Capstone project 1 Milestone Report

The problem statement for my capstone project is: How might we predict the market demands for each book genre based on the sample data we have obtained from Goodreads website. This problem is important because the market demand for a specific book genre might be low but the market supply is high, or vice versa where the market demand for a book genre might be high yet the market supply is low. In another word, my capstone project is trying to reduce the scenario where many readers out there like a specific book genre yet not many writers are interested in writing about them or publishers refuse to publish those books.

I will be using the book review information and author review information provided by Goodreads site to close down the gap between the market demand and market supply for each book genre. My clients for this project are the book writers and book publishers. Publishers will use my report to make more informed decisions on whether to say yes or no to a book genre, to a book or to a book author thus increase the publisher's reputation as well as its profit. Writers will use my report as a reference to decide which topic or genre to write about for their next book and have a clear idea on the odds that their books will make a profit or become a best seller.

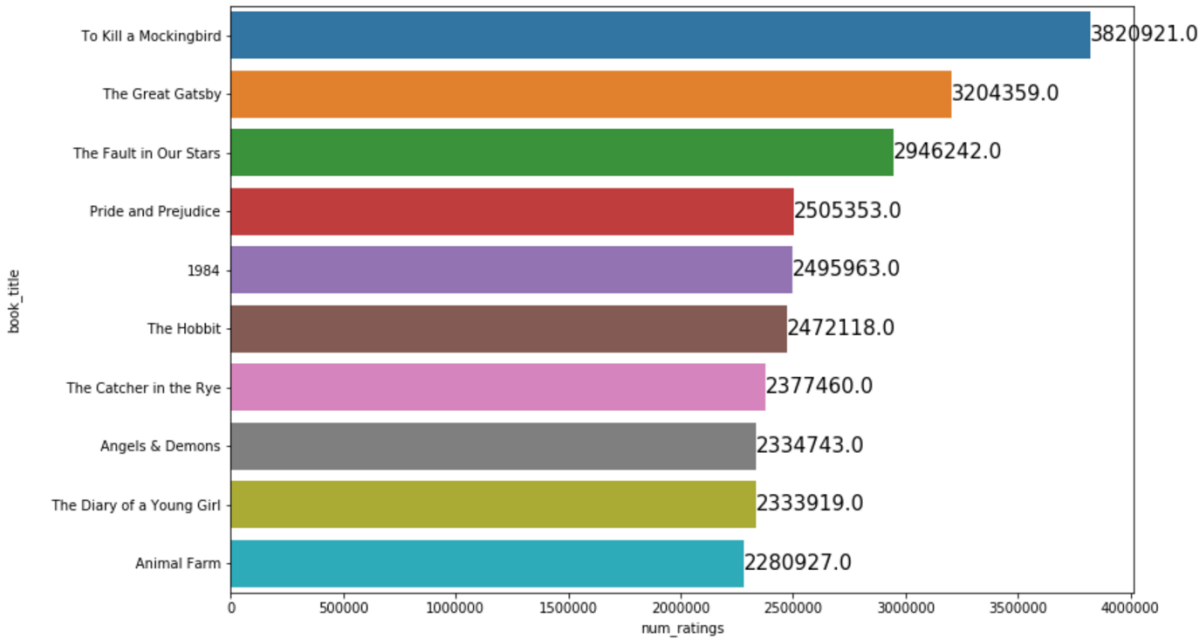
I acquired the Goodreads dataset from Kaggle. I will go through the dataset and identify the hidden relationships among various fields in the dataset such as author's rating, each genre book's rating and more to see which piece of information might be correlate with book selling and to see whether books readers like this book or genre. This will let me have an idea on the market demand for each genre or for a specific

author. Both book writers and publishers will have a clear idea on what is the demand out there in the market for each genre. This will help us close down the gap between market demand and market supplies, and slowly reaching a more optimized scenario for book market. My deliverables will include the code that I uses to predict the market demands, a paper that describe my capstone project and a Powerpoint slides that summarize my findings.

