DATA CURATION FOR DATA STORYTELLING
AMAZON BEST SELLING BOOKS ANALYSIS (2009-2019)
BY OLUWATOMISIN EBUN AROKODARE

PROJECT OVERVIEW

Amazon marketplace has been a leading e-commerce platform for book sales since July 5th,1994. It was established by Jeff Bezos and will continue to be a leader in technological innovation, customer service, and a diverse marketplace. Amazon has revolutionized book sales and distribution by growing into one of the largest and most diverse online marketplaces in the world.

The impact of Amazon on the book industry and e-commerce landscape continues to grow, despite selling items other than books, such as electronics, clothing, and home goods, amazon continues to invest in enhancing its book-selling capabilities and values books as its core business.

PROJECT GOALS

The purpose of this project is to present insights derived from ten years (2009-2019) of analysis of Amazon best-selling books data that was scrapped from Amazon's website.

DATA GATHERING

Amazon data on best-selling books was scraped from the website using Beautiful Soup and Selenium. For loops and functions were used to obtain the URLs for each page of the best-selling books. The scraped data was saved to a text file, converted into a data frame, and finally exported into a CSV file.

AMAZON BEST SELLING BOOKS CATEGORISATION:

In data curation process which is the process of organization and integration of data collected from various sources. It was observed that Amazon best book selling does not have a definite categorization of books on their website as one book can be found in different departments. This departments include Biographies & Memoirs, Christian Books & Bibles, Arts & Photography, Business & Money, Children's Books, parenting and relationship, medical books, Comics & Graphic Novels etc.).

There is a need for categorizing the books in different departments into a distinct category for exploratory analysis. I checked the Goodreads website https://www.goodreads.com/ which is the world's largest site for readers and book recommendations to get the distinct category of each book.

I then checked Kaggle for a dataset that has the title and distinct category (Nonfiction and Fiction) of the books scrapped from amazon website for 2009-2019.This dataset will be merged with the amazon best-selling books dataset scrapped from the website.

```
In [3]: ## Importing the data from kaggle that has the genre of amazon best selling books
df2=pd.read_csv('Amazon_best_selling_books_data.csv')
df2
```

Out[3]:

	title	genre
0	The Lost Symbol	Fiction
1	The Shack: Where Tragedy Confronts Eternity	Fiction
2	Liberty and Tyranny: A Conservative Manifesto	Non Fiction
3	Breaking Dawn (The Twilight Saga, Book 4)	Fiction
4	Going Rogue: An American Life	Non Fiction
...	...	...
1081	There's No Place Like Space: All About Our Sol...	Non Fiction
1082	How to Draw 101 Animals (1)	Non Fiction
1083	Simply Keto: A Practical Approach to Health & ...	Non Fiction
1084	The Outsiders	Fiction
1085	The Try Not to Laugh Challenge - Would Your Ra...	Non Fiction

1086 rows × 2 columns

DATA CLEANING

After accessing the web scrapping dataset from the amazon website, several rows in the scraped dataset for 2009-2019 have missing and null values.

My assessment of the CSV file where the scrapped data was saved helped me find the missing values using the information provided on the website and I manually added them to the file using excel to preserve the quality of the scrapped data and avoid removing data that may affect our analysis. I realized that some of the books on the website had multiple formats with different values for (cover types). Some columns were also missing (Reviews, author, and prices).

Having checked and filled all missing values, the only missing values left were those from Amazon's website that were not available and are dropped from the datasets. There are no null values in the amazon best book selling dataset between (2009-2019).

```
best_selling_books.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 607 entries, 0 to 1836
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   title          607 non-null    object 
 1   genre          607 non-null    object 
 2   author         607 non-null    object 
 3   cover_type     607 non-null    object 
 4   ratings        607 non-null    float64
 5   ranks          607 non-null    int64  
 6   price          607 non-null    float64
 7   no_of_reviews  607 non-null    object 
 8   year           607 non-null    int64  
dtypes: float64(2), int64(2), object(5)
memory usage: 47.4+ KB
```

The Final Data frame for Amazon bestselling books between 2009 and 2019 is seen below:

In [20]:
```
best_selling_books
```

Out[20]:

	title	genre	author	cover_type	ratings	ranks	price	no_of_reviews	year
0	The Shack: Where Tragedy Confronts Eternity	Fiction	William P. Young	Paperback	4.6	2	6.99	41,124	2009
1	The Shack: Where Tragedy Confronts Eternity	Fiction	William P. Young	Paperback	4.6	27	6.99	41,124	2017
6	Liberty and Tyranny: A Conservative Manifesto	Non Fiction	Mark R. Levin	Hardcover	4.8	3	20.31	5,340	2009
7	Breaking Dawn (The Twilight Saga, Book 4)	Fiction	Stephenie Meyer	Hardcover	4.7	4	20.49	25,408	2009
9	Going Rogue: An American Life	Non Fiction	Sarah Palin	Hardcover	4.6	5	6.30	1,599	2009
...	...	...	...	...	...	...	...	...	...
1832	Me: Elton John Official Autobiography	Non Fiction	Elton John	Hardcover	4.7	75	9.90	27,342	2019
1833	The Road Back to You: An Enneagram Journey to ...	Non Fiction	Ian Morgan Cron	Hardcover	4.7	76	9.99	12,563	2019
1834	The Tattooist of Auschwitz: A Novel	Fiction	Heather Morris	Paperback	4.6	77	9.39	158,421	2019
1835	People of Walmart Adult Coloring Book: Rolling...	Non Fiction	Andrew Kipple	Paperback	4.7	78	12.96	11,494	2019
1836	The Complete Cookbook for Young Chefs: 100+ Re...	Non Fiction	Americaâ□□s Test Kitchen Kids	Hardcover	4.8	80	10.49	21362	2019

607 rows × 9 columns

In [21]:
```
best_selling_books.to_csv('AMAZON BEST SELLING BOOKS CLEANED DATA 2009-2019.csv')
```

Dataset Features Description

Each year contains a list of 100 best-selling books. The whole data contains 607 rows and 9 columns in total. The following columns include:

