For my capstone project I am using the Goodreads 10k dataset. This dataset comes from a Kaggle competition in which you use the dataset to create a machine learning recommendation service. Thus, this data has been cleaned much more deliberately than other datasets. However, I still had some data cleaning that had to take place in order for me to get the data ready for analysis.
I loaded the dataset from the csv file (https://raw.githubusercontent.com/jeydion/goodbooks-10k/master/books.csv') that is provided in the Goodreads dataset. I stored the csv file as a pandas dataframe in order to work more effectively with the data. The dataset has 22 columns and is indexed by book id (I use two different id's in my index, one relating to the goodreads book ID database and another to find the position of the book in this specific pandas dataframe. I then began to clean the data. I removed several columns from the dataset that I deemed either as duplicate information or not relevant. For example three of the columns used various ID's for each book that were duplicated from another dataset. I am not using that data so these rows were omitted. There were two rows at the end that links to the cover images of each of the books. Since my analysis does not include image recognition, I omitted these as well (Though it would be an interesting side project if all the books had photo covers; creating an algorithm that would indeed judge a book by its cover).

I found there are 21 NAN values in the Year column. Since the amount is so low, I will manually replace the nan values with the correct years. I simply used Google to confirm the missing book year values.

There are 700 missing ISBNs from the DataFrame. Though there are methods for finding the missing values programatically. I have decided to drop these two columns from the DataFrame. According the isbn.org, each 10 or 13 digit number assigned to every book does not correspond with the genre. (https://www.isbn.org/faqs_general_questions). The 10 to 13 digit number also isn't used globally, despite the I standing for international. For example, some books that are published in different languages but have an English translation, have an ISBN for the English version of the books but not the original. This creates an issue when trying to compare the average ratings per book because the ratings might span across different language versions of the same book.

I will reference the books by their goodreads_book_id. This will help reference each book in the other datasets that are attached to this dataframe as well. The other datasets include additional information for each book. Most of these datasets include duplicate or irrelevant information. For instance, one dataset links each book to each user who rated that book. This would be useful if we had more data on the users, such as review frequency, average review rating and review patterns. However, the user id does not necessarily associate with each user that rated a certain book. For all the books in the dataset, they all have a user_id = 1, 2, 3 and so on. This presents the problem of there not being unique user ids for each book.

There are 1084 books that have nan values for lang (book language) upon review of all the books, they are all in english and the nan value can be replaced with 'eng'.

There are 67 books that are in Arabic and Persian, in which there might not be English translations or said translations aren't the version of the book the reviews are referring. It would be exceedingly difficult to compare the data with the english or english translation books. This ties into the issue with using ISBN's in the data. There isn't a ISBN for these books and the data associated with English translations are not part of this dataset. Also the number of reviews in general (in Arabic or otherwise) is very low (average < 200), so the data gathered from these will be limited in scope.

There weren't any outlier values other than publication year for some of the books not being in the last two hundred years. These books will remain in the dataset because they are still closely associated with many of the modern books on the list. Most of the books that predate 1800 are considered "classics" so they are commonly read in many languages and have many modern reviews.

Now that the data is cleaned I will begin the process of separating the test data; nominally 20% of the dataset, and continue the data analysis with the training set.