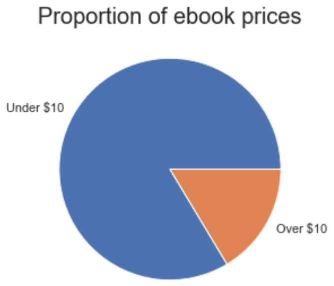
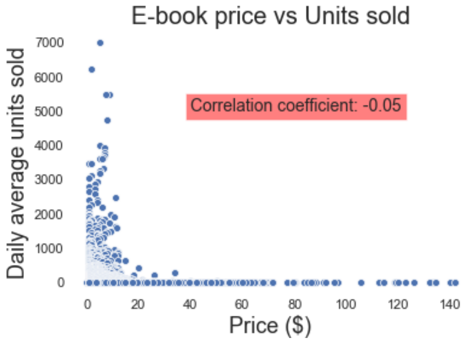
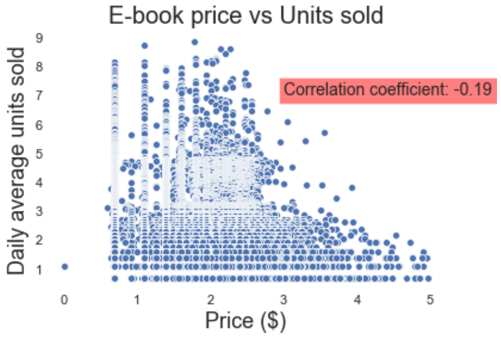
Inferential Report

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:

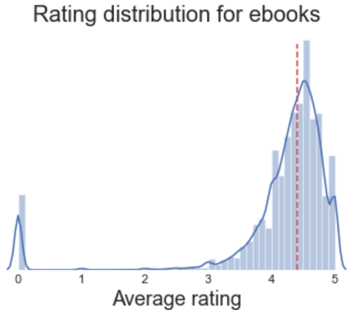
1. Price:
   a. sale_price has a large variance. Almost all books with price higher than $30 is non-fiction. One reason for this is non-fiction includes textbooks, academic books, collection books that are very expensive to produce with high markup.
   b. There is a very minor relationship between price and sales without transformation. However, with the right transformation (log), the relationship is better shown.



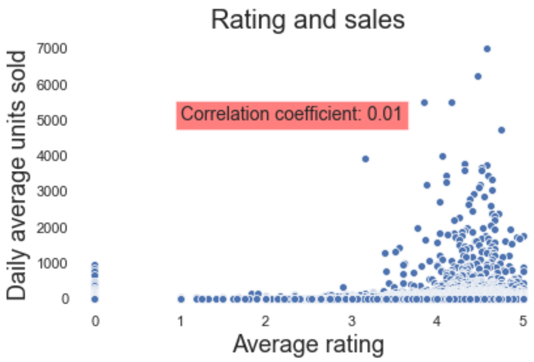


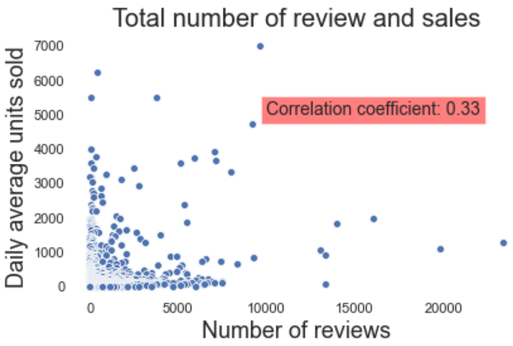


   c. Median price of 5.12 is very closed to average book price of 6.78. Over 75% of e-books has the price tag under 10 dollars.
2. Ratings and reviews
   a. Rating is scored from 1 to 5. If the score is missing, it is replaced by 0. The distribution of rating is skewed right with most of the e-books have 4+ star rating. A small set of e-books that does not have any rating and thus are rated at 0.



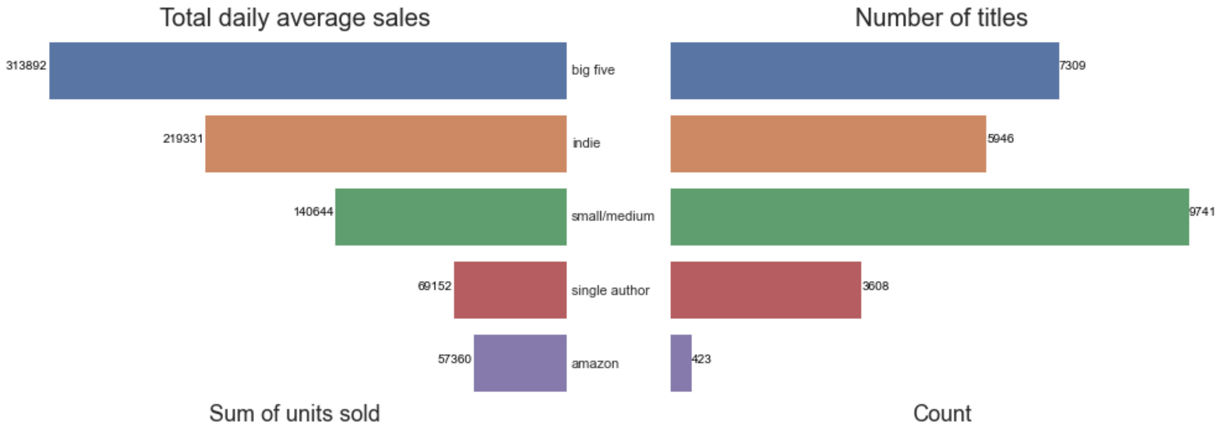
b. Rating does not seem to correlate with sales. The correlation is only at 0.01. This is very understandable as most books have high rating and thus decrease its significance.



