Data Curation Project

By

Etieneabasi Kingsley Effiong

Topic: Books Sold and Enlisted on Amazon Between 2009 And 2019

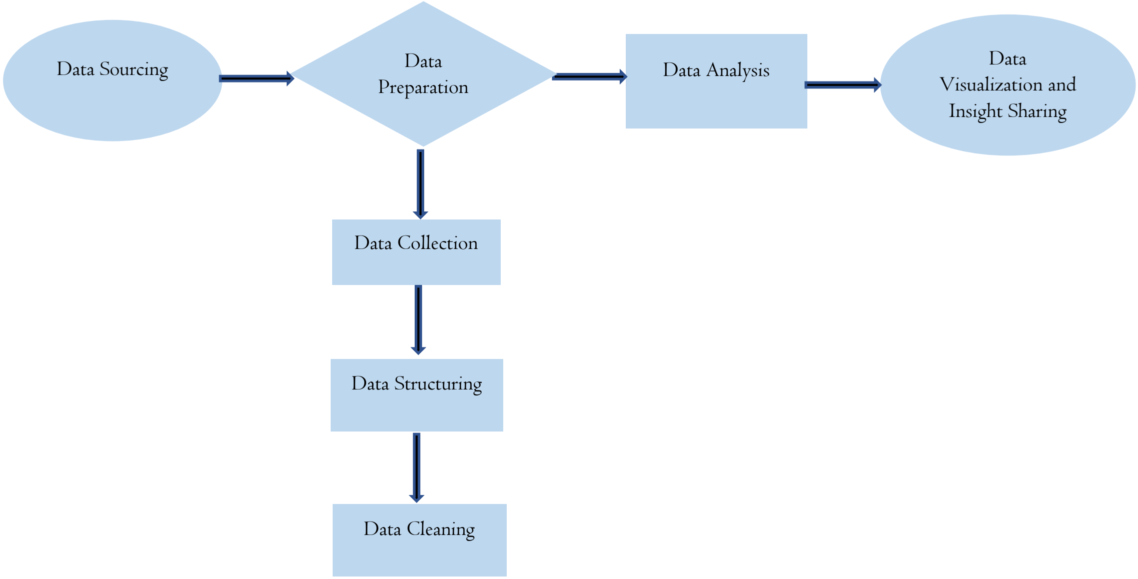
Amazon.com has grown to become one of the world's largest e-commerce platforms since its inception. Amazon has become a go-to destination for many customers due to its vast selection of books, music, videos, and software. However, the company's rapid growth has created difficulties in managing and utilizing the vast amount of data that it generates. The company collects a variety of data about each customer, including their browsing history, product purchases, product reviews, and other website interactions. Amazon values this data because it allows the company to better understand its customers and create better products for them.

Aim and Objectives

The goal of this project is to compare the number of books displayed for sale to the types of books purchased frequently by Amazon customers by successfully identifying relevant data, cleaning the data, and then transforming and providing insights.

Flow Process

The steps taken are illustrated with the flowchart below:



Data Sourcing

The website consists of books listed on Amazon books category, the customer rating for each book, the author of the book, the type of book, the price of the book etc. The data was extracted from the link - [Amazon's e-commerce platform for books].

Data Preparation

The following steps were followed in the data preparation process:

- **Data collection:** In order to determine which tags to call out while collecting data from the website, the Hypertext Markup Language (HTML) tags and elements managing the data to be scraped and curated were identified using the page inspection features on the Mozilla Firefox browser program.
- **Data structuring:** Beautiful Soup was used to extract the data, which was then exported as a Comma-Separated Values (CSV) file called 'Amazon Books.' The CSV file includes four columns: BookName, Type, Date, and Rate.
- **Data Cleaning:** The obtained dataset included books from the website between 1975 and 2022. Duplicate rows were removed, as were columns and rows that were irrelevant to the analysis. A new table was created with dates ranging from 2009 to 2019 and exported as a CSV file called 'Books Amazon.'

Data Analysis and Visualisation

After the analysing process was done, visualisation was carried out to comprehend the various datasets using plots with visual representations including histograms, box plots etc.

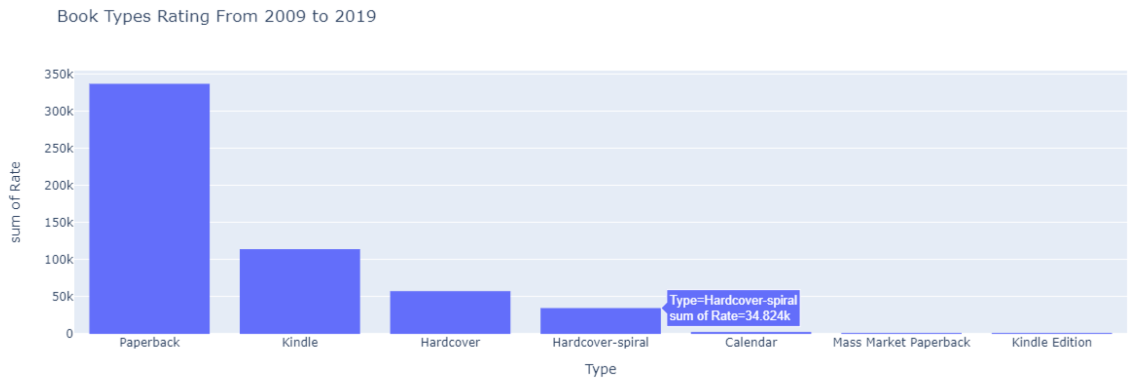


Figure 1: Histogram Chart

The figure above depicts various book types and their ratings, with ‘Paperback’ being rated the highest and ‘Kindle Edition’ rated the lowest.

