Project 2 Proposal.

For Project 2, our group is going to study the data on books. We will compare data both gathered from a Json we have of the New York Times and a CSV file of books on Goodreads.

First, an explanation of the data sets. The New York Times Bestseller list Json is simply going to list out each week and what books were on the list, with their rankings. We will go through that data and find a value for each book telling us the book’s highest-ranking and how many weeks the book was on the list. Our second data set will be a CSV file of books and their data from good reads. The data from good reads will tell us the number of user reviews, ratings, how many pages.

To make narrow the scope of the project we will be looking at fiction and non-fiction independently. We will compare a few metrics. 1) comparison for the number of weeks a book is on the NYT’s bestseller list to Good Reads average score. We will also compare the length of a book and how well it performed with the NYT’s top rankings and weeks on the list.

Important note. The New York Time’s bestseller list is sales data that is ultimately an editorial product. Because of that, The NYT’s data is acting as a place holder. We may end up using the data, however we are going to try to seek a better data source for sales data. This might simply be the number of amazon reviews. Believing that the number of reviews a book gets on Amazon the more books have been sold.

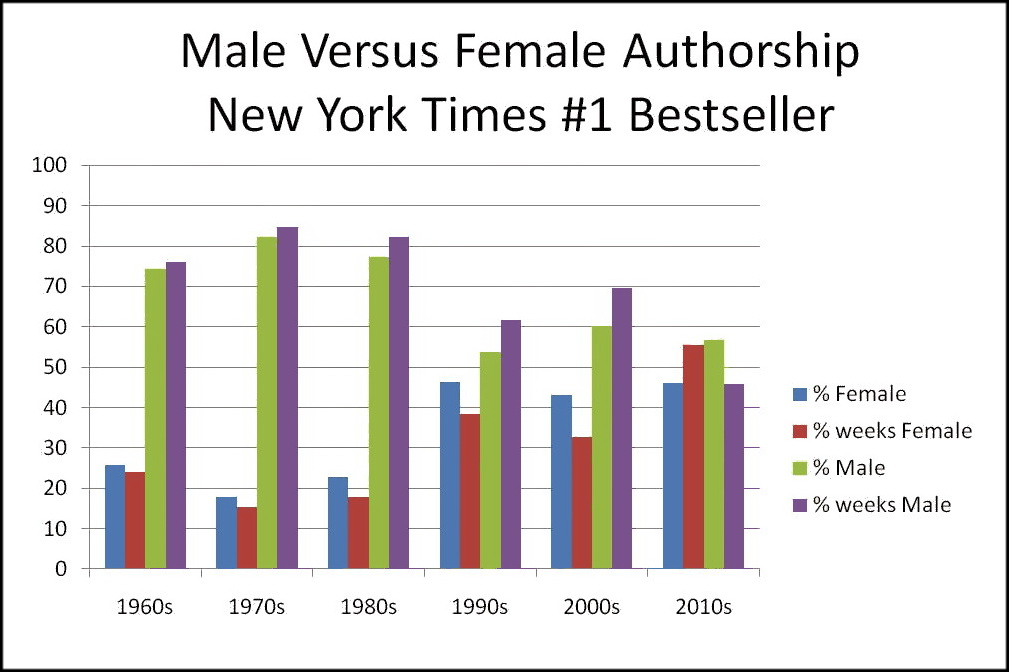
To visualize the data we will be mostly looking at Scatter plots, comparing the data mentioned above. From this, we will be looking into statistics libraries for Java and see if we can find a relation, and hopefully graph the relation, that shows the number of weeks a book is on the list and the number of pages of the book. If we can find the data, will also find relationships between data. If we continue with the new york times list. We will also look into a line graph that shows a book’s rankings throughout the list.

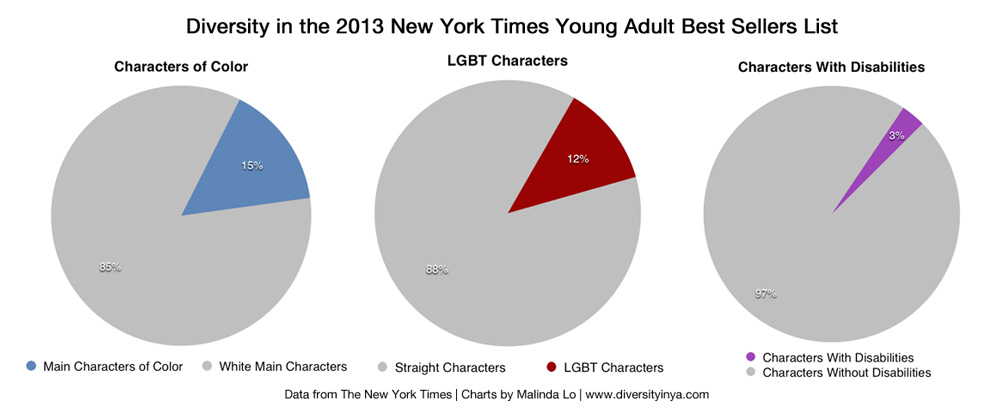
Here is a link to the data we will use:

<https://www.kaggle.com/cmenca/new-york-times-hardcover-fiction-best-sellers#nyt2.json>

<https://www.kaggle.com/jealousleopard/goodreadsbooks>

Here are some other graphs that show some of NYT’s data for inspiration.





Our GitHub

<https://github.com/ablinn91/Project-Books>