1. Problem:

What are the typical qualities of a best selling book?

Goal: Predict sales ranking of Amazon books.

1. Motivation:

Help publishers choose manuscripts that are more likely to become best sellers.

1. Data:

* Unique books from 800 books in Amazon book lists:
* 100 new york times best sellers
* 100 books to read in a lifetime
* 100 Sci-fi & Fantasy books
* 100 Biographies & Memoirs
* 100 Young Adult books
* 100 Mysteries & Thrillers
* 100 Leadership & success books
* 100 Children’s books to read in a lifetime
* For each book, store the following attributes:
* Book name
* ASIN or ISBN (used to prevent duplicating scraping)
* Average rating out of 5 star
* 1-5 star rating percentage out of all reviews
* Ranking in its sales
* The number of reviews
* A list of all review titles
* A list of all review texts
* Book list category

1. Data collection and Analysis:
2. Web scraping

* Start from a book list URL, like <https://www.amazon.com/b?node=8192263011>

And list all the URL links of books as well as their full-review links

* Loop through each review link, scrape its corresponding book link to get the Average rating, ASIN, sales ranking and its number of reviews
* Iterate through all of the review pages to scrape all reviews and add to the list of review texts and review titles
* Store all results as JSON format

1. Data Wrangling

* Organize data into the format of a data frame with each row representing one book, and its columns filled with attributes.
* For each book, count the total occurrences of each word in all review titles and review texts respectively.
* Generate attributes named with individual words for review title and review texts such as “title-enjoyable” and “text-enjoyable”, and their values are the counts of their occurrences
* The total number of columns of the final data frame depends on the total number of words that appeared in all reviews of all books

1. EDA and write a story

* Relationships among attributes other than word counts
  + What are basic statistics of the sales ranking and average review ratings?
  + Do certain book lists have higher sales ranking than others?
  + What’s the relationship between average review rating as well as the number of reviews and sales ranking?
* Which words (top 20 or 50) are the most common in review titles and in review texts?
* How do the most common words differ across sales ranking categories and average review ratings?

1. Inferential statistics

* Apply chi-squared tests on whether there was statistically significant differences in sales rankings across book lists
* Add more according to findings from EDA

1. Machine learning analysis

* Feature engineering: random forest, Lasso etc. to find the most important features
* Apply Naïve Bayes , SVM and maybe Neural Network (maybe not enough observations) to predict sales ranking using preselected features
* In each prediction, subset data into training and testing set so that the model could be optimized using cross-validation