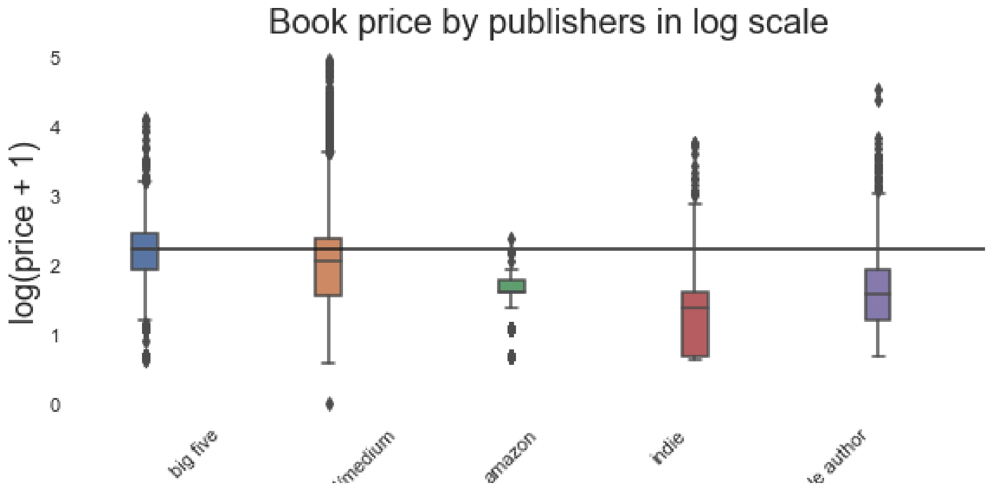


c. In contrast, number of reviews effect sales greatly with the correlation coefficient at 0.33. However, would more sales generate the number of reviews or the opposite? While it makes sense that a popular book has more people writing reviews about it, there are many ways that author can do to encourage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

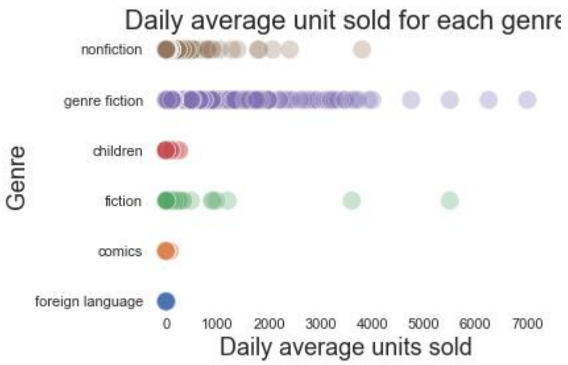
3. Publishers
   a. There are over 300 publishers in the dataset with smaller ones grouped together as indie, medium, single authors... The big five publishers are: Penguin Random House, HarperCollins, Macmillan, Hachette Book Group, and Simon & Schuster.
   b. With a more resources at hand, one would 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.

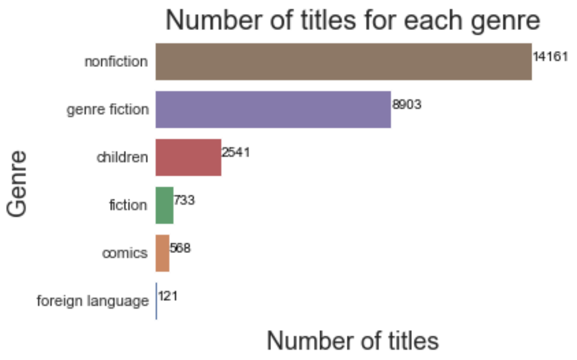


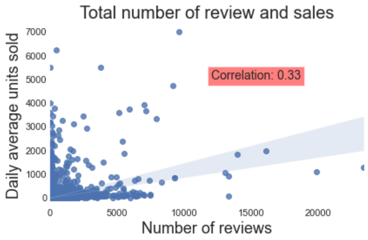
d. With the same log transform as before, the distribution of e-book price is better shown across different publishers. Big five has the highest median price with balanced variation. All other publishers have lower median price and a long tail toward the top (very high prices)