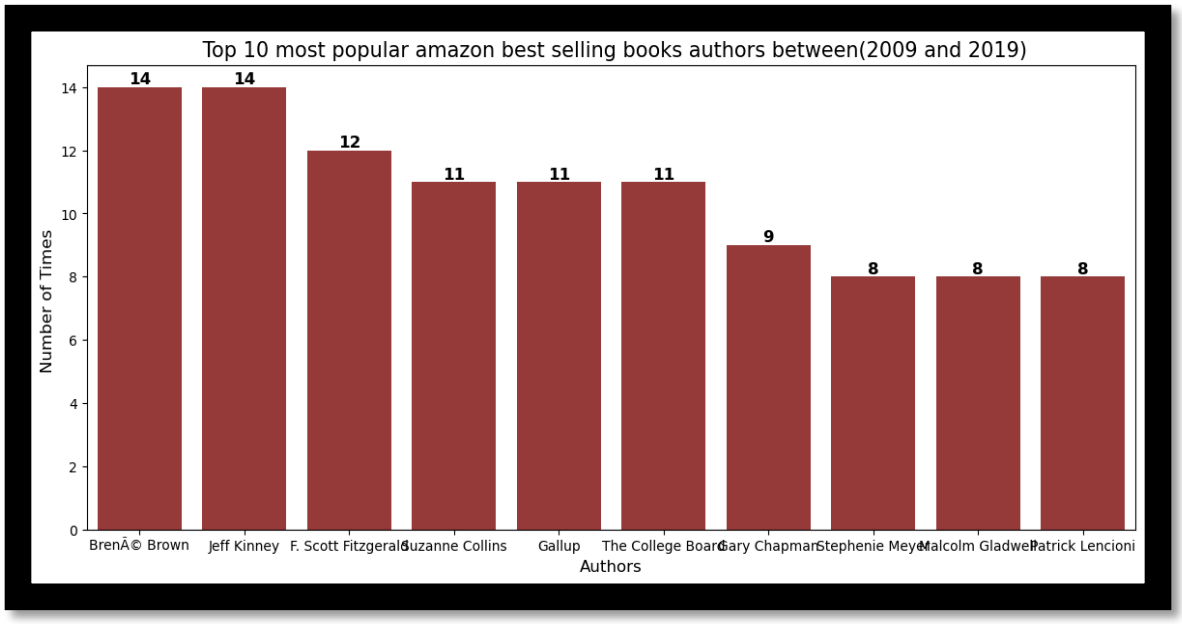
1. Title: the title of the book

2. Genre: this is the category or type of book (Fiction or Non-Fiction)
3. Author: this is the name of the author of a book.
4. Cover type: this is the cover type of each book.
5. Rating: this is the user rating that a book has.
6. Price: the price of each book
7. Rank: the ranking of each book in the top 100 best-selling for each year.
8. Year: this is the year each book appeared as a best-selling book.
9. No_of_review: this is the total reviews each book has.

Exploratory Data Analysis

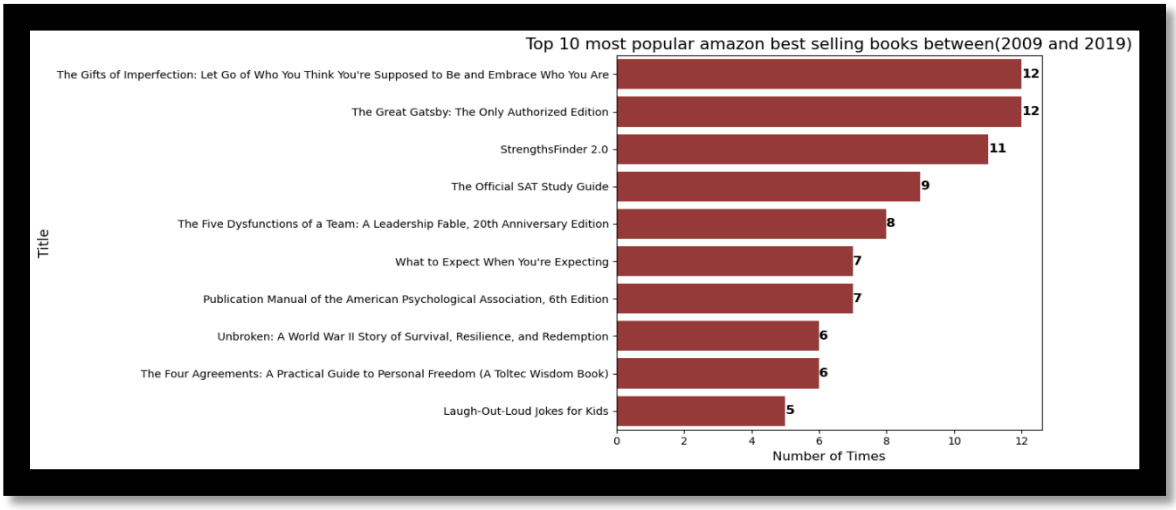
QUESTION 1. Most popular top 10 Amazon best books authors

Observation: Among the top 10 most popular Authors, Author BrenA Brown and Jeff Kinney are the top selling authors with 14 appearances in the top selling books between 2009 and 2019.



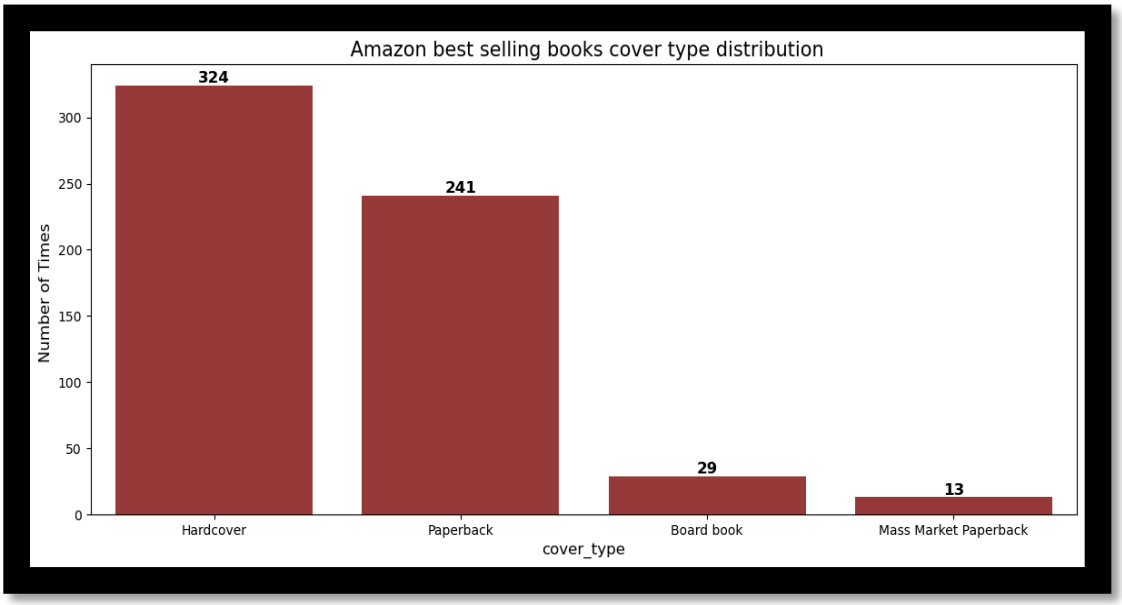
QUESTION 2. Most popular Top 10 Amazon best-selling Books

Observation: It was observed that the gift of imperfection, The great Gatsby, and Strength's finder 2.0 has the highest number of occurrences amongst the top 10 most popular books.



