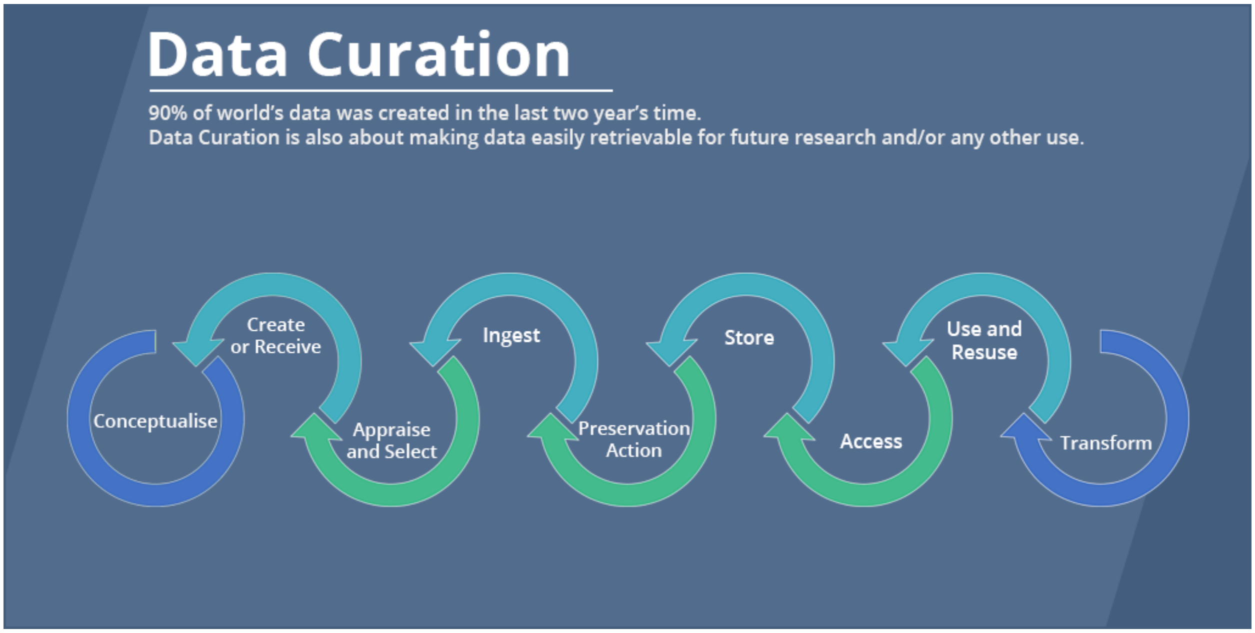
Data Curation

By Elizabeth Queen Okon

Data Curation is the process of creating, organizing, and maintaining data sets so they can be assessed and used by people looking for information.
It involves the annotation, presentation, and publication of data in such a fashion that the value is preserved for a long time. Data Curation starts with data collection either from data warehouses or lakes.



Data Curation

90% of world's data was created in the last two year's time.
Data Curation is also about making data easily retrievable for future research and/or any other use.

Steps in a Data Curation Workflow

1. Documenting the data
2. Asking questions
3. Translating into open formats
4. Assessing FAIRness
5. Cleaning and Validating

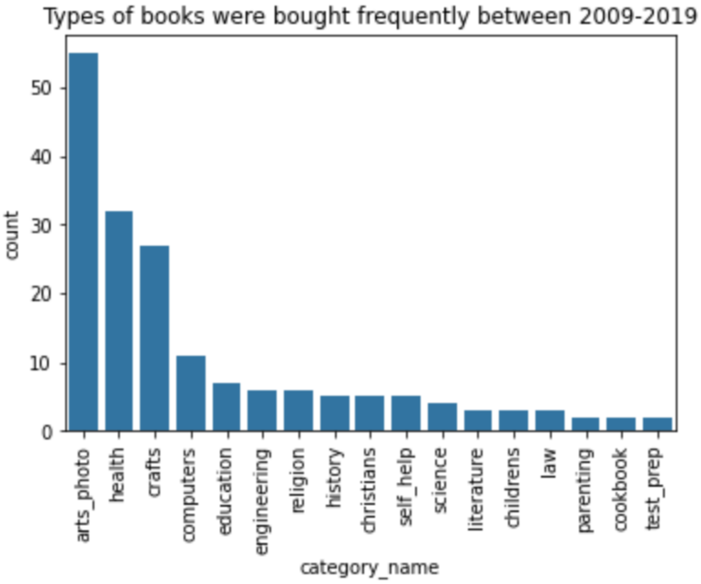
Amazon is a popular e-commerce platform where books are sold. This project aims to curate data from this platform and unravel what types of books were bought frequently between 2009 and 2019.

