Problem statement:

Writing a bestseller is hard. However, a bestseller can create a generous income stream that trickles from the author to publishers and retailers. But there are thousands of titles being published every year and most of them goes unnoticed. Moreover, the digital conversion to e-books, with its simplified publishing process, also makes this industry more competitive than ever. The problem comes from the fact that good writing is only half of the battle. Knowing what make a book popular with readers provides the remaining insights on creating the next bestseller. The goal is to help authors designing their next e-book that have a higher chance to become a financial success!

The following analysis seeks to find the best features that helps e-book sell well. What genre should it focuses on? Does rating matter? The task is to build a prediction model on unit sale of e-book and from there determine the main features effecting book sales. Two algorithms are used to model this supervised problem: linear regressions and tree-based regression.

Description of the dataset

This project will use the publishers dataset (https://corgis-edu.github.io/corgis/csv/publishers/) that includes e-book sales data from Amazon for 27k titles in 2015. Since this data comes from a standard source, it requires minimal data wrangling. However, one must address multiple redundant columns along with many outliers in the data.

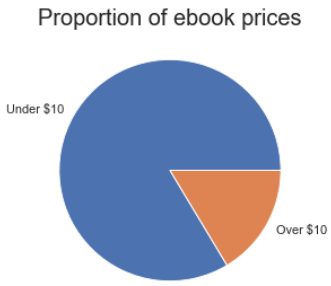
Impressions on the data and steps taken are as followed:

1. There is no missing value in the data.
2. The data is clearly labeled with correct format but the column name contains dot/space with mixing capital letters. Thus, columns' names are changed to lowercase and converted into snake case. Suffixes are also dropped.
    a. Ex: Statistics.Sale.Price => sale_price
3. Multiple columns with sales statistic including author revenue, total revenue, gross sales heavily correlate with each other. Thus, we remove all of them except for daily average units sold as the sole dependent variable. This step will ensure no collinearity exits when building the models.
4. Drop sale rank column with the same reasons as above.
5. 'sold by' column includes different names for the same company. They are adjusted to the same company.
    a. Ex: Harper Collins Christian Publishing, Publishers and Publishing become HarperCollins.
6. Outliers appear in many columns such as prices, sales and total number of reviews. Removing them will take away the usefulness of the data. Therefore, all outliers are left as is for exploratory analysis. But when building model, those columns will be log transformed to take away the long tail.
7. Lastly, the cleaned data set is renamed to book_sale.csv and saved.

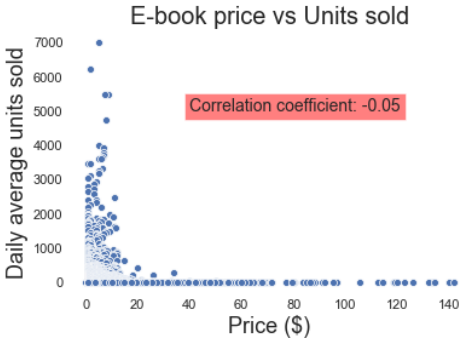
Initial findings from exploratory analysis

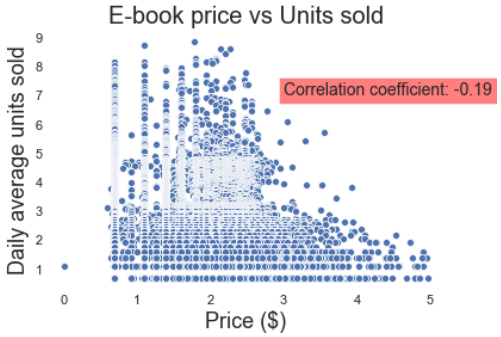
The initial findings from exploratory analysis shows 4 main features effecting e-book sales: price, ratings and reviews, publishers and genre. Most e-book titles average less than 100 units sold daily. However, the data is very skewed with about 100 titles that sell over 1000 copies daily. This is important thing to note before building a predictive model. Nevertheless, the findings are as followed:

1. Price:
    a. sale_price has a large variance. Almost all books with price higher than $30 is non-fiction. One reason for this is non-fiction includes textbooks, academic books, collection books that are very expensive to produce with high markup.



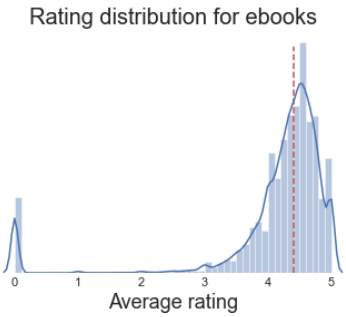
    b. There is a very minor relationship between price and sales without transformation. However, with the right transformation (log), the relationship is better shown.



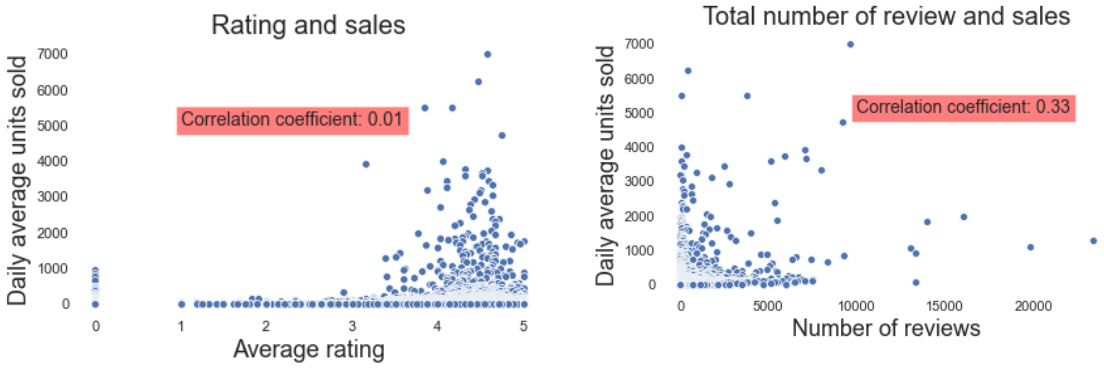


    c. Median price of 5.12 is very closed to average book price of 6.78. Over 75% of e-books has the price tag under 10 dollars.

