12-2020

Recommender Systems and Amazon Marketing Bias

Yousuf Alolama

Recommender Systems and Amazon Marketing Bias

by

Yousuf Alolama

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology
RIT Dubai

December 2020

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name**: Yousuf Alolama**

Graduate Capstone Title**: Recommender Systems and Amazon Marketing Bias**

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of committee

Name: Dr. Ehsan Warriach **Date:**

Member of committee

Acknowledgments

First and foremost, I would like to thank Allah for providing me with strength, knowledge, prowess, and opportunity to undertake this paper and to persevere and complete it adequately as without such blessings, this milestone would not have been accomplished. I take prodigious pleasure in acknowledging my mentor Dr. Ehsan Warriach and all my other professors within this great master program for recognizing my determination towards life and appreciating my efforts and my hard work. I would like to thank my mentor for his support, feed-back and passionate encouragement which helped to complete my work successfully, as well as thank chair of committee Dr. Sanjay Modak for providing me with enough time to complete my capstone project, and be there for I and other students constantly without resting. Thank you again for your hard work.

Abstract

Proper marketing requires a great understanding of customers' needs and how they can be presented to them properly so that they purchase the services provided by companies and generate high profits. Recommender systems are considered as an answer to this marketing scheme; through feeding it enough data on customers' purchase history, it can understand what customers usually procure and what items can be proposed to them so that more sales can be conducted, resulting in higher profits. However, with the increasing number of techniques and algorithms for recommender systems, it reduced the rate of its accuracy toward presenting customers with the right items. In this study, Amazon's marketing bias and its recommender system will be researched and apply the many data analytics and mining techniques to further enhance their accuracy and the rate of a customer purchasing another item based on the recommender system. Moreover, the study will determine which recommender system paradigms is most suited for Amazon.com, either Collaborative or Content-based methods.

Keywords: Recommender Systems, Collaborative-based method, Content-based method, Confirmatory Factor Analysis; Preference Elicitation.

Table of Contents

ACKNOWLEDGMENTS......II

ABSTRACT......III

LIST OF FIGURES V

LIST OF TABLES......V

CHAPTER 1 1

1.1 BACKGROUND......1

1.2 STATEMENT OF PROBLEM 2

1.3 PROJECT GOALS 3

1.4 METHODOLOGY 3

1.5 LIMITATIONS OF THE STUDY......6

CHAPTER 2 – LITERATURE REVIEW......7

CHAPTER 3- PROJECT DESCRIPTION......10

3.1 DATA SOURCES 10

3.2 PROJECT RESOURCES 11

CHAPTER 4- PROJECT ANALYSIS 13

4.1 DATA PROCESSING 13

4.2 DATA MODELING......22

4.3 RESULTS 25

CHAPTER 5 CONCLUSION 27

5.1 CONCLUSION 27

5.2 FUTURE WORK......27

BIBLIOGRAPHY......28

List of Figures

Figure 1: Improving recommender system using data-driven research (Verbert, et al., 2011)

Figure 2: Illustration of the user-item interactions (Rocca, 2019)

Figure 3: Overview of the content-based methods paradigm (Rocca, 2019)

Figure 4: importing the necessary tools from Python library

Figure 5: sample of the dataset with the 3 headings selected earlier

Figure 6: checking for missing values

Figure 7: checking for duplicate values

Figure 8: removal of duplicate values

Figure 9: visualizing the rating attribute

Figure 10: understand the rating of products per user

Figure 11: understand the reason of giving the rate to the product purchased

Figure 12: number of rates per product (book)

Figure 13: code for genre effect on book ratings

Figure 14: diagram for genre effect on book ratings

Figure 15: diagram factors that affect a book rating's

Figure 16: importing the necessary tools to initiate the testing of the chosen model

Figure 17: splitting the dataset into 2 sets

Figure 18: running the trainset

Figure 19: running the trained model against the test set

Figure 20: running the model for the recommended products

Figure 21: recommended products

Figure 22: RMSE equation

Figure 23: checking the accuracy of the model produced using the above code

Figure 24: Amazon recommender system using the collaborative-based method

Figure 25: Amazon recommender system using the user-user approach

List of Tables

Table 1: basic statistics of the dataset

Chapter 1

1.1 Background

Companies advertise their products to customers through different channels, and once they have the customers take an interest in their product, they start marketing their other products. Nonetheless, it is not as simple as it sounds, as they need to understand what made the customer purchase the first product, and what might qualify to be their next shopping item, which can become a taxing job, as each customer preference differs from one another. Initially, the companies started testing by placing some products commonly bought by a certain age group next to items that offer similar attributes to the frequently bought item; however, from a different brand and shape. Step by step, they learned what each age category preferred buying and disliked purchasing other products, allowing the companies to assort the products in a way that would suit each age category and allow for more sales to take place. Hence came the Recommender system that became a revolutionary asset to multiple Billionaire companies such as Netflix, Amazon, Ali Baba, and many more. In this study, Amazon.com will be the main subject, specifically their Marketing bias. The account of the arrangement of Amazon.com is regularly rehashed and is presently an urban legend. Jeff Bezos established the organization, software engineering, and electrical designing alumni from Princeton University. Bezos had moved to Seattle in the wake of leaving as the senior VP at D.E. Shaw, a Wall Street speculation bank. He did not think a lot about the Internet. Be that as it may, Bezos ran over a measurement that the Internet was 5 developing at 2300%, which persuaded him that it was a huge development opportunity. Not knowing significantly more, he dove into the universe of E-Commerce with no past retailing experience (Krishnamurthy, 2002). Amazon utilizes the recommender system within its marketing schemes (Hardesty, 2019), ensuring millions of customers are buying more than one product, by showing them frequently bought items with their selected products with strong reviews recommending buying those mentioned items (Martinez, n.d.).

AI and Deep Learning have an immense scope in recommender systems however, the risk of turning customers away due to wrong recommendation is high as it deals with people different personality and shopping habits and not just a classification problem. However, if applied cautiously, it can benefit the economy in tremendous manners.

1.2 Statement of problem

What do I want to solve here? Companies advertise their products to customer through different channels, to increase their profits, however, this task is not as easy as it sounds as it needs to understand the different kinds of algorithmic behaviors of each customer, and how it all can be linked to the products those companies try to sell and have a return on their investment by showcasing the right product to the right customer. Hence, the recommender system; "a number of algorithms aimed at signifying items relevant to users' preferences such as movies, texts, shopping items" (Rocca, 2019). This study will be focusing on Amazon, a worldwide organization renowned for its wide selective range of products that are delivered across the world, and virtually host numerous products, however, throughout this study, the marketing bias of Amazon will be discussed and researched to answer the following questions that question their marketing methods and how effective it is in attracting customers. The Seattle, the WA based company, has grown from a bookseller to a virtual Wal-Mart of the Web selling products as diverse as Music CDs, Cookware, Toys and Games, and Tools and Hardware. The company has also grown at a tremendous rate, with revenues rising from about $150 million in 1997 to $3.1 billion in 2001 (Krishnamurthy, 2002). The questions are as follow:

- How accurate is Amazon's marketing bias toward its customers?
- What are the factors that contribute the most toward marketing Amazon products?
- How can Amazon further enhance its Marketing tool by implementing the different methods of data analytics within its Marketing bias?

1.3 Project goals

Throughout the study conducted on Amazon Marketing bias and utilizing recommender systems from data analytics, the following are the desired deliverables from this study:

- An enhancement on Amazon marketing bias.
- Increase product sale at Amazon.com.
- Augment people awareness toward items that are similar and beneficial when bought next to their already selected item.
- Establish how accurate those recommender systems are when used in Amazon Marketing bias.
- A word document report on the findings and research done throughout the study.

While these deliverables are beneficial, they are not easy to achieve, given the limitation of the data at hand; however, the best practice will be applied within this study to achieve the maximum possible results to meet the expected outcome.

1.4 Methodology

This paper aims to study the marketing bias and recommender system usage in Amazon and how accurate it is and which recommender system paradigms are most valuable for the said company and how it can be enhanced through applying what has been taught in the data analytics master degree program. However, it is pertinent to understand that without data there can be no foundation to base our study upon, so the first approach is to research for Amazon available datasets that can be used for this study, as the Gantt chart shows in the below section titled "Project timeline" duration of 40 days was assigned as this task requires a lot of time and focus when researching for the right dataset to work with. Next, the methods to be implemented and tested on the dataset by segmenting the dataset, profile it, apply some data mining techniques to solve the questions imposed above. The student then will be checking which recommender system paradigm is being used from the below figures and ascertain which approach is used and how it can be enhanced through the different data mining techniques previously learned.