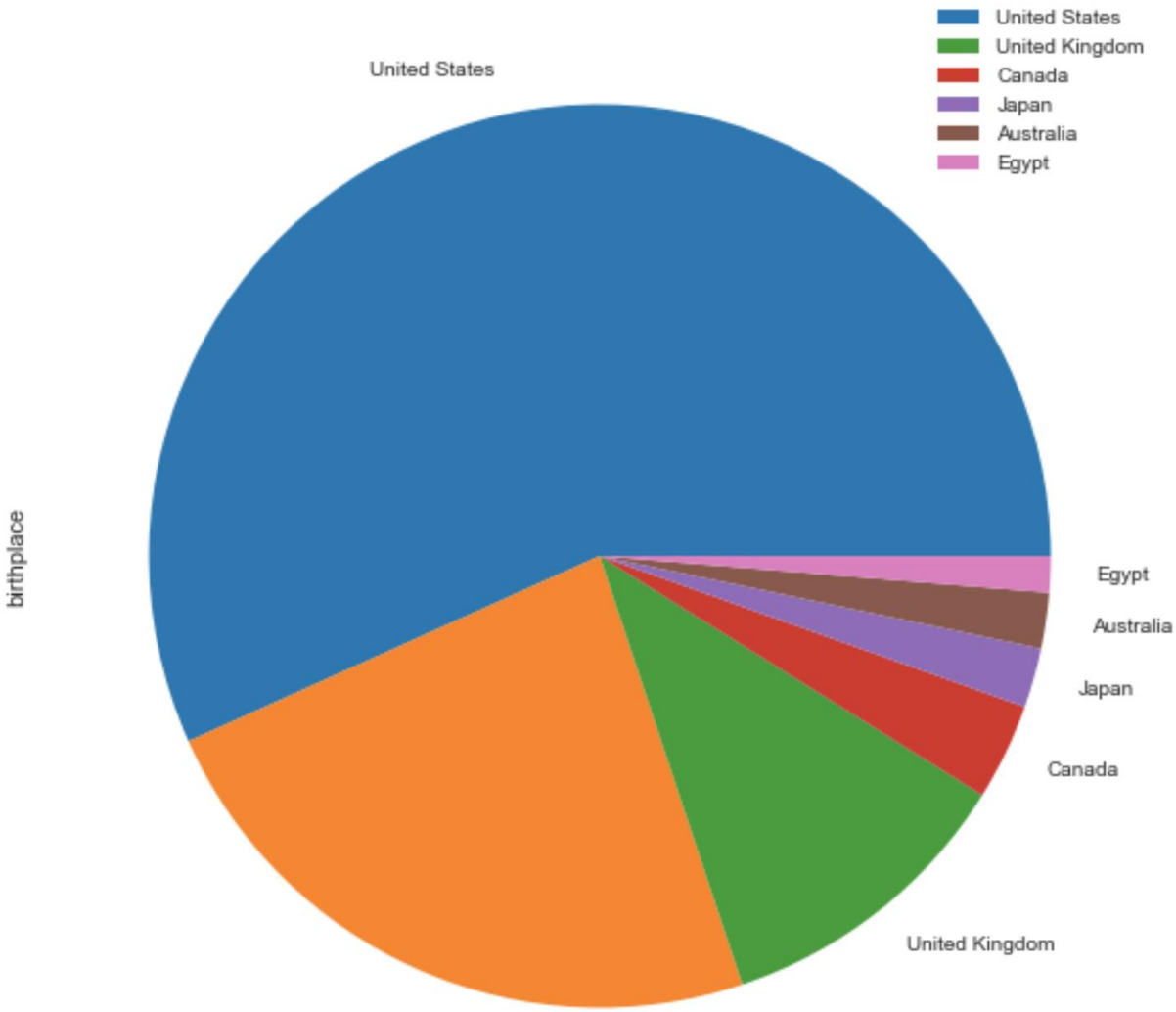
The data I obtained from Kaggle website was fairly clean. After I loaded the csv data into Pandas dataframe, I dropped all the unessential fields. I noticed that many columns had "\n" in them so I applied the replace method to replace all the "\n" with empty string. A few columns contain spaces before and after its value so I apply strip function to remove the extra spaces. There are some filling word data like "by" seems to appear out of nowhere and they do not fit with other values in the column so I replace them with empty string as well. In addition, many book titles repeated more than one times and I applied drop_duplicates method to remove those duplicate rows.

After I did the data analysis part, I found some interesting insights. The books and authors with the highest average rating are the ones with fewer total review count. More than half of the authors listed on Goodread database are from USA, the site is English so it is not surprising. The books with the highest number of review, the highest number of rating are fiction genre. The author with highest author rating count are fiction. The total number of fiction books written by female are 1995 and 2979 by male. Romance and fiction genre have the most authors. In the sample, we have a total of 7642 female authors and a total of 8708 male authors. The average author rating has a mean of 3.958102141, with the minimum value of 1.82

and maximum value of 5. The author rating count has a mean of 154213.3758, with the minimum value of 6 and maximum value of 21117318. The book average rating has a mean of 3.951413456, with the minimum value of 0 and maximum value of 5. The number of rating has a mean of 26996.7526, with the minimum value of 0 and maximum value of 3820921. The number of review has a mean of 1455.380122, with minimal value of 0 and maximum value of 147696. Below is the top 10 most rated books.



Below is the pie chart based on the author's birthplace. The orange ones are null. Most likely from USA as well.



The variables that are particularly significant in terms of explaining the answer to my project question: How might we predict the market demands for each book genre based on the sample data we have obtained from Goodreads website. Columns num_reviews, num_ratings, genre_1 and genre_2 and author_genre columns, these columns are important to my inferential statistics. Especially num_reviews and num_ratings columns, they show how popular a book is to the readers, to the public. The number of reviews and number of ratings give us an idea how many people most likely have read the books and are willing to spend their valuable time to write a review for the book.

I will use bootstrap inferential statistics to resample the sample data we have from Goodreads. I will run 10,000 res-sampling to find the correlation between num_review and num_ratings columns for each genre to predict its average rating. In addition to these two columns, I will also use its author's rating count as well as its author's average rating columns to help me with the statistics modeling.