2. Ratings and reviews
    a. Rating is scored from 1 to 5. If the score is missing, it is replaced by 0. The distribution of rating is skewed right with most of the e-books have 4+ star rating. A small set of e-books that does not have any rating and thus are rated at 0.

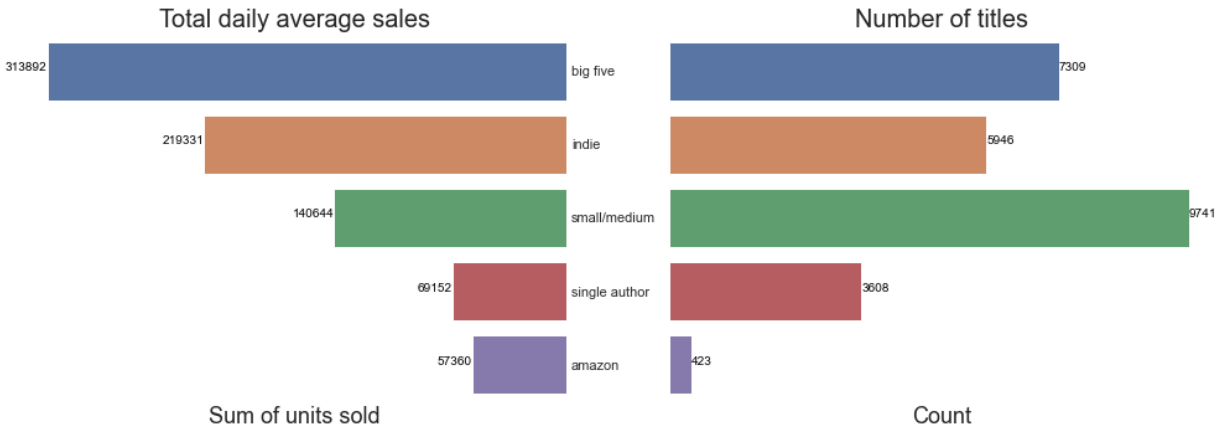


b. Rating does not seem to correlate with sales. The correlation is only at 0.01. This is very understandable as most books have high rating and thus decrease its significance.

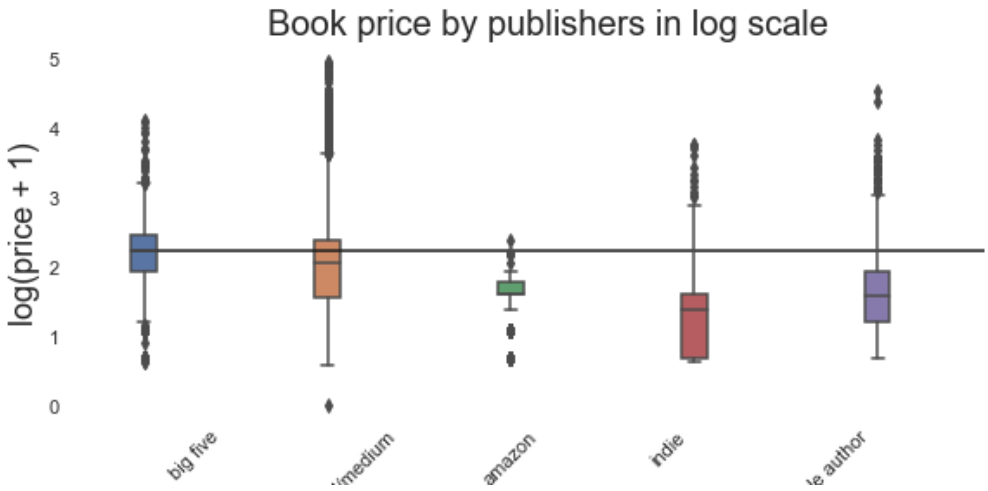


c. In contrast, number of reviews effect sales greatly with the correlation coefficient at 0.33. However, would more sales generate the number of reviews or the opposite? While it makes sense that a popular book has more people writing reviews about it, there are many ways that author can do to encourage more people writing the review about the book too. Thus, number of reviews is a legit factor in our analysis.

3. Publishers
   a. There are over 300 publishers in the dataset with smaller ones grouped together as indie, medium, single authors... The big five publishers are: Penguin Random House, HarperCollins, Macmillan, Hachette Book Group, and Simon & Schuster.
   b. With a more resources at hand, one would expect e-book by big 5 do better than the average e-book and it is true:
      i. big five has less titles count but a much better sale figure
      ii. big five has a higher median starting price
   c. However, trying to get published by the big 5 is impractical because they do not accept author manuscript without literary agency.



d. With the same log transform as before, the distribution of e-book price is better shown across different publishers. Big five has the highest median price with balanced variation. All other publishers have lower median price and a long tail toward the top (very high prices)



4. Genre
   a. non-fiction and genre fiction dominate the market in both number of titles and sales. Only these two have multiple titles that sales on average 1000+ copy daily.
   b. Trying to be popular in niche genre such as children (Dr.Seuss) and comics (Stan Lee) is much harder compared to other genres.