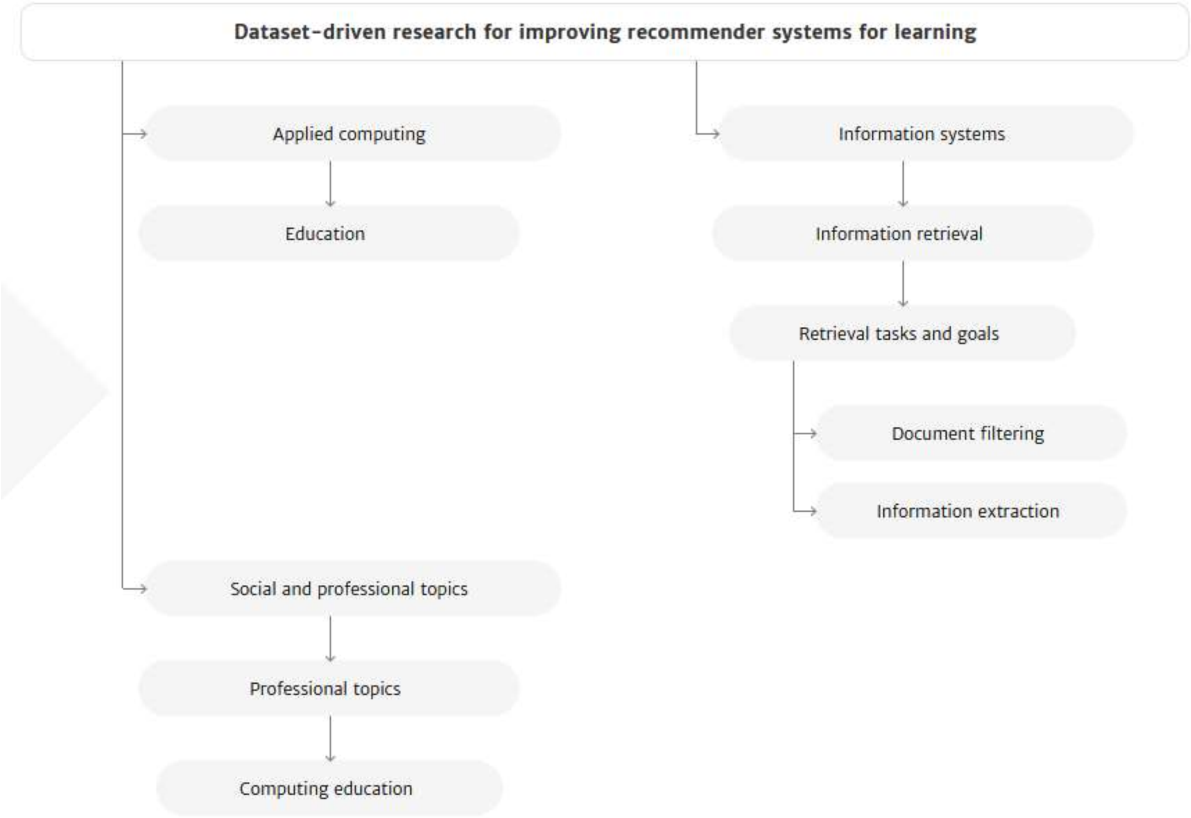


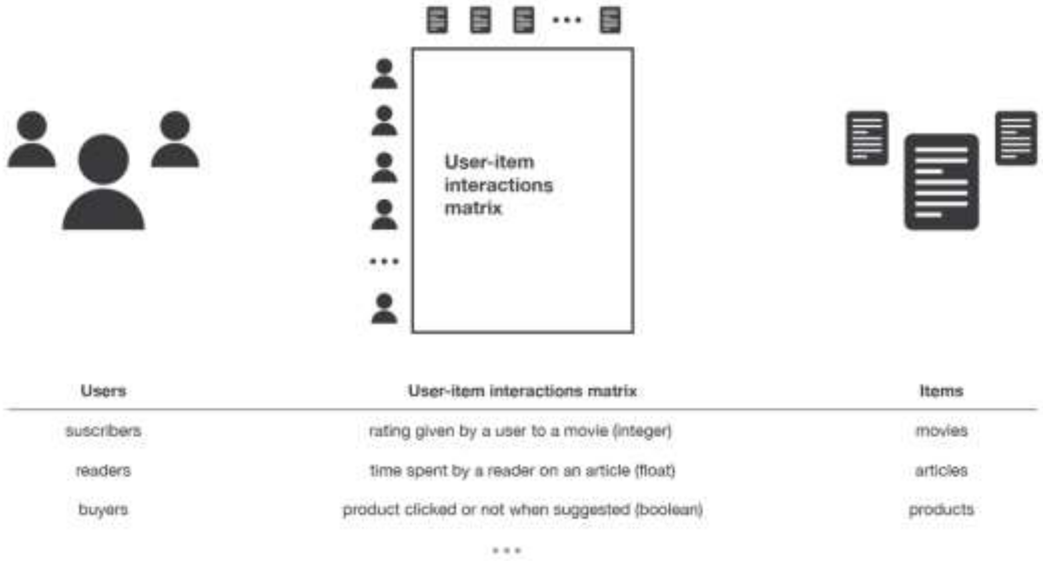
Figure 1: Improving recommender system using data-driven research (Verbert, et al., 2011)

The tools that can be utilized for the study are the following with each entailing what they will be used for:

- R studio: will handle the dataset and will host the different data analytics and data mining techniques to visualize and answer the imposed questions.
- Python: will handle the dataset in case r does not comprehend the amount of data.
- Google search engine: research the many articles, codes, reading materials, datasets, etc.

- Lecture slides from previous courses: utilize the codes and techniques taught, the writing methods, and how to describe charts and plots.

To be able to distinguish which model is being followed in the recommender system, the following figure will allow the student to comprehend which methodology is used and how to tackle it:



Users	User-item interactions matrix	Items
suscribers	rating given by a user to a movie (integer)	movies
readers	time spent by a reader on an article (float)	articles
buyers	product clicked or not when suggested (boolean)	products
	...	

Figure 2: Illustration of the user-item interactions (Rocca, 2019)

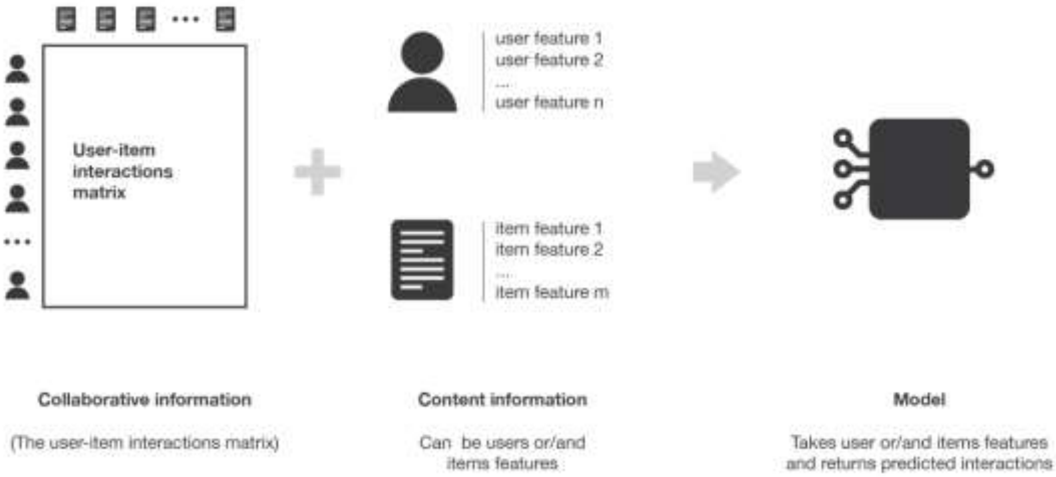

User-item
interactions
matrix
user feature 1
user feature 2
...
user feature n
item feature 1
item feature 2
...
item feature m
Collaborative information
(The user-item interactions matrix)
Content information
Can be users or/and items features
Model
Takes user or/and items features and returns predicted interactions

Figure 3: Overview of the content-based methods paradigm (Rocca, 2019)

1.5 Limitations of the Study

There are not many datasets available that are published on the known search engines as Amazon is very conservative of its data, and only allow certain information to be known to public. An attempt to get data from Amazon through emailing their customer service to check whether they will be willing to share their data for master thesis and further escalate it to University staff to get permission was turned down. Nevertheless, the dataset available for this study turned to be sufficient.

Chapter 2 – Literature Review

Throughout the world, there are companies that require systems that would enhance customers aware of their variety of products they offer, however without a good system implemented in those companies' sale sites, it is very difficult to achieve maximum profits possible. Hence, the recommender systems come to light and guide those companies' customers toward the right product and several other products accompanying it, further enhancing the company's profit rate and allow for products to be sold, rather than sitting in warehouses and costing the company more to ensure they are still in good condition. Zaier, Godin, and Faucher feel that “Recommender systems are considered as an answer to the information overload in a Web environment. Such systems recommend items (movies, music, books, news, web pages, etc.) that the user should be interested in” (Zaier, Godin, & Faucher, 2008). It is, as they say, the recommender systems are indeed a solution used by multiple globally renowned organizations in their marketing schemes to maximize profits, and people tend to utilize said recommender systems to buy, watch, and visit places, therefore increasing the profits gained by those organizations. The recommender system is extremely valuable to today's applications, and it is widely utilized in people's daily life routines without them realizing how much of a help it is (Zaier, Godin, & Faucher, 2008).

In addition, Authors of the book “Evaluating Recommender Systems with User Experiments” believe that user experience is needed when experimenting with algorithms for recommender systems to produce a best-fit algorithm for a group of people based on certain criteria established ahead of the experiment, and further add “Proper evaluation of the user experience of recommender systems requires conducting user experiments.” (Knijnenburg & Willemsen, 2015). Recommender systems have two major paradigms which are Collaborative and Content-based, and they are described as follow:

- Collaborative methods: are strategies that depend exclusively on the past connections recorded among clients and things to deliver new proposals. The author then states that “the main idea that rules collaborative methods is that these past user-item interactions are sufficient to detect similar users and/or similar items and make predictions based on these estimated proximities." (Rocca, 2019). It is further subdivided into memory-based and model-based approaches. The author further states that "They represent a powerful method for enabling users to filter through large information and product spaces." (Ekstrand, Riedl, & K, 2011).

- Content-based methods: this method not only uses the collaborative method techniques but adds the use of additional information provided by the users and the items. The author states that "If we consider the example of a movies recommender system, this additional information can be, for example, the age, the sex, the job or any other personal information for users as well as the category, the main actors, the duration or other characteristics for the movies (items)." (Rocca, 2019). Furthermore, the author states that "Most content-based recommender systems use textual features to represent items and user profiles, hence they suffer from the classical problems of natural language ambiguity." (Gemmis, Lops, Musto, Narducci, & Semeraro, 2015). Moreover, content-based methods do not have to deal with the difficult start-up needed for collaborative methods, which gives it an advantage over the collaborative methods. A content-based filtering system chooses things dependent on the connection between' s the substance of the elements and the client's inclinations rather than a collaborative filtering framework that picks things dependent on the relationship between' s kin with similar affinities (Meteren & Someren, 2000).

