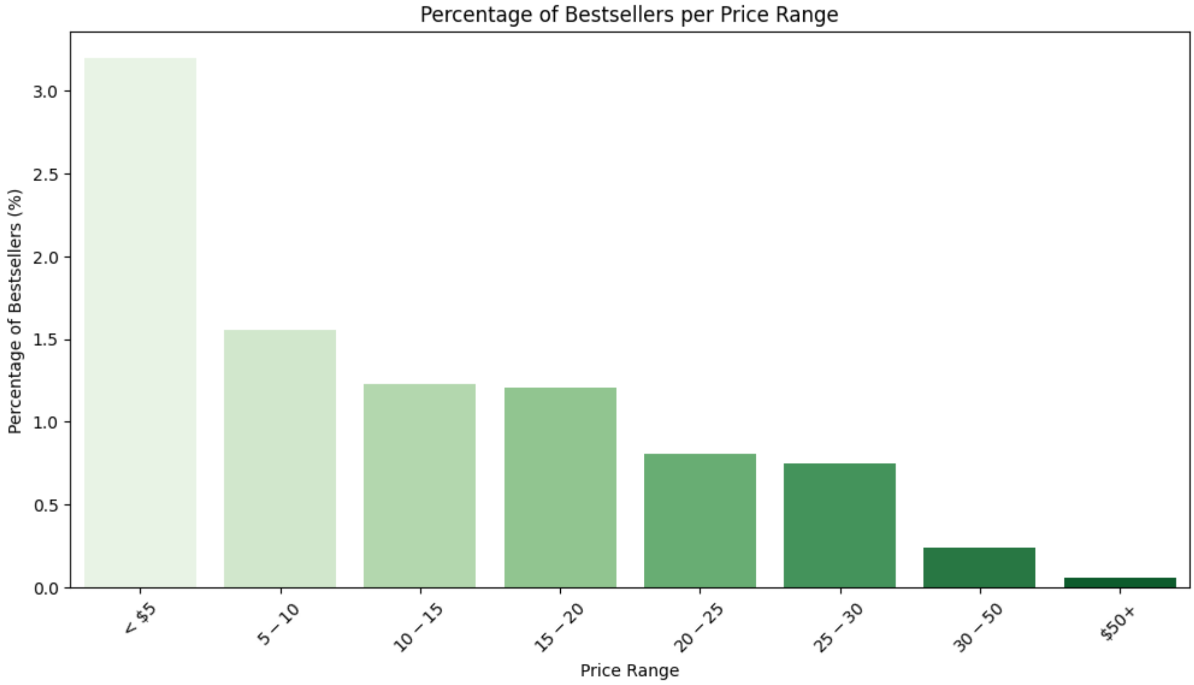
This may be because more accessible books tend to have a bigger public therefor the more opportunities to be valued higher.

Best Sellers



Img. 7

Looking at the best Sellers per Category we can easily see that if we want to publish a kindle book to be a best seller our best chances are with the Non Fiction and the Literature and Fiction categories.



Img. 8

Continuing with the Best seller’s analysis we can see that as well as choosing the correct category the best way of publishing a kindle book to be a best seller the best option is to price it in a moderate price range and the lower the better.

Hypotheses

All the following hypotheses were deduced from the later analysis

Hypothesis 1:

- H_0: Books priced below $5 are not more likely to be bestsellers compared to books priced $5 or above.
- H_1: Books priced below $5 are more likely to be bestsellers compared to books priced $5 or above.

Hypothesis 2:

- H_0: Certain categories have a higher proportion of bestsellers compared to other categories.
- H_1: There is no significant difference in the proportion of bestsellers across different categories.

Hypothesis 3:

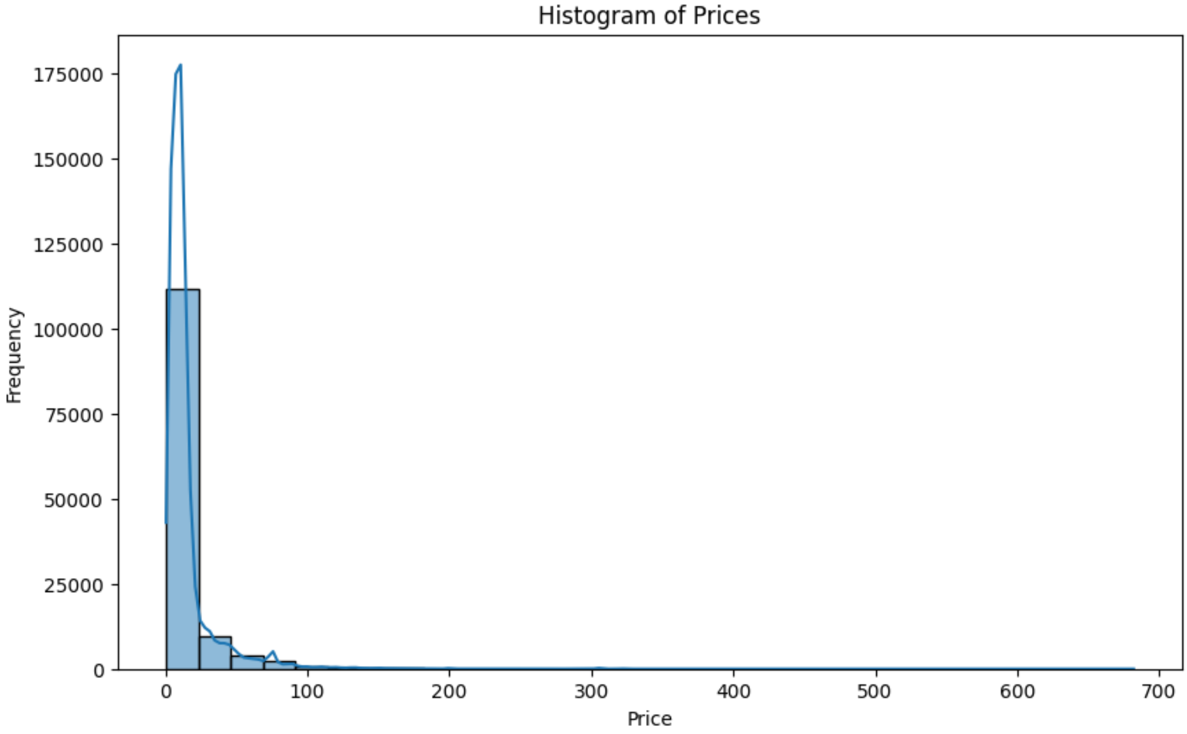
- H_0: There is no significant difference in the average price across different categories.
- H_1: At least one category has a significantly different average price compared to others.

Hypothesis Significance Testing

Hypothesis 1:

- H_0: Books priced below $5 are not more likely to be bestsellers compared to books priced $5 or above.
- H_1: Books priced below $5 are more likely to be bestsellers compared to books priced $5 or above.

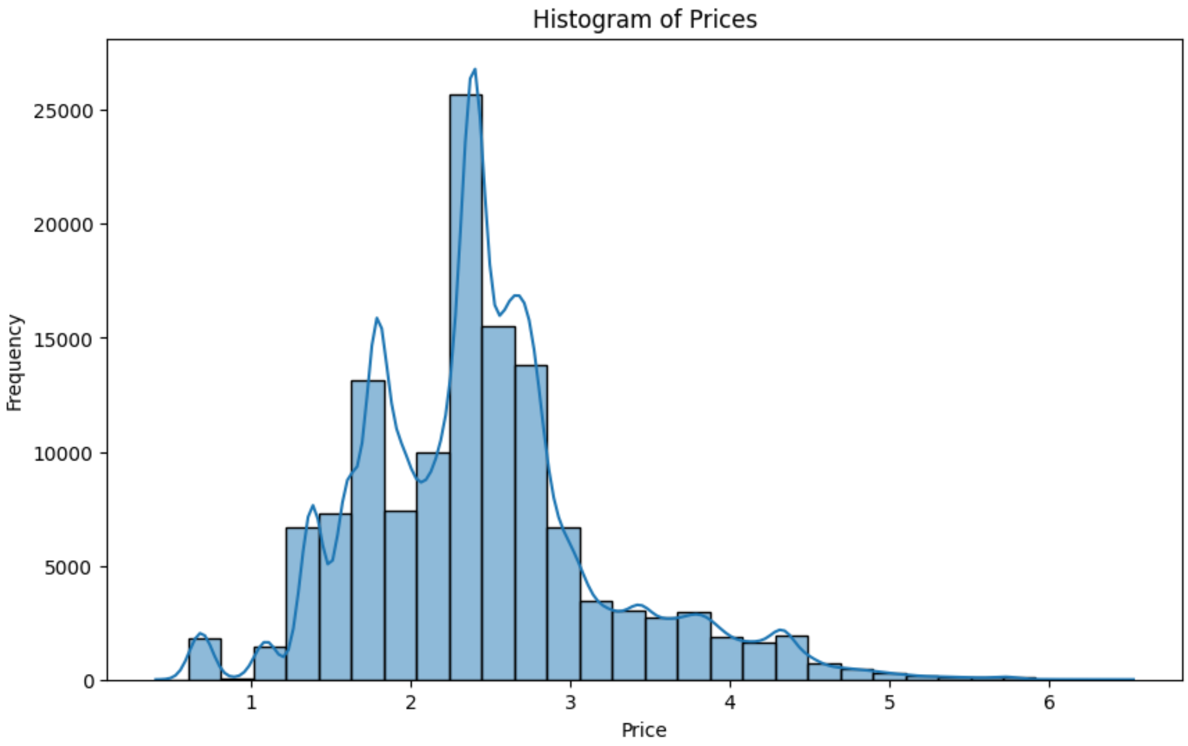
As the first step in the hypothesis testing, I did a revision of the distribution of my data since several tests assumed a normal distribution.



Img. 9

As we can see on the results the price distribution is heavily skewed to the right for that reason I made a log transformation to the price row to have a more normally distributed dataset.