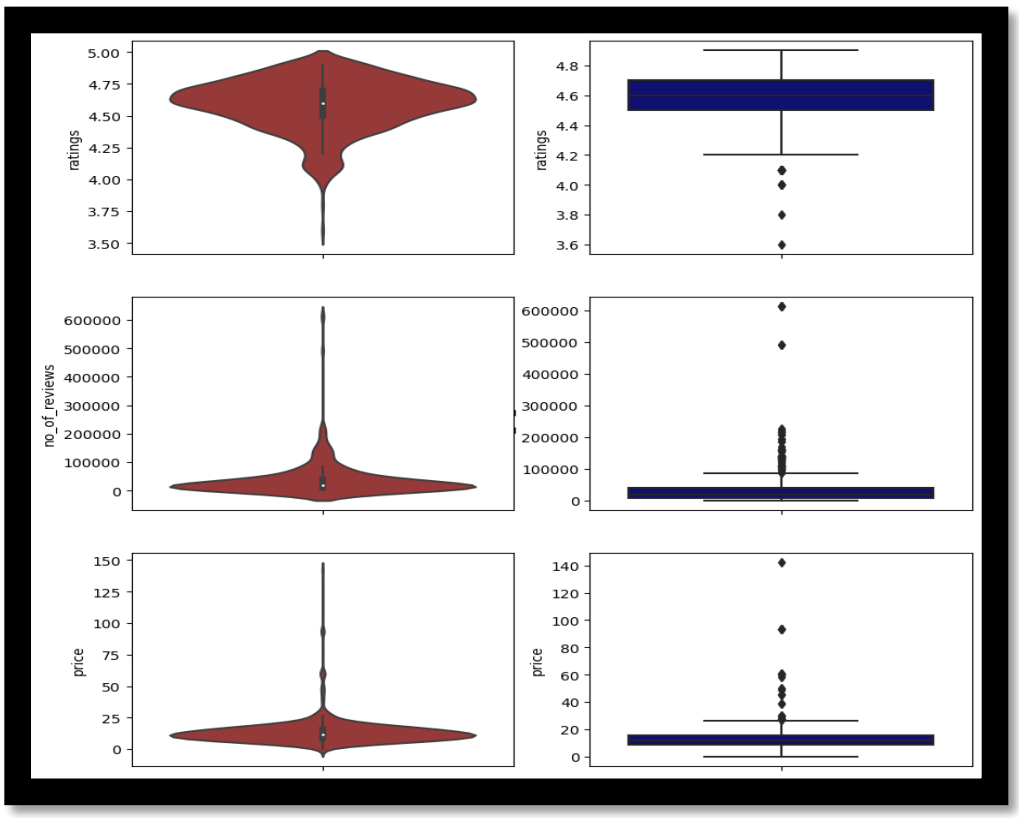
QUESTION 3. Amazon best selling book cover type distribution

Observation: According to Amazon's best-selling book list between 2009 and 2019, the most popular cover type is hardcover (324 titles); the least popular cover type is Mass Market paperback (13 titles).



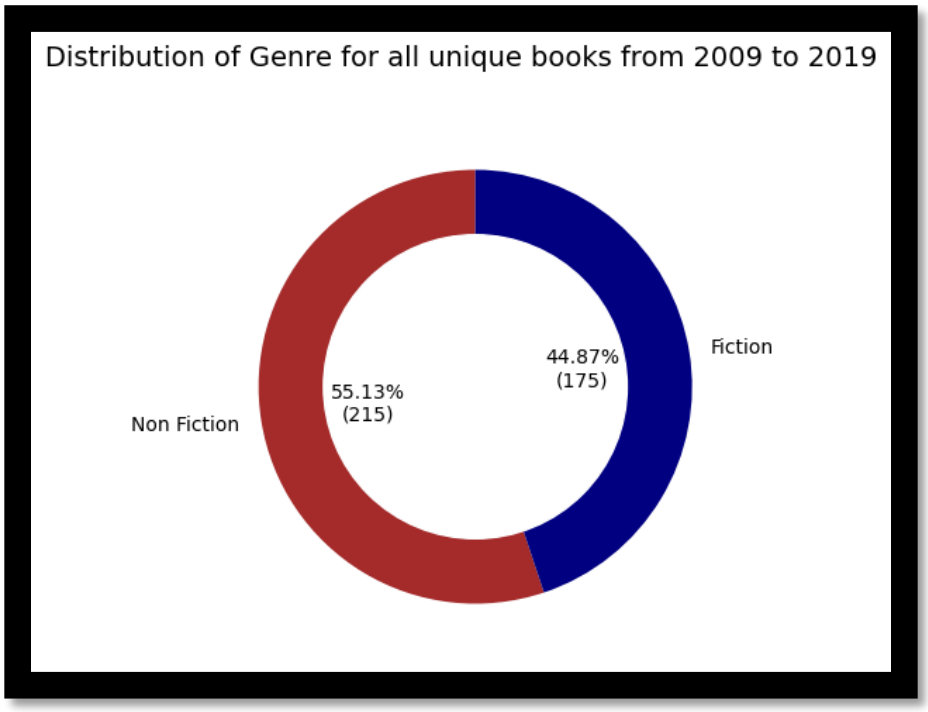
QUESTION 4. Distribution of Amazon best selling books Rating, Reviews and Price

Observation: The box plot and violin plots show that top outliers in the ratings variable were books that had ratings below 4.0 ratings. It was observed that the outliers in the number of reviews variable is a book that had about 600,000 reviews. The outlier in the price range is a book that was sold for about 140 dollars. It was observed that most user ratings occurred between 4.5 and 4.8, most occurring price range was between 10 and 18 dollars, and most occurring number of reviews were below 40,000.