Market basket analysis is commonly used by large shopping centres to understand the relation between the different items they sell and use those discovered relations to increase their sales. The process occurs when multiple items are bought together in a single transaction in a frequent manner, thus allowing retailers to understand how to market their products and increase sale. It is usually mistaken with the recommender systems. Market-Basket Analysis is a technique in which buyers' preferences are studied in order to find the relationship between various goods in their market basket. By considering the products often bought together by buyers, the exploration of these relationships will help the merchant establish a sales strategy (D. H. Setiabudi, G. S. Budhi, I. W. J. Purnama and A. Noertjahyana). It then creates if and then scenario rules at the transaction stage, which lists all products purchased by the customer in a single purchase. The technique establishes the relationship between which products have been purchased and which other product has been purchased(s).

Recommender systems have to methods that are widely known and used around the world; Content-based and Collaborative. In modern age, and a narrower sense, collaborative filtering is a method of making automatic predictions about a user's interests by collecting preferences or taste information from many users. The underlying assumption of the collaborative filtering approach is that if person A has the same opinion as person B on an issue, A is more likely to have B's opinion on a different issue than a randomly chosen person.-based method, throughout this study we will be discussing the latter methodology in detail.

Chapter 3- Project Description

3.1 Data sources

The dataset that will be used for the Capstone project will consist of the following:

- The total number of samples present in the dataset are 1,050,000.
- The total number of features present in the dataset are 10.

The dataset contains multiple variables from products that are sold at both Amazon and Modcloth, which can showcase the marketing bias done by Amazon. The dataset source come from (McAuley, Lakkaraju, & Leskovec, 2013) at which there were multiple dataset for recommender systems, however the selected dataset was the most relevant to the topic of this paper. It will be retrieved from the following source:

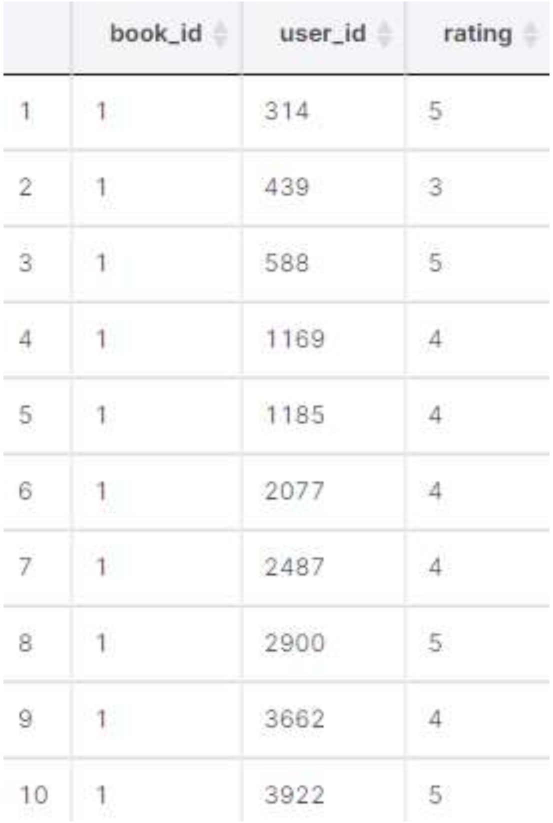
https://cseweb.ucsd.edu/~jmcauley/datasets.html#market_bias

another focus will be on the Collaborative-based method, for a specific set of products throughout the dataset originally obtained from above mentioned source. I will be discussing the usage of the specified product, which is Books, as it is a subject of interest of mine and many of my fellow book readers and purchasers.

Some of the basic statistics of the dataset is shown in the table below:

	ModCloth	**Amazon electronics**
Reviews	99,893	1,292,954
Items	1,020	9,560
Users	44,783	1,157,633

The below table showcases a sample of the product specified earlier on, and the rating for said book by other users with the range of rating going from 1 to 5:



	book_id	user_id	rating
1	1	314	5
2	1	439	3
3	1	588	5
4	1	1169	4
5	1	1185	4
6	1	2077	4
7	1	2487	4
8	1	2900	5
9	1	3662	4
10	1	3922	5

As for the metadata that are being used within the dataset are the following:

- Ratings
- Product images
- User identities
- Item sizes
- User genders
- Book ID

3.2 Project Resources

In order to start processing the dataset at hand, it is essential that the following be established to have a perfectly running software and programs, as running the program for the dataset will require a set amount of memory and graphic processing units that require certain physical assets to be there when the work starts:

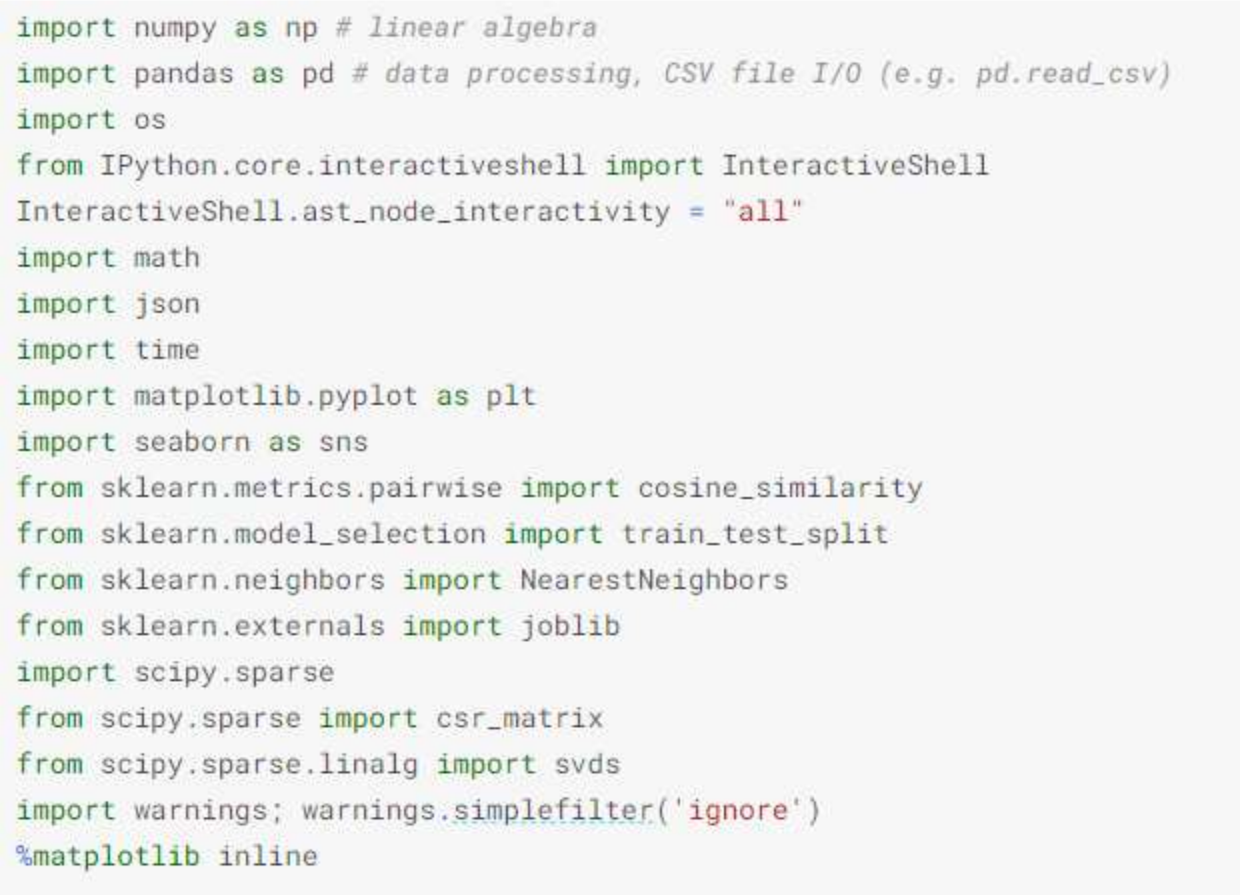
- Hardware Requirements
    - Personal laptop/Desktop PC
    - Intel i7 or higher
    - 16GB Ram or higher
    - NVIDIA GEFORCE RTX
- Software Requirements
    - R studio (tools as well)
    - Python (latest version)
    - Jupyter Notebook
    - Note pad ++
    - Snipping Tool

Chapter 4- Project Analysis

4.1 Data Processing

The first step in data processing is to check whether the dataset at hand have the missing values, outliers, unknown variables, etc. and try to eliminate them from the dataset therefore not affecting the exploratory data analysis phase and the modeling phase. The following steps are taken to prepare the dataset:

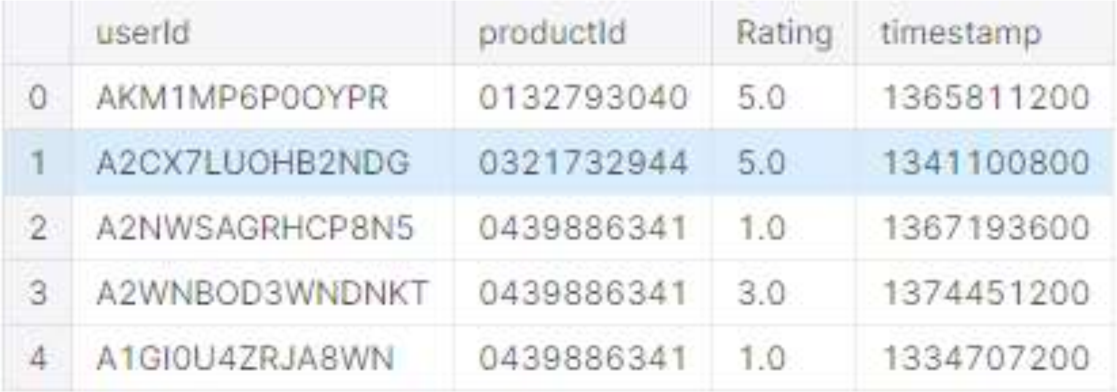
- Load the library and tools: for this step, the following was input so that I could advance in the processing step of our study, as the I required the following libraries imported from Python to start working on the data available within the set meant to be used for this specific study.



```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import os
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
import math
import json
import time
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.model_selection import train_test_split
from sklearn.neighbors import NearestNeighbors
from sklearn.externals import joblib
import scipy.sparse
from scipy.sparse import csr_matrix
from scipy.sparse.linalg import svds
import warnings; warnings.simplefilter('ignore')
%matplotlib inline
```

Figure 4: importing the necessary tools from Python library

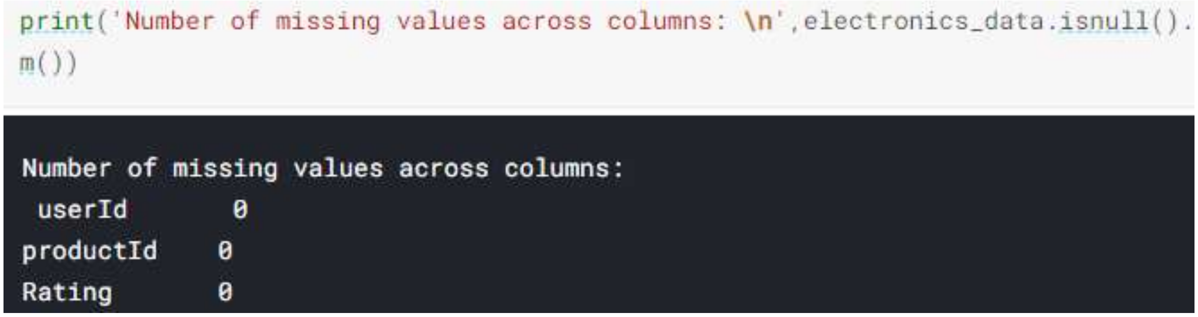
- Data reduction:
    - Through this step, the dataset variables are reduced and mainly focus on the following attributes as I try to aim my focus on this study to produce some sense from the data obtained:
        - userid: Each user is marked with a unique id.
        - rating: rating of products based on users.
        - Book ID: the id for books purchased by users and rated.



	userId	productId	Rating	timestamp
0	AKM1MP6P0OYPR	0132793040	5.0	1365811200
1	A2CX7LUOHB2NDG	0321732944	5.0	1341100800
2	A2NWSAGRHCP8N5	0439886341	1.0	1367193600
3	A2WNBOD3WNDNKT	0439886341	3.0	1374451200
4	A1GI0U4ZRJA8WN	0439886341	1.0	1334707200

Figure 5: sample of the dataset with the 3 headings selected earlier

- Missing values:
    - The following input showcases the missing values if any, however, when checked whether this dataset had any missing value, it turned out to have none, which was a reassuring factor that both ensured that the dataset values are good to go and no need to filter out the missing data.

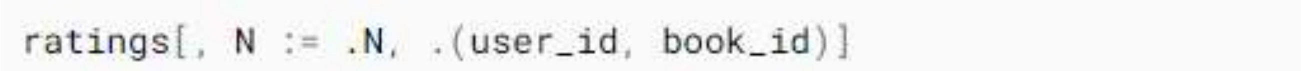


```
print('Number of missing values across columns: \n',electronics_data.isnull().
m())
```

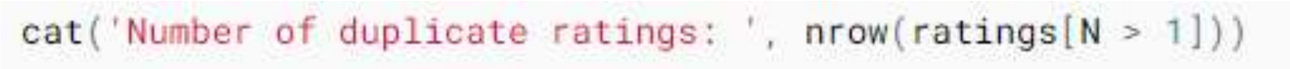
```
Number of missing values across columns:
 userId       0
productId    0
Rating       0
```

Figure 6: checking for missing values

- Removing the duplicate data: in this step I will be removing the duplicate data from the rating column within the dataset, so that when I sample the data, it produces accurate results that make sense and can be modified into graphs and charts.



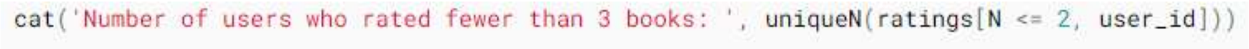
```
ratings[, N := .N, .(user_id, book_id)]
```



```
cat('Number of duplicate ratings: ', nrow(ratings[N > 1]))
```

Figure 7: checking for duplicate values

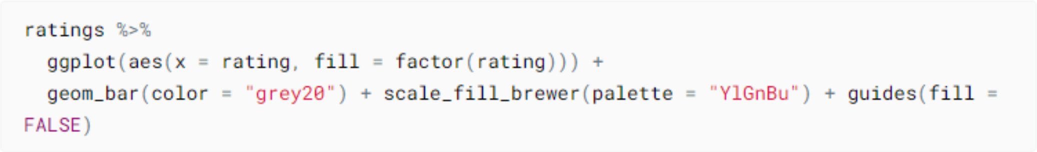
Turns out there are about 4487 duplicate rating that exists within the dataset which will be dropped out of the dataset to reduce the error value in the upcoming steps.



```
cat('Number of users who rated fewer than 3 books: ', uniqueN(ratings[N <= 2, user_id]))
```

Figure 8: removal of duplicate values

- Visualize data:
  - Examining the distribution of the rating attribute through the following code input, which indicate that usually users provide a good rating to the purchased products, while it is rare for users to rate products negatively as shown in the bar chart below:



```
ratings %>%
  ggplot(aes(x = rating, fill = factor(rating))) +
  geom_bar(color = "grey20") + scale_fill_brewer(palette = "YlGnBu") + guides(fill =
FALSE)
```

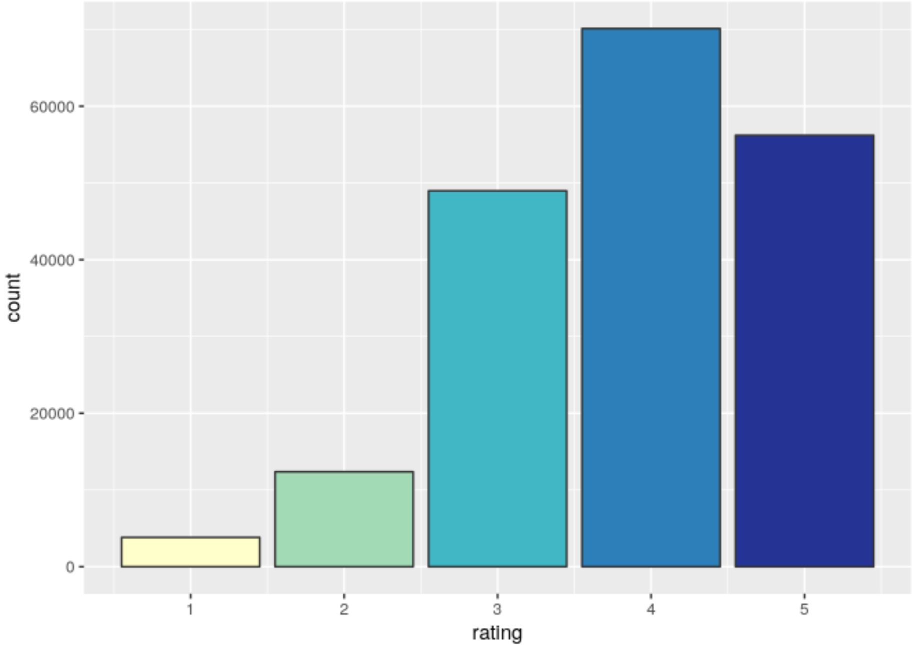
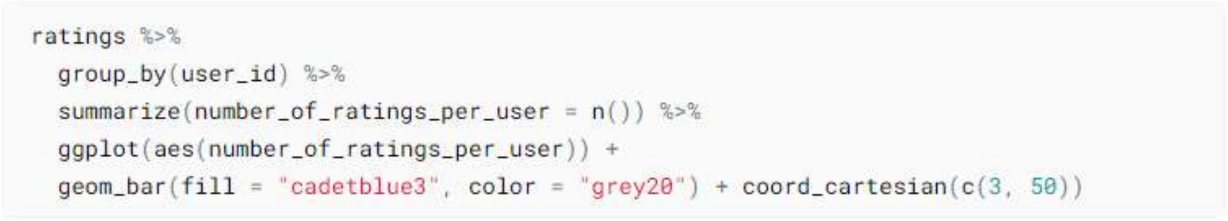


Figure 9: visualizing the rating attribute

- Inspecting the number of rating provided by the user through the following code input, which indicates the number of ratings done per user keeping in mind that duplicate ratings have been removed earlier on, hence the diagram below should show us an accurate graph of how many rates are done through users:
    - For example
        - User a: 13 ratings
        - User b: 21 ratings, etc.



```
ratings %>%
  group_by(user_id) %>%
  summarize(number_of_ratings_per_user = n()) %>%
  ggplot(aes(number_of_ratings_per_user)) +
  geom_bar(fill = "cadetblue3", color = "grey20") + coord_cartesian(c(3, 50))
```