Transformed dataset:



Img. 10

With that out of the way I decided to test my first hypothesis with a z-test as it is used for large enough samples and that it is designed to test hypothesis about the difference between two populations.

I obtained the following values from the cleaned dataset:

- p1: 0.04000923734195485
  - The proportion of bestsellers in the below $5 group
- p2: 0.012988408002506379
  - The proportion of bestsellers in the $5 or more group
- n1: 17321
  - The total number of books priced below $5
- n2: 111715
  - The total number of books priced $5 or more
- x1: 693
  - number of bestsellers in the below $5 group
- x2: 1451
  - number of bestsellers in the $5 or more group

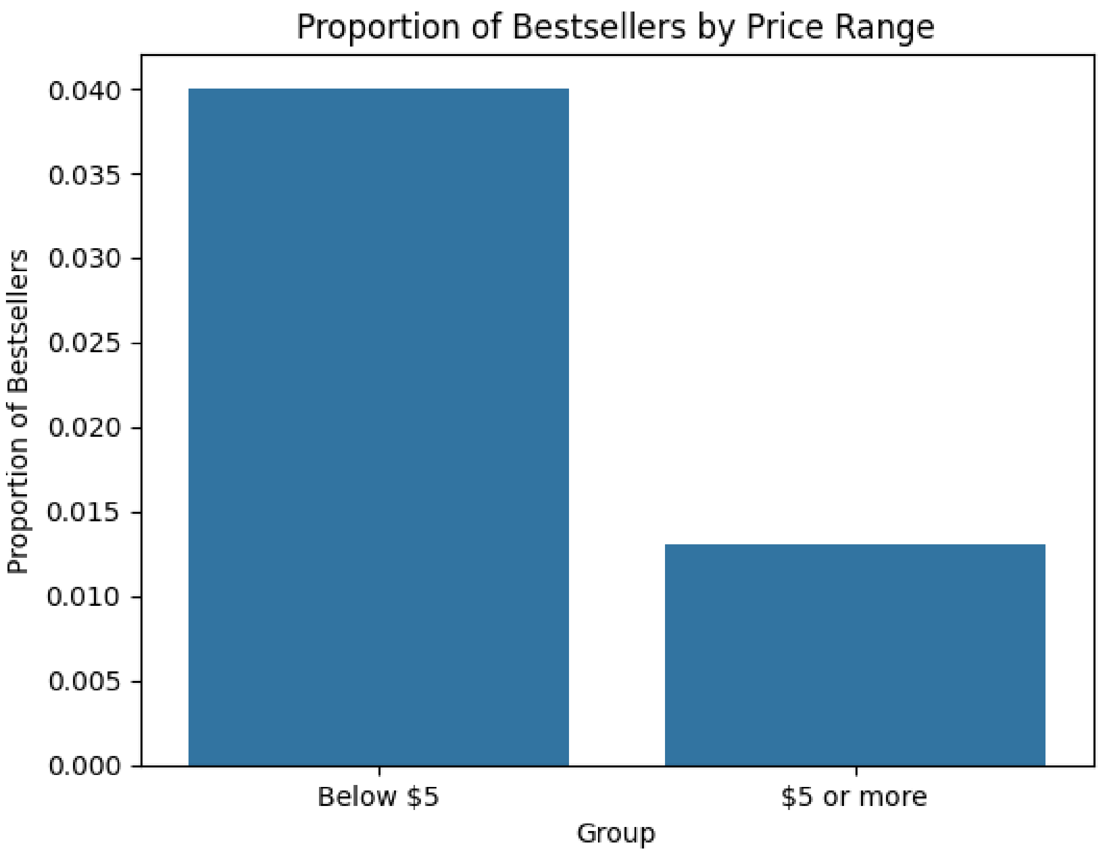
With this data I was able to calculate a pooled proportion of both populations to calculate the standard error and finally I was able to calculate the z score and p-value which I tested with an alpha of .05

All of this means that if one proportion is larger than the other by a significant size in which the p-value (calculated with the z score) is smaller than the alpha the null hypothesis will be rejected.

After conducting the significance test the null hypothesis was rejected leaving us with the Alternative.

In this hypothesis the proportion of the Bestsellers books priced below 5 is 4% approximately and the proportion priced 5 or above is lower than 1.3%. This means that it is indeed more likely to have a Bestseller book if it is priced under $5 as confirmed by the significance test.

To reinforce the results of this testing here is a histogram of the proportions in which we can clearly see the differences.



Img. 11

Next Steps in Analyzing the data

For the next steps in analyzing the data I would recommend implementing several analyses about how the publication date can affect the price and reviews of the books as well as maybe recollecting some more information for the reviews row since it has a lot of missing values.

We can also do analysis on how the authors perform with their books based on stars and reviews to determine the most liked authors and so much more.

Dataset Quality

I think this dataset was well collected even though it has a lot of missing values it is interesting in the way that all of this information is about eBooks of which I couldn’t find more data about.

I also think that if a company wanted to boost their eBooks this dataset would be a useful tool and even more practical with a sales row.