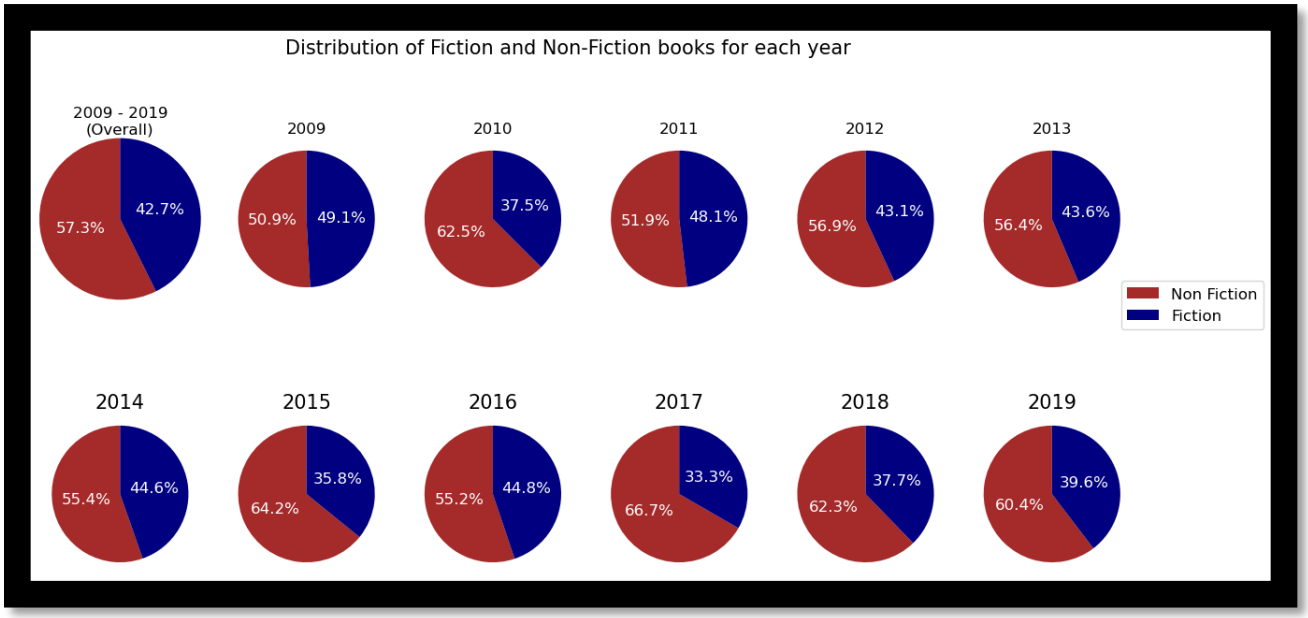
QUESTION 5. Amazon best selling unique books distribution of Genre.

Observation: The Amazon best book selling with Non-Fiction category was the most popular category from 2009 to 2019, with 55.13% (percent) of the 215 books being Non-Fiction and 45.6% (percent) being Fiction.



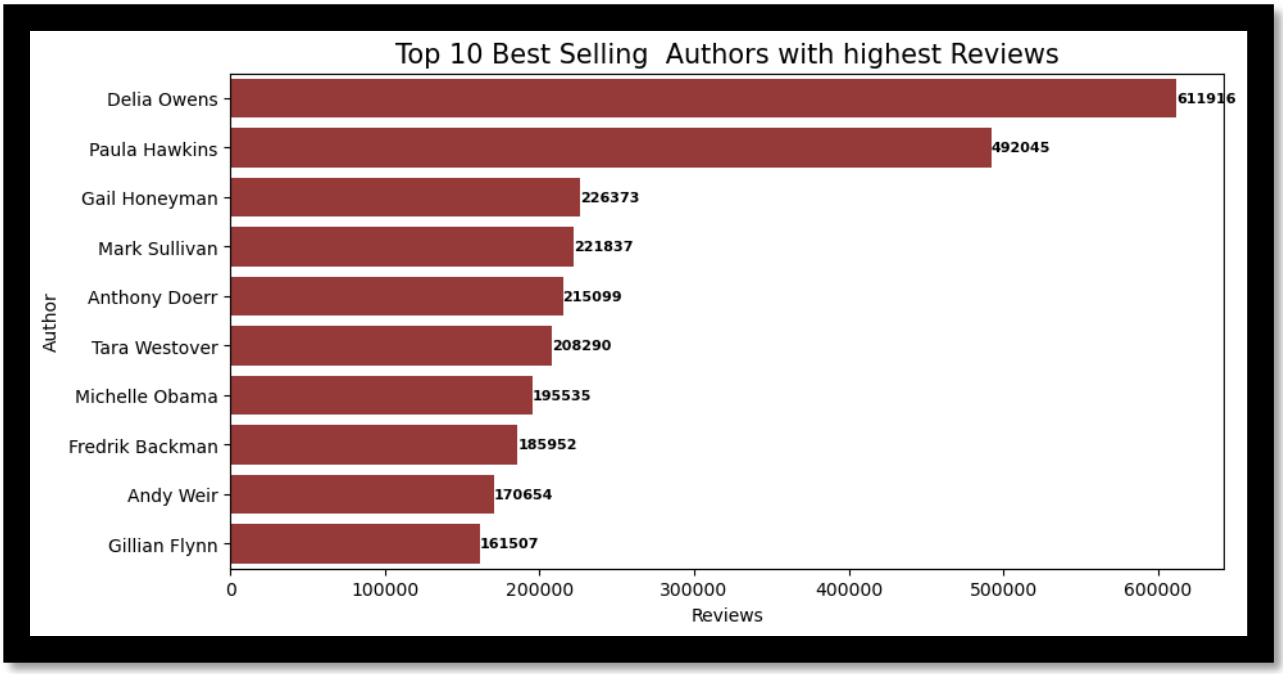
QUESTION 6. Amazon best selling unique books distribution of Genre for each year.

Observation: It was observed that the highest percentage (66.7%) of Nonfiction books were sold in 2017 and the highest percentage of fiction books were sold in 2009. However, the lowest percentage of books for Nonfiction were sold in 2009 and lowest percentage of fiction books sold are in 2007.