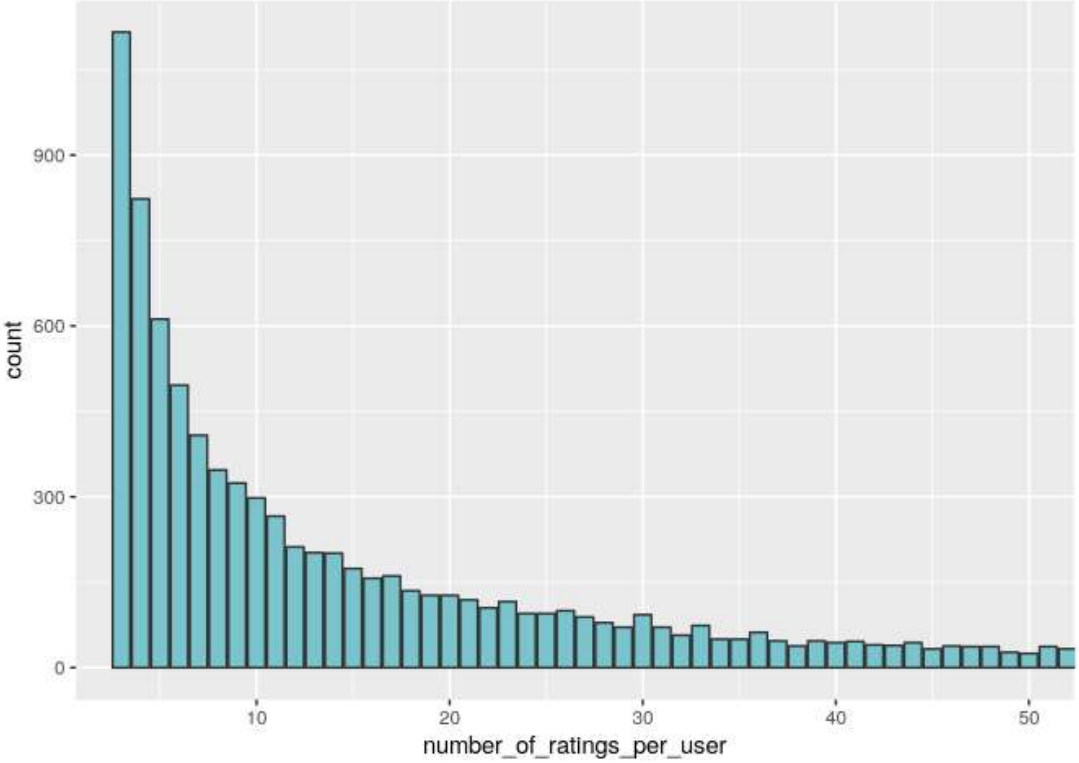


Figure 10: understand the rating of products per user

However, there are a number of users who rate their products based on the quality of the written material, whereas others rate based on shipping service, item packaging, how fun the book was for them when they have read, etc. this all can be witnessed in the next diagram, which will contribute to the methodology I will be enforcing later on the dataset to understand how it can benefit users who wish to purchase products that peak to their interests accurately:

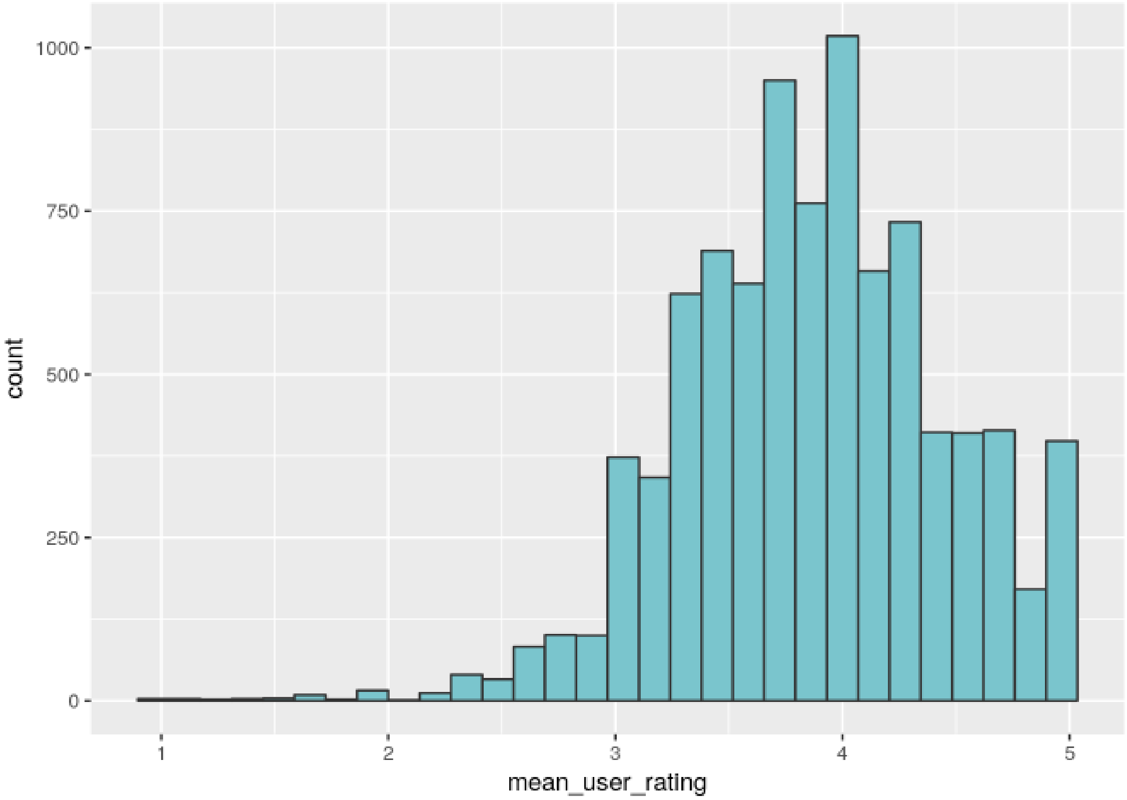
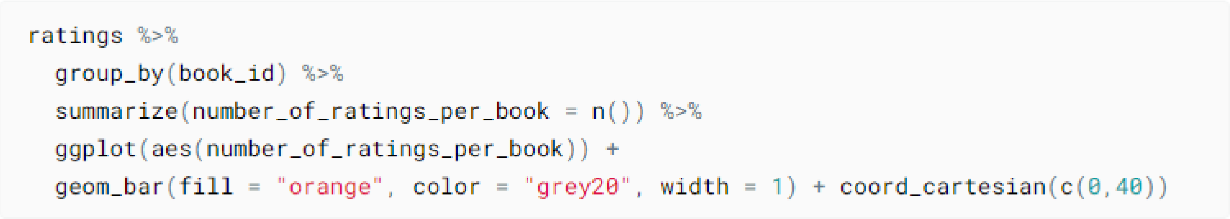


Figure 11: understand the reason of giving the rate to the product purchased

- Getting the number of total rating per book through the following code input, which indicates that an item rated by users usually tend to be between 18 to 23 rates, which is entails that the example below is correct:
  - As the number of books purchased increases the number rates increases as well albeit in a small manner.
  - The books purchased usually have around 18 to 23 ratings



```
ratings %>%
  group_by(book_id) %>%
  summarize(number_of_ratings_per_book = n()) %>%
  ggplot(aes(number_of_ratings_per_book)) +
  geom_bar(fill = "orange", color = "grey20", width = 1) + coord_cartesian(c(0,40))
```

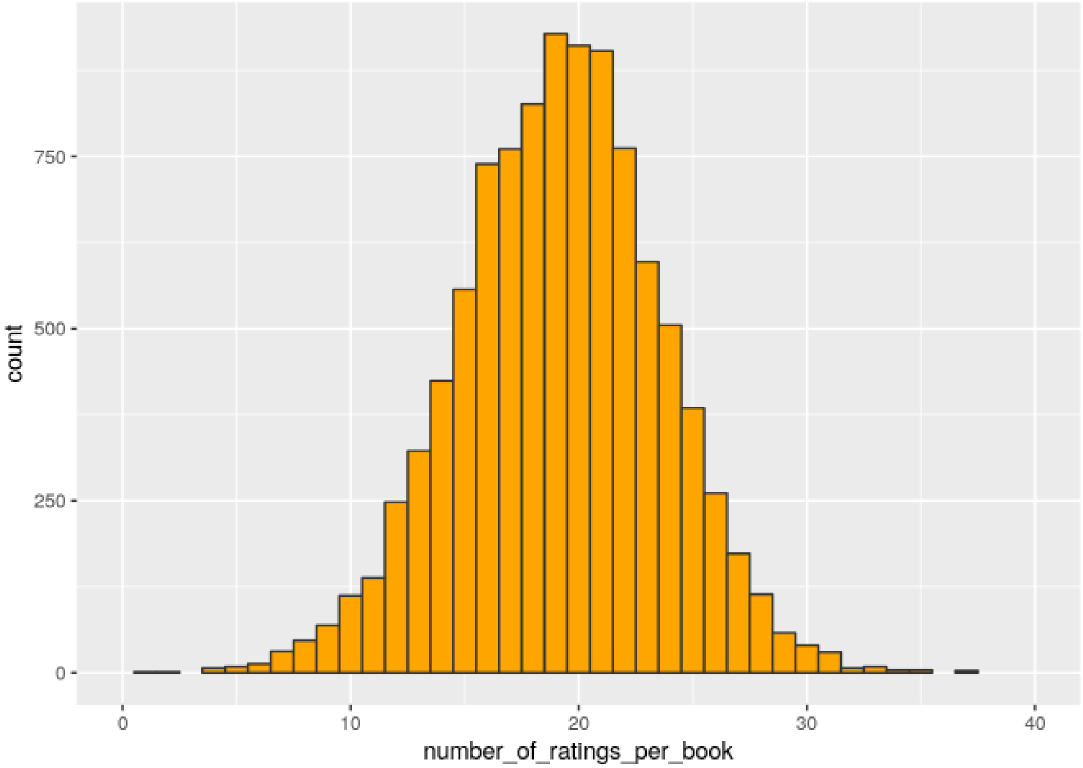
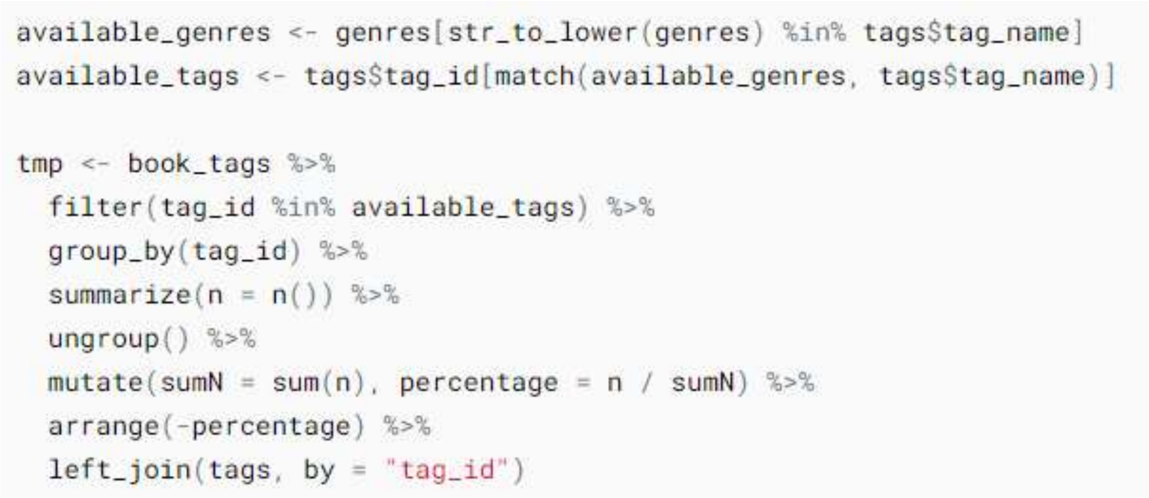


