# Book Recommender System

Janhavi Shembade, Data Science Trainee @almabetter

# ABSTRACT:

Today the amount of information on the internet grows very rapidly and people need some instruments to find and access appropriate information. One of such tools is called a recommendation system. Recommendation systems help to navigate quickly and receive necessary information. Generally they are used in Internet shops to increase the profit. This paper proposes a quick and intuitive book recommendation system that helps readers to find appropriate books to read next. The overall architecture is presented with it’s detailed description. We used a collaborative filtering method based on Pearson correlation coefficient. Finally the experimental results based on the online survey are provided with some discussions.

*Keywords- recommendation system, collaborative filtering*

**INTRODUCTION:**

Nowadays the amount of information, especially on the Internet, grows very rapidly. Finding necessary information becomes more difficult. Recommendation systems aim to solve this kind of problems.With the help of them one can quickly access relevant information without searching the web manually. As such many websites today benefit from recommendation systems to promote and sell their products. There is a wide range of products like music, movies, articles and etc. that can be recommended to the customer based on their profiles in internet shops or even social networks, browsing history such as visited links, browsing activity like number and time of visits and other online behavior. Online shops are increasing their sales using such technologies.

In this paper we propose using recommendation systems for recommending books. We developed a system, which learns user preferences by asking to rate books and choosing favorite categories and then generate the list of books user most probably would like to read.

**BUSINESS UNDERSTANDING:**

During the last few decades, with the rise of Youtube, Amazon, Netflix, and many other such web services, recommender systems have taken more and more place in our lives. From e-commerce (suggest to buyers articles that could interest them) to online advertisement (suggest to users the right contents, matching their preferences), recommender systems are today unavoidable in our daily online journeys.

In a very general way, recommender systems are algorithms aimed at suggesting relevant items to users (items being movies to watch, text to read, products to buy, or anything else depending on industries). Recommender systems are really critical in some industries as they can generate a huge amount of income when they are efficient or also be a way to stand out significantly from competitors. The main objective is to create a book recommendation system for users.

**DATA DESCRIPTION:**

# THERE ARE 3 CSV FILE FOR BOOK RECOMMENDER SYSTEM

# Users:-

Contains the users. Note that user IDs (User-ID) have been anonymized and map to integers. Demographic data is provided (Location, Age) if available. Otherwise, these fields contain NULL values.

**Books:-**

Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given (Book-Title, Book-Author, Year-Of-Publication, Publisher), obtained from Amazon Web Services. Note that in the case of several authors, only the first is provided. URLs linking to cover images are also given, appearing in three different flavors (Image-URL-S, Image-URL-M, Image-URL-L), i.e., small, medium, large. These URLs point to the Amazon website.

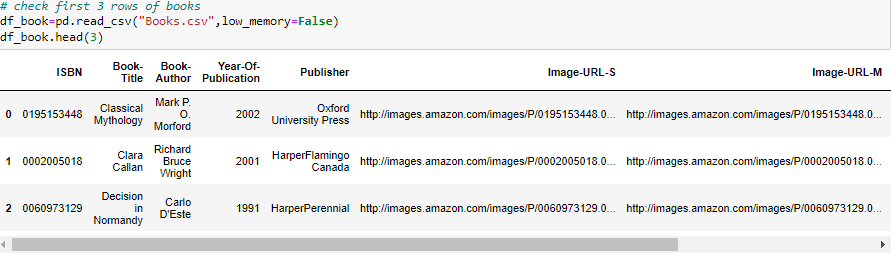
# Ratings:-

Contains the book rating information. Ratings (Book-Rating) are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