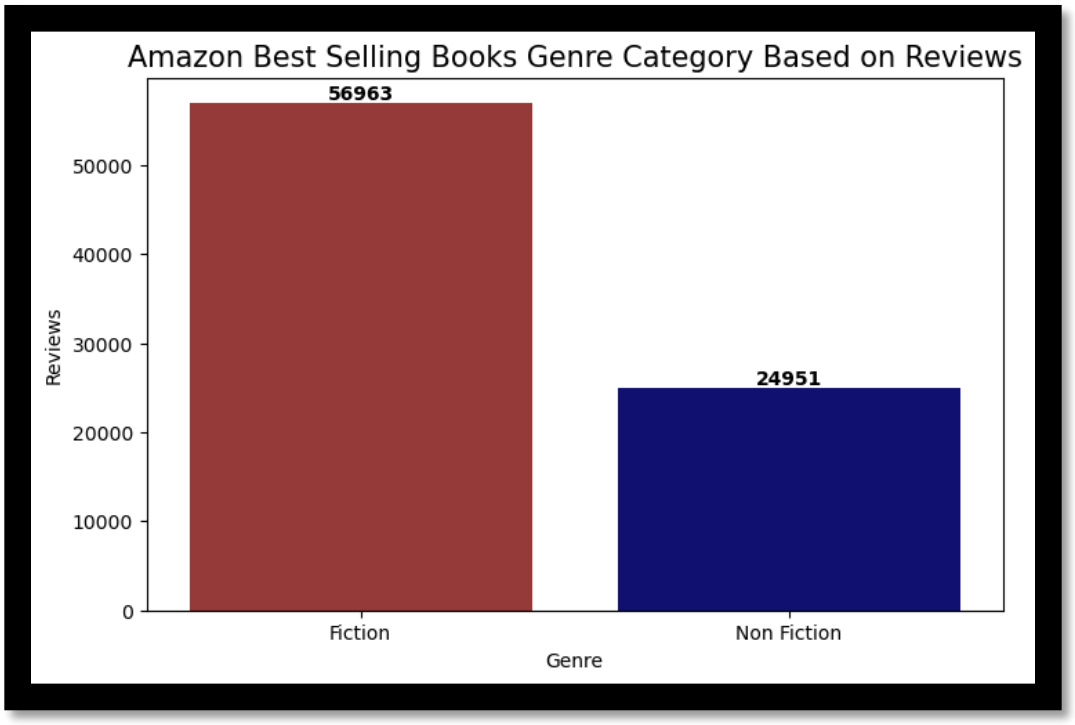
QUESTION 7. Top 10 best selling Authors with highest Reviews

Observation: Paula Hawkins and Delia Owens are the two best selling authors with the highest number of books reviews in the top 10. However, we can see that Delia Owens is that one outlier that appears to have had the highest number of reviews, approximately 611,916 which is much higher than the number of reviews of other authors.



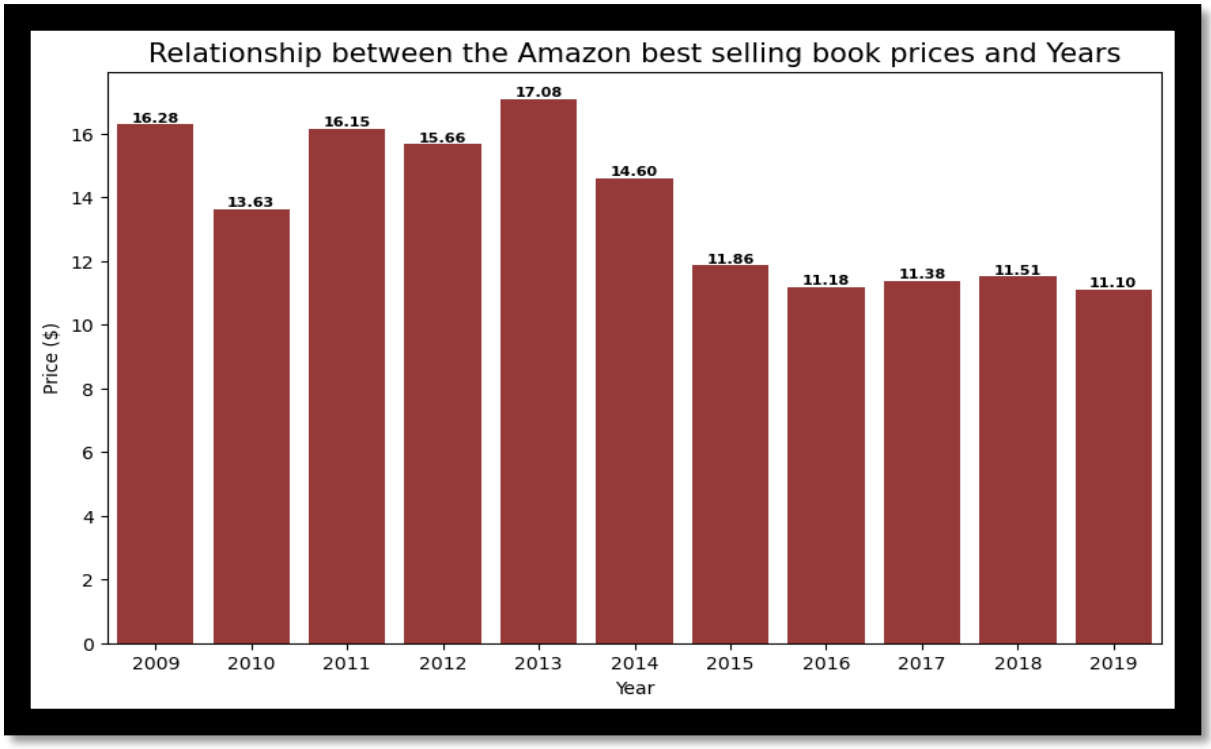
QUESTION 8. Amazon best selling books Genre category based on Reviews.

Observation: It was observed that the top 10 best-selling fiction titles received approximately 56,963 reviews on average, while the top 10 best-selling non-fiction titles received approximately 24,952 reviews on average. We can see that Fiction best-selling books received more reviews than non-fiction between 2009 and 2019.



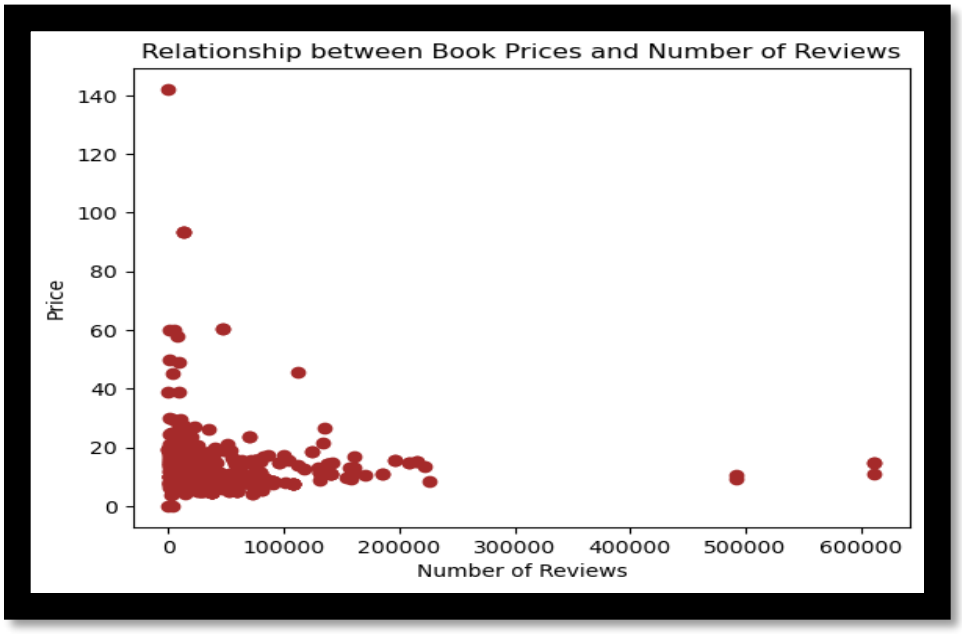
QUESTION 9. Relationship between the Amazon best selling book prices and Years

Observation: The average price of Amazon's best-selling books has decreased over the years; in 2013 it was around 17.08 dollars, but by 2019, it was approximately 11. 10dollars. Overall, cost of books was decreasing with Years.