Figure 12: number of rates per product (book)

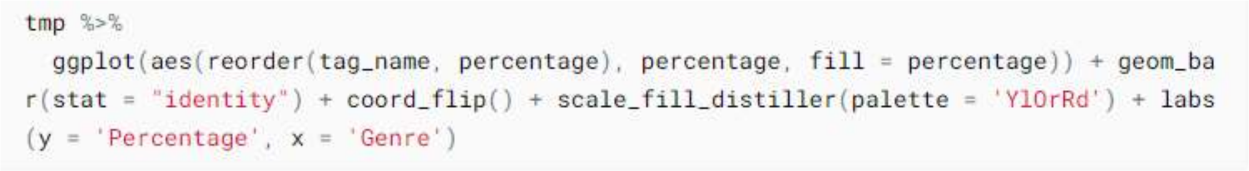
- Inspecting whether genre affects the rating of books that are rated by users: the following step will check if the data within the set is affected by the book genre, as the diagram below show that it only accepts the tags that are official and not the personal tags used to rate books based on genre as this would both require time and effort that I am not willing to spend within this study due to those two being extremely limited and cannot be bargained.



```
available_genres <- genres[str_to_lower(genres) %in% tags$tag_name]
available_tags <- tags$tag_id[match(available_genres, tags$tag_name)]

tmp <- book_tags %>%
  filter(tag_id %in% available_tags) %>%
  group_by(tag_id) %>%
  summarize(n = n()) %>%
  ungroup() %>%
  mutate(sumN = sum(n), percentage = n / sumN) %>%
  arrange(-percentage) %>%
  left_join(tags, by = "tag_id")
```

Figure 13: code for genre effect on book ratings



```
tmp %>%
  ggplot(aes(reorder(tag_name, percentage), percentage, fill = percentage)) + geom_ba
r(stat = "identity") + coord_flip() + scale_fill_distiller(palette = 'YlOrRd') + labs
(y = 'Percentage', x = 'Genre')
```

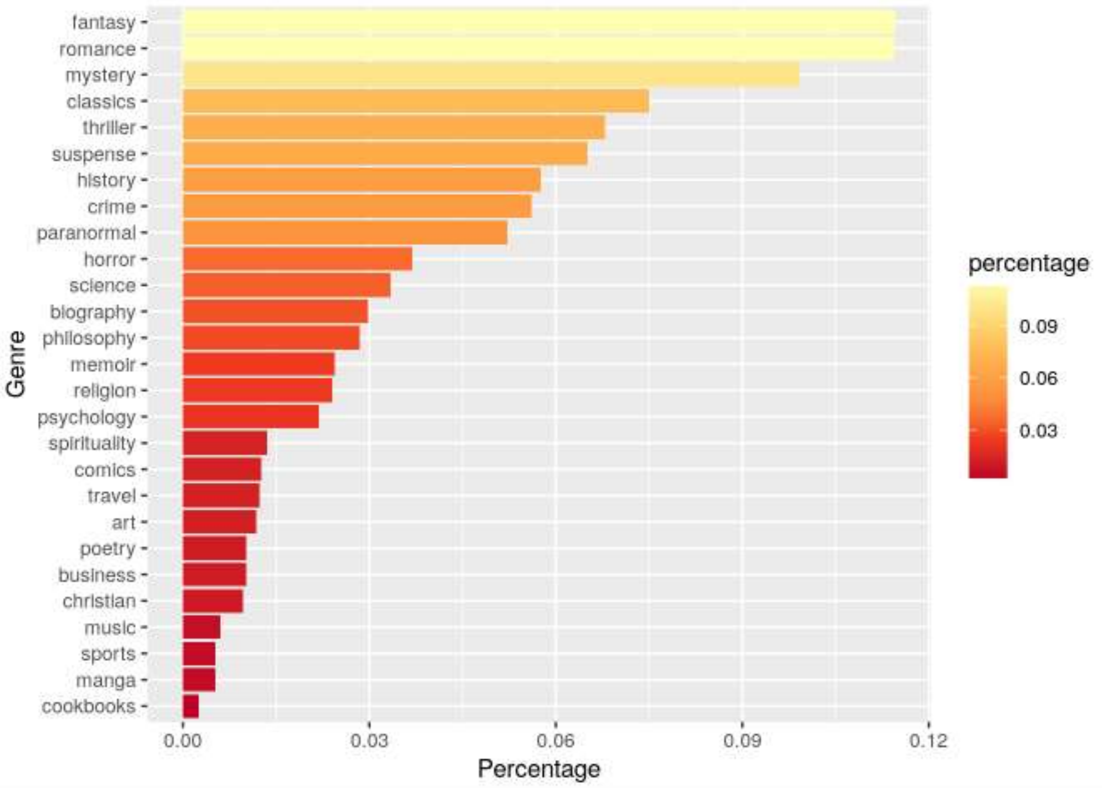
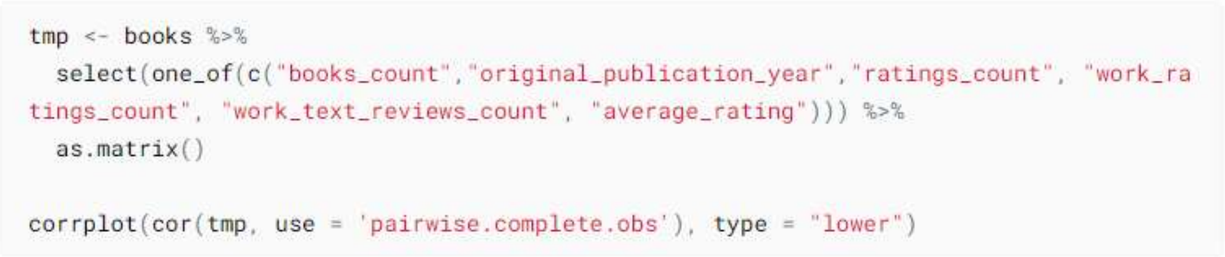


Figure 14: diagram for genre effect on book ratings

- Inspecting what are the factors that affect a book rating when purchased by a user and rated later on. The following code input and diagram show that there are some small correlations between the factors however they do not amount to much when deciding the reasons that affect a user rating behavior.



```
tmp <- books %>%
  select(one_of(c("books_count","original_publication_year","ratings_count", "work_ra
tings_count", "work_text_reviews_count", "average_rating"))) %>%
  as.matrix()

corrplot(cor(tmp, use = 'pairwise.complete.obs'), type = "lower")
```