**DATASET PREPARATION:**

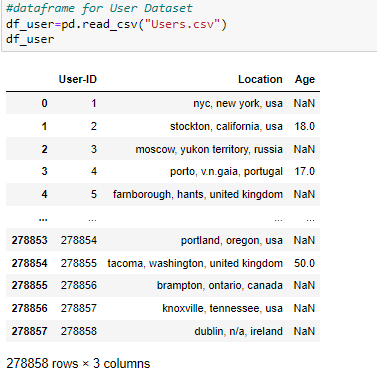
**Book Data:-**

There are 8 columns and 271360

****

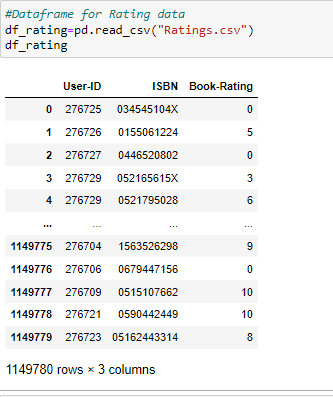
**User Data:-**

There are 3 columns and 278858 rows.



**Ratings Data:-**

There are 1031136 3 columns

****

**EXPLORATORY DATA ANALYSIS:**

If we want to explain EDA in simple terms, it means trying to understand the given data much better, so that we can make some sense out of it. It was also used to produce a value distribution and identify missing values, and outliers.

EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures. In this chapter, we are going to discuss the steps involved in performing topnotch exploratory data analysis

In statistics, A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task.EDA in Python uses data visualization to draw meaningful patterns and insights

* **DATA ANALYSIS:**

This is one of the most crucial steps that deals with descriptive statistics and analysis of the data. The main tasks involve summarizing the data, finding the hidden correlation and relationships among the data, developing predictive models, evaluating the models, and calculating the accuracies. Some of the techniques used for data summarization are summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, searching, grouping, and mathematical models.

* **DATA CLEANING**

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

Irregularities are of different types of data.

* Missing Values:-So In this dataset contains 0 values we need to fill with its mean value
* Incorrect Format
* Incorrect Headers:- All headers are in good format
* **DATA TRANSFORMATION:**

Data transformation is the process of normalizing and aggregating the data to further improve the efficiency and accuracy of data mining.

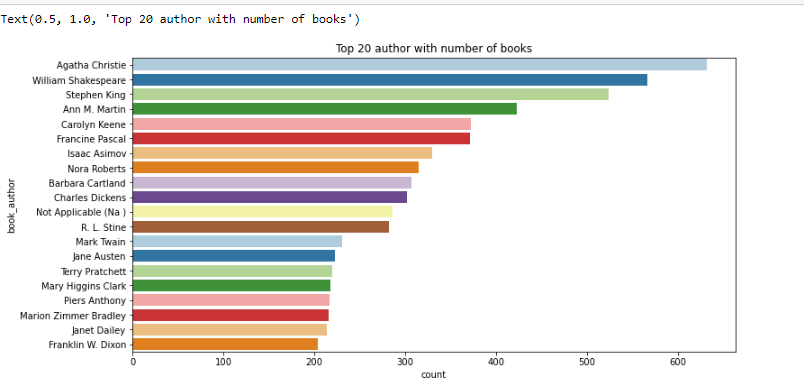
* **DATA DEDUPLICATION:**

It is very likely that your dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset.

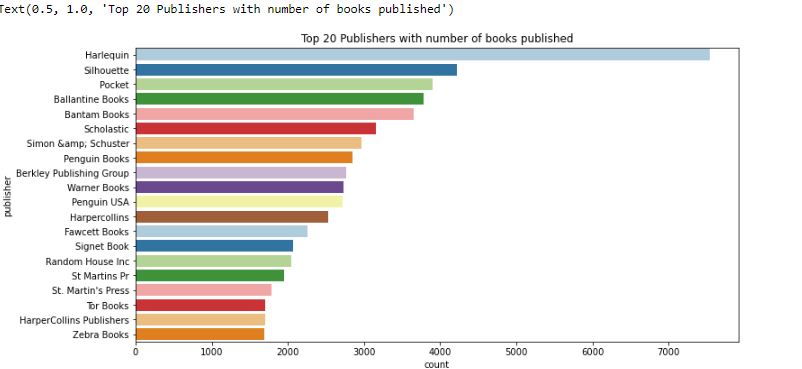
So there is no duplicate value present

* **Data Visualization:-**

**Below graph shows the Top 20 Authors with number of books**



* **Among top 20 Authors the highest number of books has been hold by Agatha Christie. Agatha Christie is leading at top with more than 600 counts, followed by William Shakespear**
* **Top 20 Publishers with number of books published**