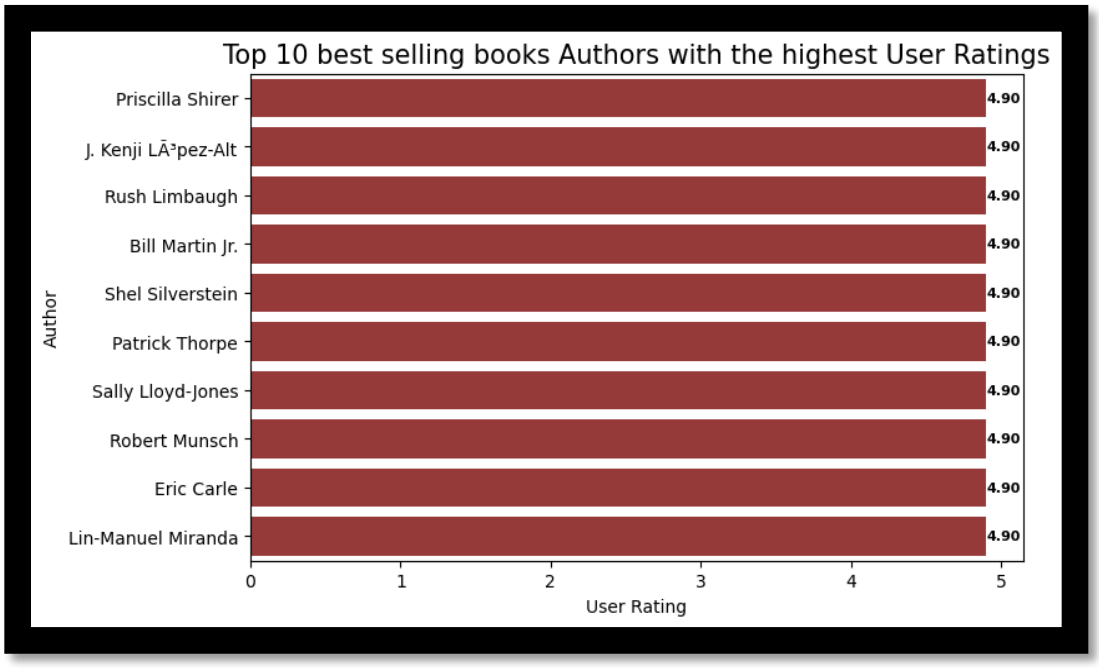
QUESTION 10. Relationship between Amazon best selling book Prices and number of reviews

Observation: From the scatterplot, there is a negative correlation between prices and the number of reviews. This indicates that as the number of reviews increases, the price decreases slightly, but the correlation is not very strong, suggesting that the relationship is not significant. In other words, a correlation does not imply causation, so we cannot conclude that a decrease in prices directly results from an increase in reviews. We also observed that there are outliers in the prices of books which are above 140 dollars and number of reviews which is above 600,000 reviews.



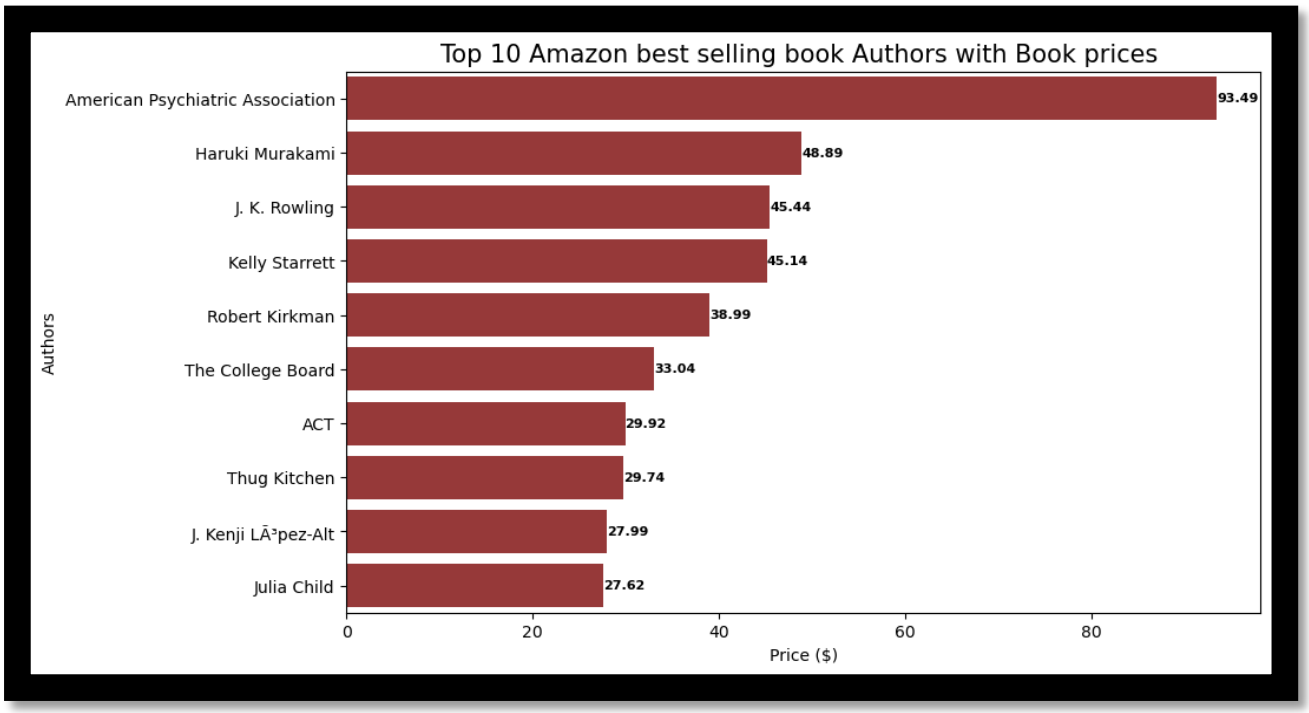
QUESTION 11. Top 10 best selling books Authors with the highest User Ratings

Observation: The top 10 amazon best selling books Authors with the highest User ratings has 4.90 stars for their books.