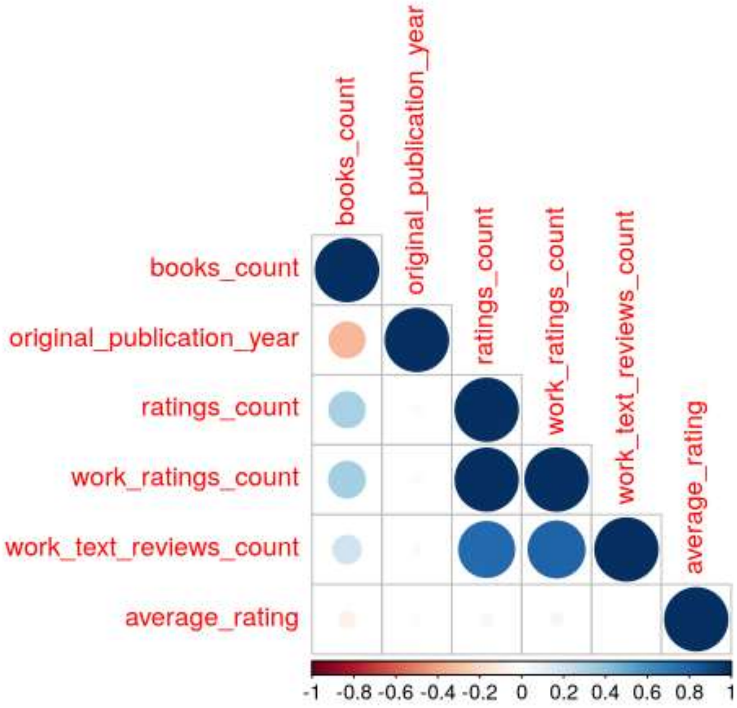


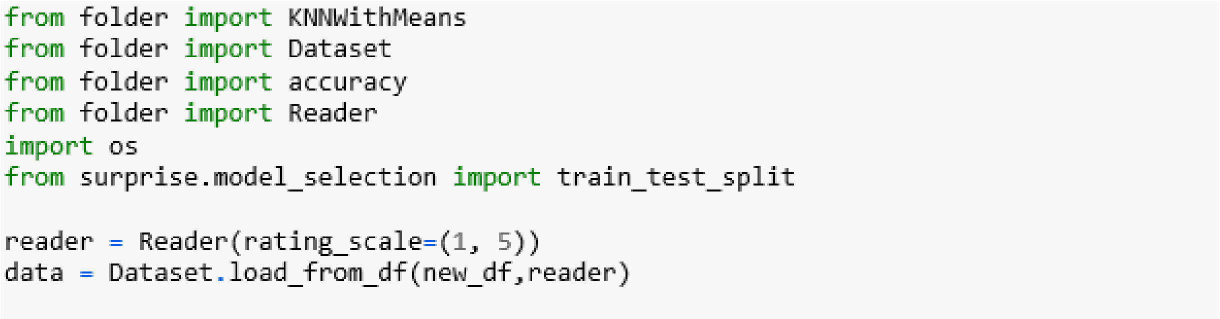
Figure 15: diagram factors that affect a book rating's

4.2 Data Modeling

In this step, the dataset is ready to be utilized in the model that have been selected earlier on "Collaborative-based model" and understand how its multiple techniques can enhance Amazon marketing bias and upsurge the profits. Collaborative filtering for recommendation systems is widely used, and through this study the usage of a collaborative filtering technique based on the premise that people with similar preferences have the similar tastes. Recommender systems have to methods that are widely known and used around the world; Content-based and Collaborative. In modern age, and a narrower sense, collaborative filtering is a method of making automatic predictions about a user's interests by collecting preferences or taste information from many users. The underlying assumption of the collaborative filtering approach is that if person A has the same opinion as person B on an issue, A is more likely to have B's opinion on a different issue than a randomly chosen person.-based method, throughout this study we will be discussing the latter methodology in detail.

In other words, in order to predict how someone will rate an item, it utilizes historical item ratings of like-minded people. The following steps are implemented to check the dataset with the established model:

- Import the needed tools and read the dataset through the following code:

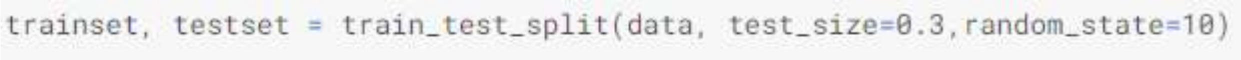


```
from folder import KNNWithMeans
from folder import Dataset
from folder import accuracy
from folder import Reader
import os
from surprise.model_selection import train_test_split

reader = Reader(rating_scale=(1, 5))
data = Dataset.load_from_df(new_df,reader)
```

Figure 16: importing the necessary tools to initiate the testing of the chosen model

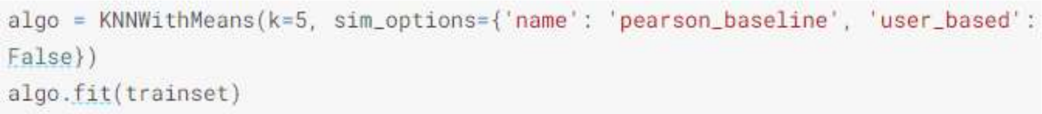
- Splitting the dataset to two sets; trainset and test set: through this step I will be recreating the data into two sets that will be used for model training and model testing. This will ensure that the trained model will be ready when tested against the testing set and hence produce accurate results for this study on how effective recommender systems for Amazon marketing bias are.



```
trainset, testset = train_test_split(data, test_size=0.3,random_state=10)
```

Figure 17: splitting the dataset into 2 sets

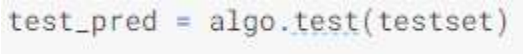
- Running the trainset: in this step, the dataset is will be running the training set which will allow the system to learn how to collaboratively work on recommending products (books for example) to users based on their preferences, hence allowing them to purchase more items and increase sale for Amazon.



```
algo = KNNWithMeans(k=5, sim_options={'name': 'pearson_baseline', 'user_based':
False})
algo.fit(trainset)
```

Figure 18: running the trainset

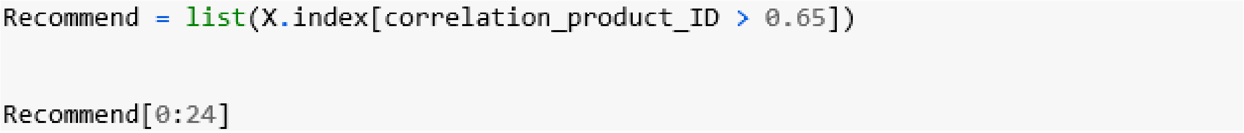
- Running the trained model against the test set: after getting the set ready with the training step, it is time to test it against the test set that I have prepared earlier on when splitting the dataset into two parts. This step will enlighten us toward how accurate the model built so far when running tests on it.



```
test_pred = algo.test(testset)
```

Figure 19: running the trained model against the test set

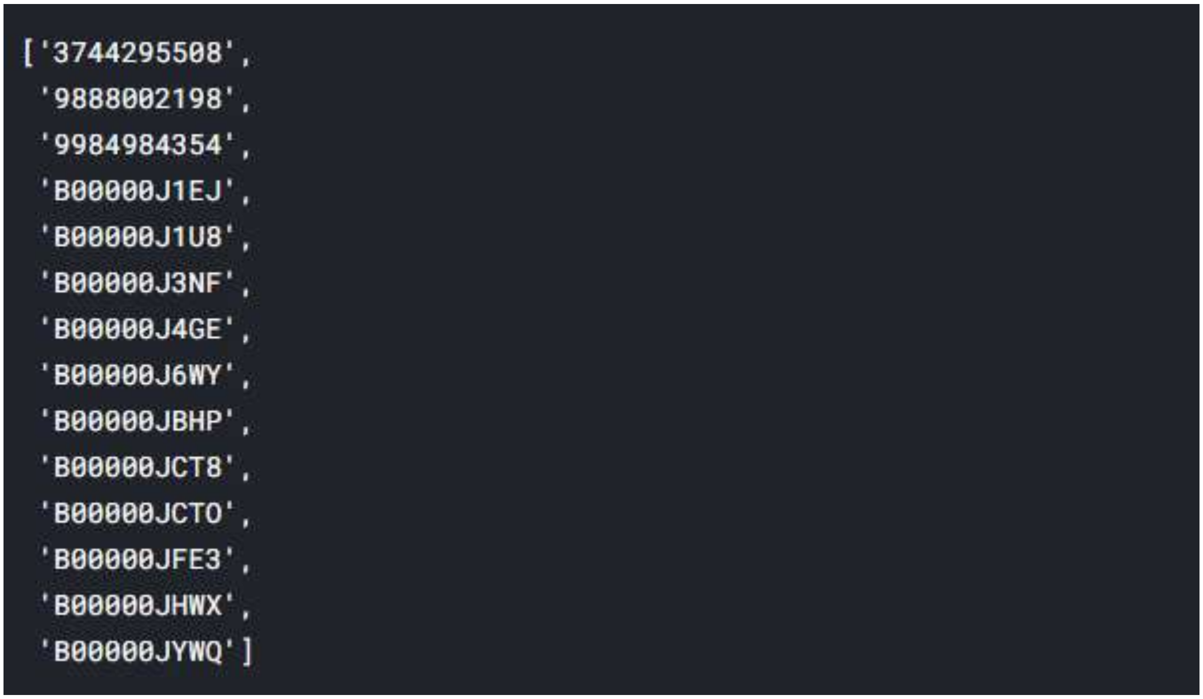
- Running the latest model and inspecting the top-rated items in sequence correlating to other users with similar preferences:



```
Recommend = list(X.index[correlation_product_ID > 0.65])


Recommend[0:24]
```

Figure 20: running the model for the recommended products



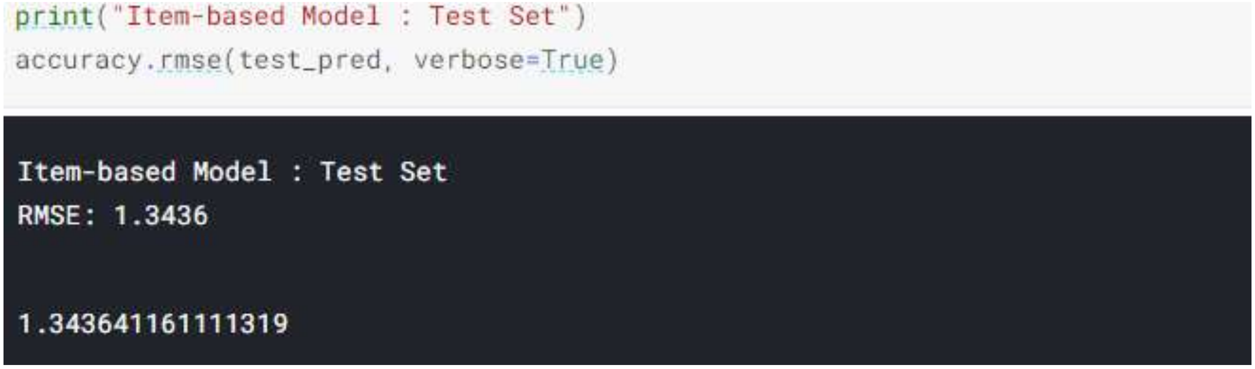
```
['3744295508',
 '9888002198',
 '9984984354',
 'B00000J1EJ',
 'B00000J1U8',
 'B00000J3NF',
 'B00000J4GE',
 'B00000J6WY',
 'B00000JBHP',
 'B00000JCT8',
 'B00000JCTO',
 'B00000JFE3',
 'B00000JHWX',
 'B00000JYWQ']
```