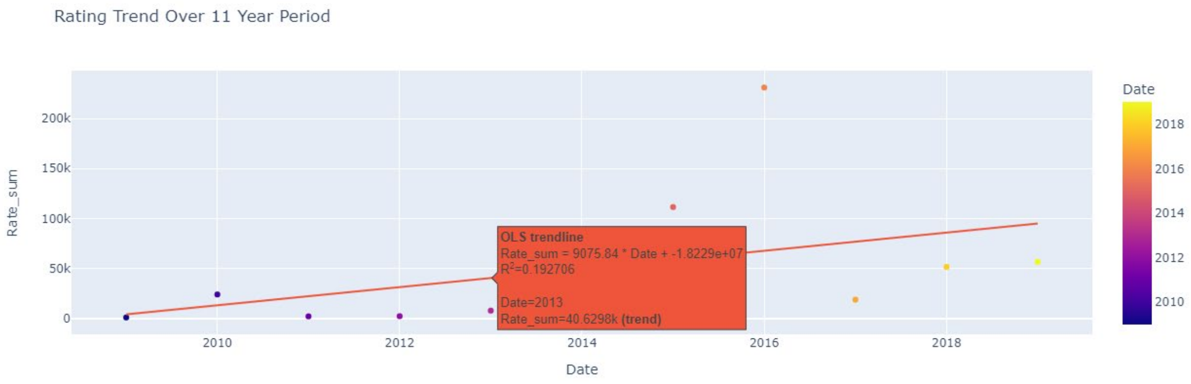


Figure 2: Scatter Plot
The plot above shows the number of ratings received each year. Hovering over the trendline reveals the line's equation, rate sum, and its R-squared value (coefficient of determination).

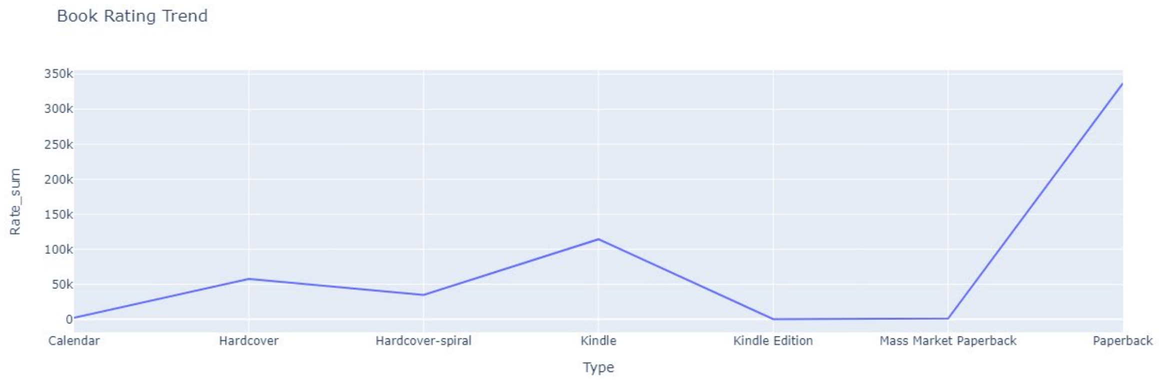


Figure 3: Line Plot
The line plot depicts the pattern of rating received according to the type of book over the years in an ascending order.

Results

Based on the curated data, the highest rated book is titled 'It Ends with Us: A Novel (1)', with rating from 16,7843 customers. 'Paperback' is the most popular book type, receiving rating from 337,128 customers from 2009 to 2019. Also, the year 2016 received the highest number of ratings which was from 230,980 customers, followed by the year 2015 that received a total rating from 111,565 customers.

On the other hand, there are 7 books that fell within the least rated with a rating of 1, containing only paperback and kindle book types with the year 2019 occurring the most. The book type with the lowest rating from 2009 to 2019 is "Kindle Edition," which has a rating sum of 33. The year 2009 received the least rating of 1,246 from customers.

Finally, a total of 546,762 customers rated 140 books between the years 2009 and 2019.

Conclusion and Recommendation

This article has looked at the ratings of 140 books consisting of various book types over a period of 11 years. Data that was obtained followed the checklist provided by the Data Curation Network (DCN) curators and can be said that it met the FAIR (Findable, Accessible, Interoperable, Reusable) evaluation.

Further information on all analysed and visualised data including the codes to execute the project, can be accessed via GitHub.