Yutong Wu

U98858372

Collaborators: none

Final Project write-up

I choose the dataset

In this project, I aim to investigate

The data was collected in the summer of 2006. https://snap.stanford.edu/data/amazon-meta.html

This is the summary statistics for the dataset.

Category: Music

Number of Products: 103144

Average Sales Rank: 154302.87

Average Review Rating: 4.37

Category: Video

Number of Products: 26131

Average Sales Rank: 31713.51

Average Review Rating: 4.11

Category: Toy

Number of Products: 8

Average Sales Rank: 12921.38

Average Review Rating: 4.51

Category: Video Games

Number of Products: 1

Average Sales Rank: 339.00

Average Review Rating: 2.33

Category: Baby Product

Number of Products: 1

Average Sales Rank: 1017.00

Average Review Rating: 4.67

Category: Software

Number of Products: 5

Average Sales Rank: 2016.20

Average Review Rating: 4.46

Category: Book

Number of Products: 393561

Average Sales Rank: 630906.88

Average Review Rating: 4.16

Category: Sports

Number of Products: 1

Average Sales Rank: 4684.00

Average Review Rating: 4.20

Category: CE

Number of Products: 4

Average Sales Rank: 66277.50

Average Review Rating: 3.00

Category: DVD

Number of Products: 19828

Average Sales Rank: 25399.30

Average Review Rating: 4.05

There are several categories in the dataset, but some categories have few products. Under the consideration of memory management and stack overflow, I decided to pick the top three categories with most products, which are Book, Music, and Video.

Since not all products have similar products, I will exclude those product without similar.