Figure 21: recommended products

- Running the accuracy code in order to understand how accurate the model turned out to be with the trainset and test set against each other: this step will require an evaluation tool to test how accurate the built model is, and I have selected the Root Mean Square Error tool which will determine how accurate the model is based on the value produced after running the RMSE code input. The following shows the equation used when applying the Root Mean Square Error within the code:

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N]^{1/2}$$

Figure 22: RMSE equation



```
print("Item-based Model : Test Set")
accuracy.rmse(test_pred, verbose=True)
```

```
Item-based Model : Test Set
RMSE: 1.3436


1.343641161111319
```

Figure 23: checking the accuracy of the model produced using the above code

4.3 Results

Finally, the aforementioned steps produced the desired results of the study, which is a measure of how accurate the recommender system will be when implanted in the Amazon marketing bias system which turned out to be of the following value:

- Root Mean Square Error also known as RMSE = 1.3436 which is very accurate.

In this study, I have used a dataset on product sale at Amazon and directed my focus on Book sale and checked whether implementing a recommender system using the Collaborative-based method would accurately recommend books to other users so that they would purchase it and increase sale for Amazon. Implementing recommender systems in today shopping and retail industry is a must, as data value keeps increasing, industries need to adhere the need to utilize said data to their full potential and provide their customers with items that best matches their preferences and increase their profits alongside their customers' happiness. The below figure showcases Amazon recommender system with high accuracy and how it can recommender to users based on their preferences:

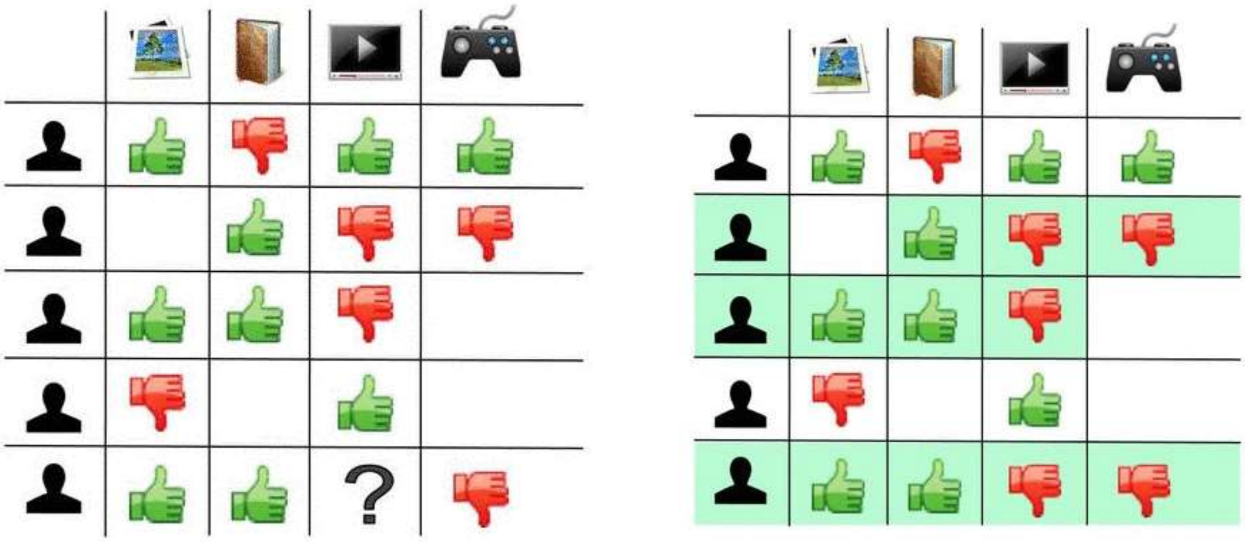
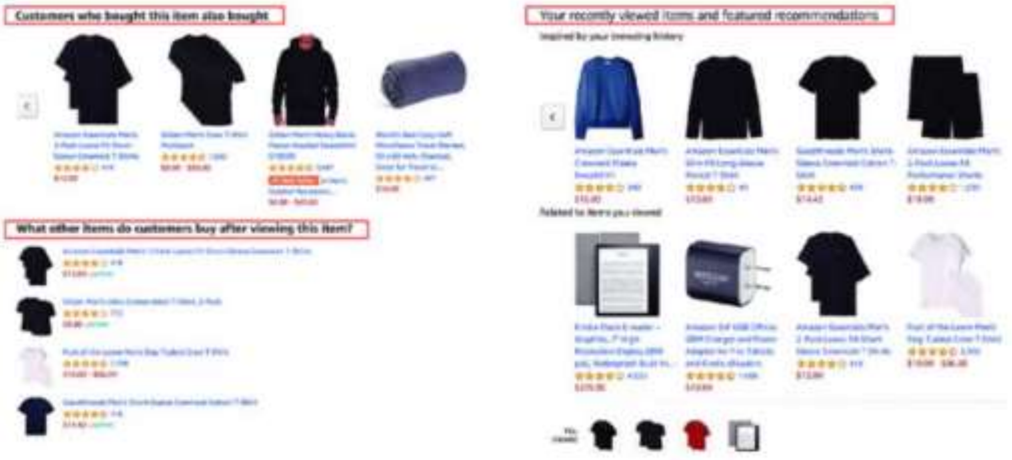


Figure 24: Amazon recommender system using the collaborative-based method



Customers who bought this item also bought

What other items do customers buy after viewing this item?

Your recently viewed items and featured recommendations

Inspired by your browsing history

Related to items you viewed

Figure 25: Amazon recommender system using the user-user approach

Chapter 5 Conclusion

5.1 Conclusion

In conclusion, marketing bias occurs often in every business, however, not every business draw best the potential of having their products marketed based on the consumer behavior as this provides a wider range of products sale. Proper marketing requires a great understanding of customers' needs and how they can be presented to them properly so that they purchase the services provided by companies and generate high profits. Recommender systems are considered as an answer to this marketing scheme; through this study it have been evident that the recommender system is a great fit for amazing marketing bias, as it provide their customers with great recommendation based on their preferences, this is established through feeding it enough data on customers' purchase history, their product purchases, rating of purchased products, which all contribute to the system understanding of what customers usually procure and what items can be proposed to them so that more sales can be conducted, resulting in higher profits. As it turns out the accuracy rate of the recommender system using the collaborative-based model is relatively high, which is great and show that Amazon takes care of its consumers and its marketing scheme.

5.2 Future Work

If there is any chance in the future I would like to further research the recommender systems on other marketing brands as I found that this topic is quite interesting and full of potential to increase profits for other brands when they market their products to consumers. Furthermore, if there are more datasets that corresponds to the above study, I would very much like to test them with the collaborative-based model and see it contribute more to the item-item approach or the user-user approach.

Bibliography

Anand, S. (2019). Amazon recommendation system | Data Science and Machine Learning. Retrieved 18 December 2019, from https://www.kaggle.com/general/126224

D. H. Setiabudi, G. S. Budhi, I. W. J. Purnama and A. Noertjahyana, "Data mining market basket analysis' using hybrid-dimension association rules, case study in Minimarket X," 2011 International Conference on Uncertainty Reasoning and Knowledge Engineering, Bali, 2011, pp. 196-199, doi: 10.1109/URKE.2011.6007796.

Ekstrand, M. D., Riedl, J. T., & K, J. A. (2011). *Collaborative Filtering Recommender Systems.* Minnesota: Foundation and Trends.

Gemmis, M. d., Lops, P., Musto, C., Narducci, F., & Semeraro, G. (2015). *Semantics-Aware Content-Based Recommender Systems.* Boston, MA: Springer.

Hardesty, L. (2019, 11 22). *The history of Amazon's recommendation algorithm*. Retrieved from https://www.amazon.science: https://www.amazon.science/the-history-of-amazons-recommendation-algorithm

Knijnenburg, B. P., & Willemsen, M. C. (2015). *Evaluating Recommender Systems with User Experiments.* Boston, MA: Springer.

Krishnamurthy, S. (2002). *CASE #1- AMAZON.COM- A BUSINESS HISTORY.* Washington DC: E-COMMERCE MANAGEMENT: TEXT AND CASES.

Martinez, M. (n.d.). *Amazon: Everything you wanted to know about its algorithm and innovation*. Retrieved from https://www.computer.org: https://www.computer.org/publications/tech-news/trends/amazon-all-the-research-you-need-about-its-algorithm-and-innovation

McAuley, J., Lakkaraju, H., & Leskovec, J. (2013). Understanding the interplay between titles, content, and communities in social media. Retrieved from https://cseweb.ucsd.edu: https://cseweb.ucsd.edu/~jmcauley/datasets.html#market_bias

Meteren, R. v., & Someren, M. v. (2000). *Using Content-Based Filtering for Recommendation.* Amsterdam: University of Amsterdam.

Rocca, B. (2019, 06 03). *Introduction to recommender systems*. Retrieved from https://towardsdatascience.com/: https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada

Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikari , R., & Duval, E. (2011). *Dataset-driven research for improving recommender systems for learning.* Belgium: LAK.

Zaier, Z., Godin, R., & Faucher, L. (2008). *Evaluating Recommender Systems.* Florence, Italy : IEEE.