QUESTION 12. Top 10 Amazon best selling book Authors with Book prices

Observation: It can be observed that the author with highest priced books is American Psychiatric Association who charges 93.49 dollars. However, American Psychiatric Association is an outlier having the maximum price of above 80dollars.

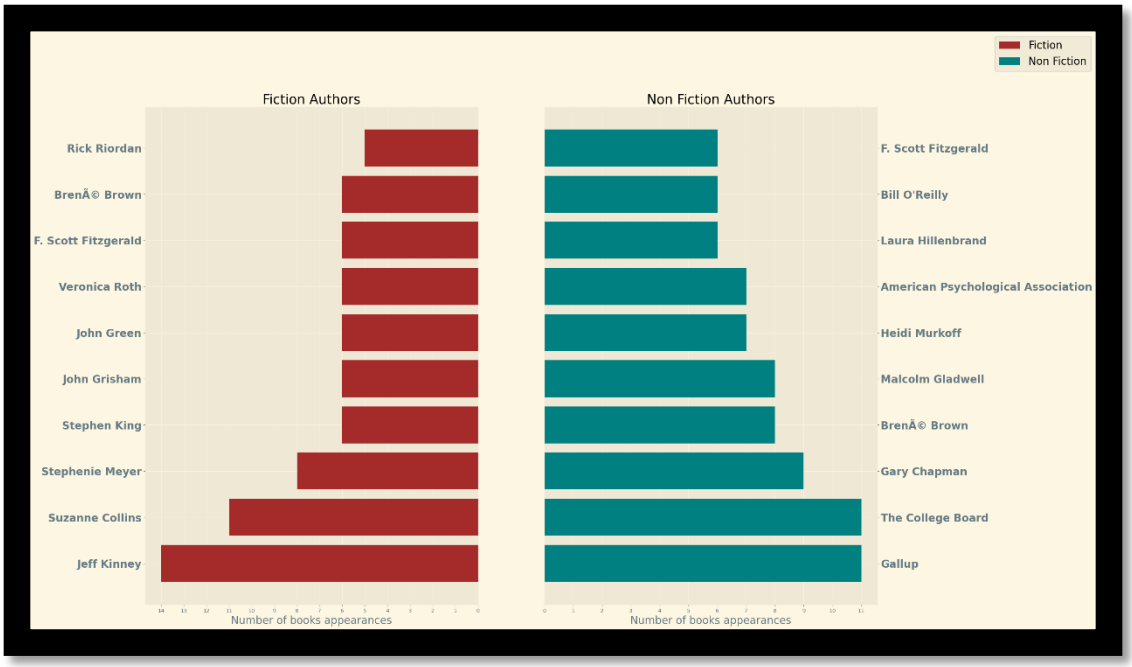


QUESTION 13. Is there any correlation between the amazon best selling books variables?

Observation: From the amazon best books selling, the heatmap shows that year has a positive correlation with the book's ratings. A strong correlation exists between the number of reviews and the year, suggesting that the number of reviews tends to increase over time, however, the other correlations are not significant enough to draw strong conclusions about their relationships.

QUESTION 14. Top 10 Best Selling Authors based on Genre and Number of books Appearances.

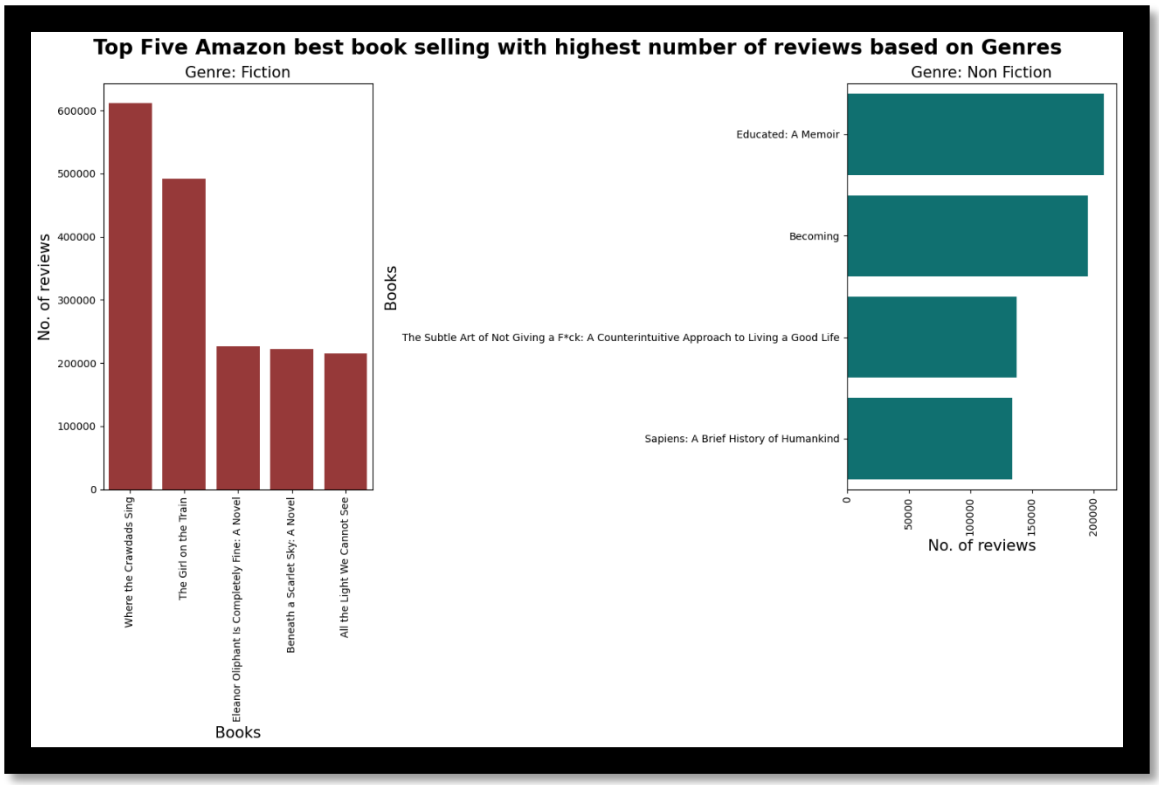
Observation: Among the top 10 Best Selling Authors based on Genre and Number of their books appearances
Jeff Kinney and Suzanne Collins rank the top two best selling authors for Fiction genre category between 2009 and 2019. The top two best selling authors in the Non-Fiction genre category between 2009 and 2019 are Gallup and the College Board Authors.



