Questions

1. **Why choose book description instead of looking though genre?**
    a. Book descriptions are much more varied and unique whereas genres show much less diversity of differences to discern between.
2. **Would your content based system allow for multiple filters (e.g., description and genre and length)?**
    a. As it currently stands no. This would be something to expand upon in the future.
3. **Why drop 0 ratings from the first dataset, but not the second?**
    a. The number of) ratings in the first dataset was quite large and thus would have had a greater impact on the results, particularly since the method used on the first dataset was concerned with ratings to make the recommendations. The second dataset have much fewer 0 ratings (~300) and the recommendations were not based on the ratings so those books were left in.
4. **Why not redo the rating-based analysis with the second dataset and see if you get better results?**
    a. This could certainly be done, however, the methodology used for the first dataset is considering each user's rating of each book whereas the second dataset only has a single average rating for the book with no user variation. Also, the analysis showed that the results from the content-based recommendation far surpassed that of the rating-based method so there does not seem to be a point in doing so.
5. **By grouping the second dataset by title, aren't you potentially erasing titles? You mentioned same times by different authors – these would be merged together and only one would survive.**
    a. This is certainly true. However, to avoid grouping different titles together would have required a lot more data cleanup than this project had time for. There were many variations both in the titles and in the authors listed even within the same book. Each similar title would have to be cleaned and a common definition agreed upon to represent it. Unfortunately, merging the titles as I did may have eliminated a few titles from the database, but is outweighed by the vastly reduced the number of duplicated entries.
6. **For the ratings based system you mentioned having only a few reviews may skew the average. Could you take the mode average instead of mean average?**
    a. This suggestion would work if more titles had more reviews. Taking the mode of 1 or 2 ratings is potentially worse than taking the average. This could however be an interesting spin on the analysis for the books that do have more ratings.
7. **Are the databases limited to in-print books? Do they include classic literature?**
    a. The databases do include a wide variety of books both in and out of print. However, books in the database all have ISBN numbers which suggests that the books need to have been in print (or reprint) recently. Many of the

classics have been reprinted numerous times and are represented in these datasets.

8. **What happens if a book I input isn't in the dataset?**
    a. The recommender will return an error. If the book is not in the dataset, the recommender will have nothing to reference your search to and won't know what to do.
9. **How quickly can new books be added to the system?**
    a. New books can be added immediately, even prior to publish, in the content-based system since all you need is the information about the book (e.g., Title, Author, Summary). Adding new titles to the rating-based system is a bit trickier since you need readers to read and rate it before it adds value in the dataset. When to add it becomes a matter of defining your acceptable threshold. Do you add it once the book has at least one rating or do you wait until it has "critical mass" of ratings for a good average?
10. **What are the "unnecessary columns" for the content based system?**
    a. The unnecessary columns in the second dataset that were dropped included:
        i. book_edition
        ii. book_format
        iii. book_isbn
        iv. book_pages
        v. book_rating
        vi. book_review_count
        vii. image_url

    These columns, such as pages and rating, could be useful for future iterations of the recommender system, but were not used in the current one I tested. The columns and their corresponding data still live in the original dataset.