Problem Statement (Academic Project)

Say you are a manager at a media company and want to start a new product line to boost your revenue. Being a media company, you want to get into either ebooks, movies, TV shows or similar such product categories in the entertainment industry.

Since you are starting a new product line, you want to be sure about the choice of products you'll buy (and sell). You have **three options** of product categories to choose from - CDs and Vinyl, Movies and Ebooks (Kindle).

Specifically, you want to use the gigantic dataset to identify:

1. Which product category has a larger market size
2. Which product category is likely to be purchased heavily
3. Which product category is likely to make the customers happy after the purchase

Since you do not have the actual sales data of the products, you will have to use some 'proxy-metrics' to estimate the aforementioned metrics.

For example, you can use the 'number of reviews' as a proxy for the number of products sold (i.e. the ratio of number of reviews of two product categories will reflect, approximately, the ratio of the number of units sold - you are not trying to estimate the absolute numbers anyway, you want to compare the metrics across categories).

Using similar logic, to estimate the market size, you can use the number of reviewers as a proxy.

Similarly, you can define some metrics to proxy customer satisfaction as well. For instance, if a customer has written a long review and rated the product 5, it indicates that he/she is quite happy with the product.

In short, you are trying to use data about reviews to estimate some key metrics, which will help you identify the product category you should be investing in.

Broad Methodology

Data Pulling

First, grab the dataset. Here is the home page of the data. We will be using the 5-core data, where each user and product have at least 5 reviews. Locate the three categories mentioned and find the links for these.

Data Dictionary

Attribute	Data Type	Description
reviewerID	string	ID of the reviewer, e.g. A2SUAM1J3GNN3B
asin	string	ID of the product, e.g. 0000013714
reviewerName	string	Name of the reviewer
helpful	array	Helpfulness rating of the review, e.g. 2/3
reviewText	string	Text of the review
overall	double	Rating of the product
summary	string	Summary of the review
unixReviewTime	long	Time of the review (UNIX time)
reviewTime	string	Time of the review (raw)

Helpfulness score

Each review comes with a 'helpful' variable, as shown in the image below. This variable is calculated by collecting other people's votes on the question "Was this review helpful to you?"

The dataset presents this score in the form of a tuple, where the first integer is the number that voted as helpful and the second one is the total number of votes. For example, if the helpfulness score reads "10 of 15 people found this helpful", the tuple will be (10, 15). For a given tuple, the helpfulness score will be (10/15). A good start is to consider those reviews where at least 10 people have voted (i.e., the 2nd number in the tuple is greater than 10).
You may also divide the reviews into bins of increasing length (of review text). Now, find the average helpfulness score for all reviews in that particular bin. Can you spot a trend? Are longer reviews more helpful on an average? Should you give a higher weightage to longer reviews?

Next, you could examine the average helpfulness score over the 5 different rating levels, and explore if the helpfulness scores vary.

These are only suggestive ideas to get you started; you may choose to follow any methodology that you think can help achieve the business objective.

Results expected

You are expected to conduct exploratory analysis to find answers to the questions mentioned above and recommend one product category (out of the three shortlisted) which you'll be investing in.

Along with the code, your comments should clearly demonstrate your thought process, summarise key results and ultimately use the results to make a decision.

Write all your code in one well-commented R file. You need to submit only the R file (no presentation or report), and hence, comments are really important for evaluation.