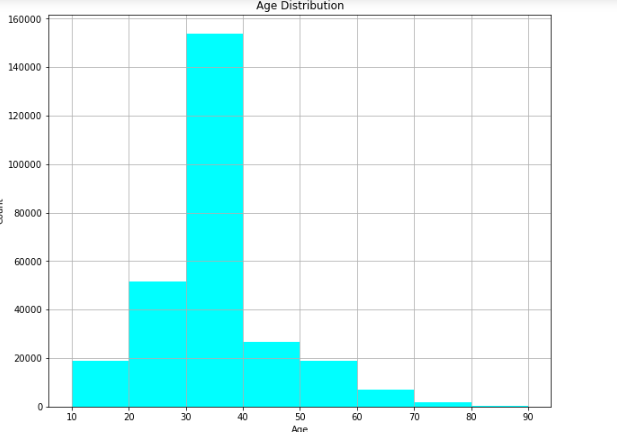


**Among top 20 Books Publishers the top 20 number of book publisher is Harlequin**

* **GRAPHICAL REPRESENTATION OF THE RESULTS:**

This is an essential step as the result analyzed from the dataset should be interpretable by the business stakeholders, which is one of the major goals of EDA.

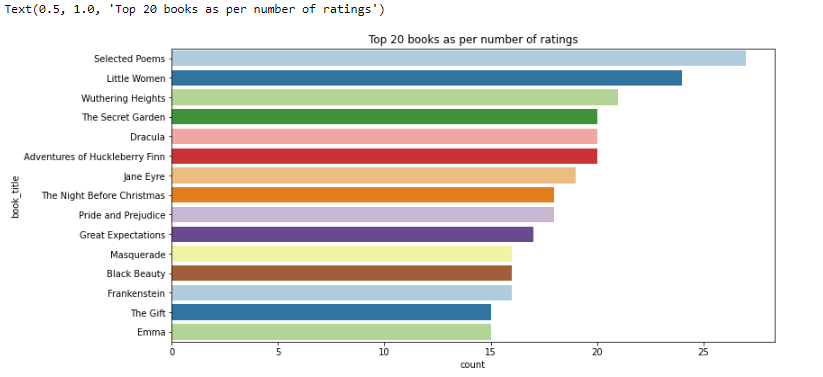
**Age Distribution:-**

****

**Most User are from the age of 10 to 70 years of age and max users age is in 30 to 40**

**Ratings:-**

**Top 20 books as per number of ratings**

****

Top 20 book which having highest number of ratings are given to Selected poems followed by Little women.

Popular Based Filtering:-

As the name suggests Popularity based recommendation system works with the trend. It basically uses the items which are in trend right now. This is particularly useful when you don't have past data as a reference to recommend product to the user. It is not tailor fit for any particular group of audience or movie.

Collaborative Filtering Based Recommender System:-

Collaborative Filtering is the most famous application suggestion engine and is based on calculated guesses; the people who liked the product will enjoy the same product in the future.

Collaborative Filtering (CF) is the most popular and widely used approach for Recommender System.

which tries to analyze the user's interest over the target item on the basis of views expressed by other like-minded users.

**CONCLUSIONS:**

Agatha Christie has 600 books which is the top one author who has a large number of book.rather than William Shakespeare is a great author but it is followed by agatha.

Harlequin has the most number of books published, followed by Silhouette.

Number of Books published yearly are between 1950 - 2005.

Most of the users are in 20-30 and 30-40 prefer more books.

As per ratings "Selected Poems" has been rated most followed by "Little Women". The countplot shows users have rated 0 the most, which means they haven't rated books at all.

The top 10 books recommendation as per ratings with top "The lovely Bones: A novel" with 707 book ratings. But this are not based on some recommendation system. They are top 10 books as per ratings.