- Data Collection

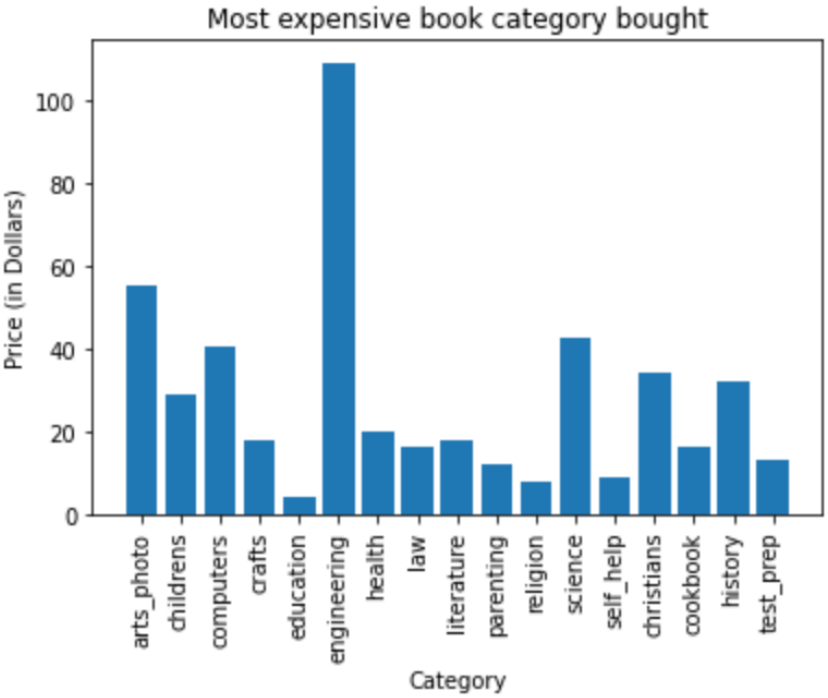
  For this project, data was scrapped off Amazon.

The collected data was then cleaned, analyzed, and visualized to answer some questions which are as follows :

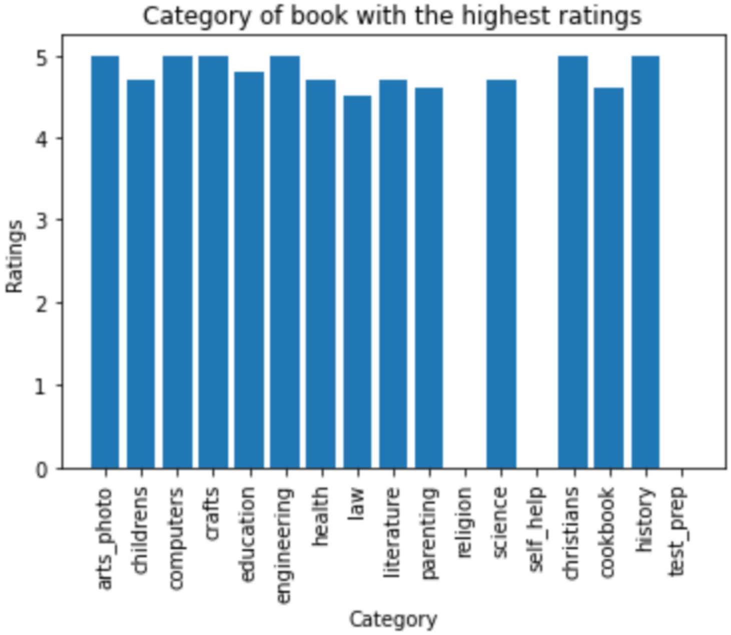
1. **What types of books were bought frequently between 2009 and 2019?**
   After analyzing and visualizing the data set it was discovered that Arts / Photo was the type of book bought frequently between 2009 and 2019. This can be seen in the image below.



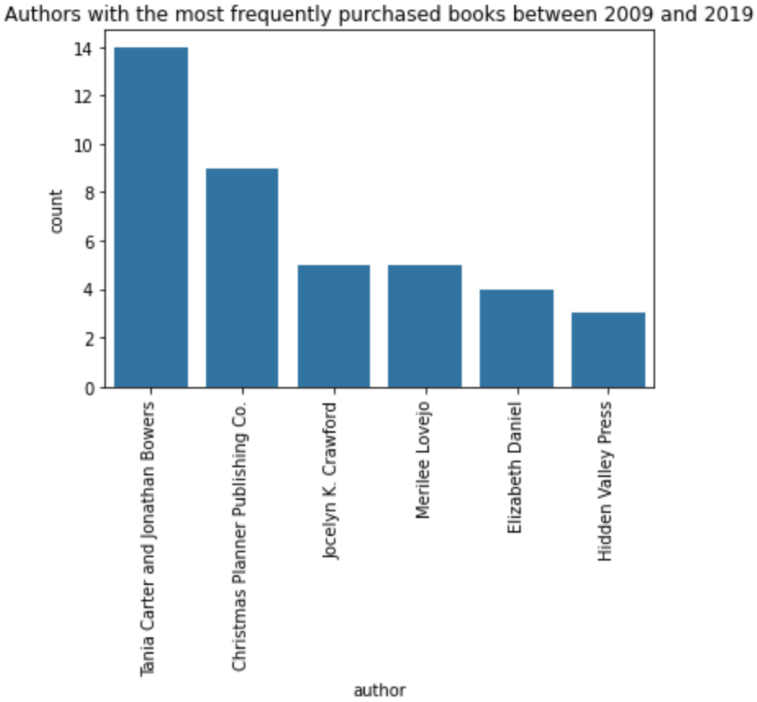
2. **What category of books was the most expensive?**
   I was interested in getting some additional pieces of information so I went further to analyze the data. I found out the most expensive category of books was Engineering books.



3. **What category of books had the highest rating?**
After analyzing, I found that Arts / Photo, Computers, Crafts, Christians, and History books had the highest ratings with 5 out of 5 stars.



4. Finally, I was interested in finding out the authors with the most frequently purchased books between 2009 and 2019. I discovered that Tania Carters and Jonathan Bowers had the highest number of books bought with a count of 14.



Why is Data Curation important?

Data Curation is an important part of data management. It allows companies to store sustainable and accessible data to share and apply self-service analytics.