QUESTION 15: Amazon Top 5 Selling Books Based on Number of Reviews

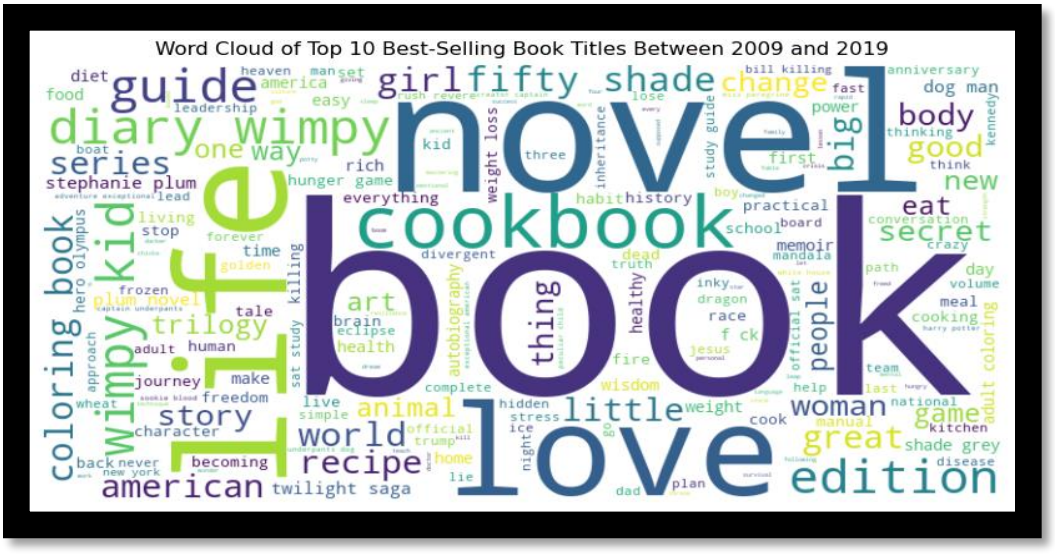
Observation: Amazon top 5 best Fiction selling books based on user reviews are Where the Crawdads Sing, The Girl on the train, Eleanor Oliphant is completely fine: A novel, beneath a Scarlet Sky and All the light we cannot see. However, Where the Crawdads Sing is that outlier with the highest number of reviews of about 611,916 reviews.

Amazon top 5 best Non-Fiction selling books based on user reviews are Educated: A Memoir, Becoming, The Subtle Art of Not Giving a Fuck: A Counterintuitive Approach to Living a Good Life, Sapiens: Brief History of Humankind. However, Where the Educated: A Memoirs is that outlier with the highest number of reviews of about 208,290 reviews.



QUESTION 16: A word Cloud of the Top Best Selling Books Between 2009 and 2019

Observation: The Word cloud for amazon best selling books provides at-a-glance view and It makes it easier to communicate the main themes or trends in huge text sets to a wider audience when they can be summarized and visualized in such an appealing and straightforward way.



CONCLUSION

- Among the top 10 most popular Authors, Author BrenA Brown and Jeff Kinney are the top selling authors with 14 appearances in the top selling books between 2009 and 2019.

- It was observed that the gift of imperfection, The great Gatsby, and Strength's finder 2.0 has the highest number of occurrences amongst the top 10 most popular books.

- According to Amazon's best selling book list between 2009 and 2019, the most popular cover type is hardcover (324 titles); the least popular cover type is Mass Market paperback (13 titles).

- The box plot and violin plots show that top outliers in the ratings variable were books that had ratings below 4.0 ratings. It was observed that the outliers in the number of reviews variable is a book that had about 600,000 reviews. The outlier in the price range is a book that was sold for about 140 dollars. It was observed that most user ratings occurred between 4.5 and 4.8, most occurring price range was between 10 and 18 dollars, and most occurring number of reviews were below 40,000.

- The Amazon best book selling with Non-Fiction category was the most popular category from 2009 to 2019, with 55.13% (percent) of the 215 books being Non-Fiction and 45.6% (percent) being Fiction.

- It was observed that the highest percentage (66.7%) of Nonfiction books were sold in 2017 and the highest percentage of fiction books were sold in 2009. However, the lowest percentage of books for Nonfiction were sold in 2009 and lowest percentage of fiction books sold are in 2007.

- Paula Hawkins and Delia Owens are the two best selling authors with the highest number of books reviews in the top 10. However, we can see that Delia Owens is that one outlier that appears to have had the highest number of reviews, approximately 611,916 which is much higher than the number of reviews of other authors.

- It was observed that the top 10 best-selling fiction titles received approximately 56,963 reviews on average, while the top 10 best-selling non-fiction titles received approximately 24,952 reviews on average. We can see that Fiction best-selling books received more reviews than non-fiction between 2009 and 2019.

- The average price of Amazon's best-selling books has decreased over the years; in 2013 it was around 17.08 dollars, but by 2019, it was approximately 11. 10dollars.Overall, cost of books was decreasing with Years.

- From the scatterplot, there is a negative correlation between prices and the number of reviews. This indicates that as the number of reviews increases, the price decreases slightly, but the correlation is not very strong, suggesting that the relationship is not significant. In other words, a correlation does not imply causation, so we cannot conclude that a decrease in prices directly results from an increase in reviews. We also observed that there are outliers in the prices of books which are above 140 dollars and number of reviews which is above 600,000 reviews.

- The top 10 amazon best selling books Authors with the highest User ratings have 4.90 stars for their books.

- It can be observed that the author with highest priced books is American Psychiatric Association who charges 93.49 dollars. However, American Psychiatric Association is an outlier having the maximum price of above 80dollars.

- From the amazon best books selling, the heatmap shows that year has a positive correlation with the book's ratings. A strong correlation exists between the number of reviews and the year, suggesting that the number of reviews tends to increase over time, however, the other correlations are not significant enough to draw strong conclusions about their relationships.

- Among the top 10 Best Selling Authors based on Genre and Number of their book's appearances-Jeff Kinney and Suzanne collins rank the top two best selling authors for Fiction genre category between 2009 and 2019.The top two best selling authors in Non-Fiction genre category between 2009 and 2019 are Gallup and the College Board Authors.

- Amazon top 5 best Fiction selling books based on user reviews are Where the Crawdads Sing, The Girl on the train, Eleanor Oliphant is completely fine: A novel, beneath a Scarlet Sky and All the light we cannot see. However, Where the Crawdads Sing is that outlier with the highest number of reviews of about 611,916 reviews.

- Amazon top 5 best Non-Fiction selling books based on user reviews are Educated: A Memoir, becoming, The Subtle Art of Not Giving a Fuck: A Counterintuitive Approach to Living a Good Life, Sapiens: A Brief History of Humankind. However, Where the Educated: A Memoirs is that outlier with the highest number of reviews of about 208,290 reviews.

- The Word cloud for amazon best selling books provides at-a-glance view and it makes it easier to communicate the main themes or trends in huge text sets to a wider audience when they can be summarized and visualized in such an appealing and straightforward way.