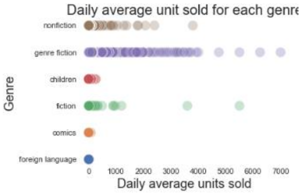


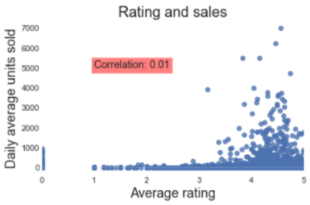
4. Genre
    a. non-fiction and genre fiction dominate the market in both number of titles and sales. Only these two have multiple titles that sales on average 1000+ copy daily.
    b. Trying to be popular in niche genre such as children (Dr.Seuss) and comics (Stan Lee) is much harder compared to other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

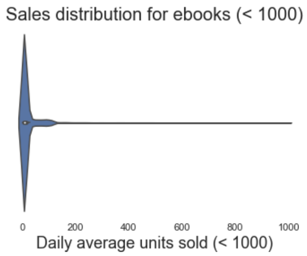


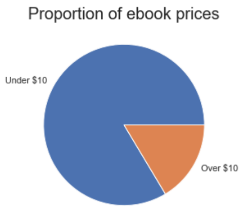


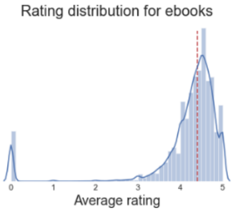

Total number of review and sales
Daily average units sold
7000
6000
5000
4000
3000
2000
1000
0
Correlation: 0.33
0
5000
10000
15000
20000
Number of reviews

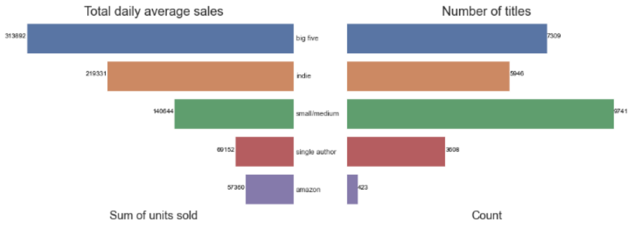

Daily average unit sold for each genre
Genre
nonfiction
genre fiction
children
fiction
comics
foreign language
0
1000
2000
3000
4000
5000
6000
7000
Daily average units sold

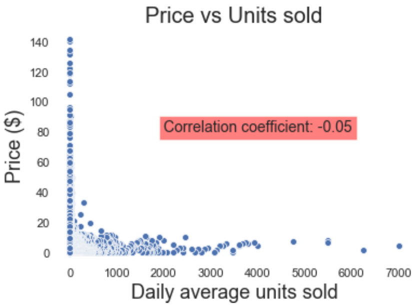

Rating and sales
Daily average units sold
7000
6000
5000
4000
3000
2000
1000
0
Correlation: 0.01
0
1
2
3
4
5
Average rating

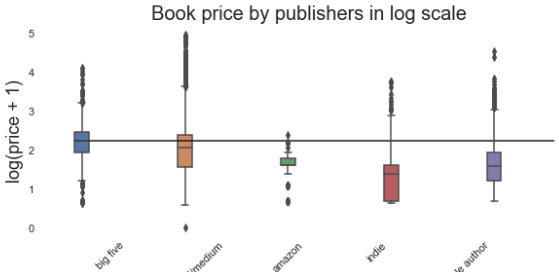

Sales distribution for ebooks (< 1000)
0
200
400
600
800
1000
Daily average units sold (< 1000)

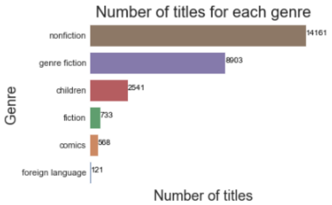

Proportion of ebook prices
Under $10
Over $10

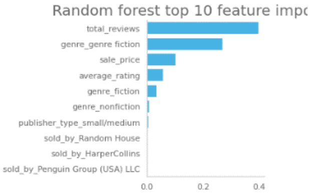

Rating distribution for ebooks
0
1
2
3
4
5
Average rating


Total daily average sales
Number of titles
313892
219331
140644
69152
57360
big five
indie
small/medium
single author
amazon
7309
5946
9741
3608
423
Sum of units sold
Count


Price vs Units sold
Price ($)
140
120
100
80
60
40
20
0
Correlation coefficient: -0.05
0
1000
2000
3000
4000
5000
6000
7000
Daily average units sold


Book price by publishers in log scale
log(price + 1)
5
4
3
2
1
0
big five
/medium
amazon
indie
le author


Number of titles for each genre
Genre
nonfiction
genre fiction
children
fiction
comics
foreign language
14161
8903
2541
733
568
121
Number of titles


Random forest top 10 feature impo
total_reviews
genre_genre fiction
sale_price
average_rating
genre_fiction
genre_nonfiction
publisher_type_small/medium
sold_by_Random House
sold_by_HarperCollins
sold_by_Penguin Group (USA) LLC
0.0
0.2
0.4