



FELLOWSHIP PROGRAM

ASSIGNMENT COVER SHEET

THIS FORM MUST BE AT THE FRONT OF EACH ASSIGNMENT
CANDIDATES MUST KEEP A COPY OF THEIR ASSIGNMENT

Candidate to complete the following section (and update details in header and footer):

	COURSE: Data Analytics Applications
DATE DUE: Monday, 11 April 2022 at 12:00pm (AEST)	

PLAGIARISM

By submitting your assignment, you are implicitly stating that the work is your own.

Remember that an important aspect of being a professional actuary is to always act with integrity.

Committing plagiarism by copying another person's work or not properly referencing other sources used in your assignment is a breach of the Integrity principle under the Actuaries Institute's Code of Conduct.

Be aware that your assignment may be vetted using Turnitin.



Question 1: Characteristics of the Book Publishing Industry

Word count excluding references and headings: 1498

Characteristic 1: Composition of the book publishing industry

A key characteristic of the book publishing industry is the oligopolistic nature and influence of publisher scale on book sales.

Historically, the primary participants within the industry were the ‘Big 6’ publishers, namely:

1. Penguin
2. Random House
3. HarperCollins
4. Hachette Book Group
5. Simon & Schuster
6. Macmillan

Book Publisher	2021 US Market Share
Penguin Random House	40%
HarperCollins	20%
Hachette Book Group	10%
Simon & Schuster	10%
Macmillan	5%

The oligopolistic nature of the book publishing industry is evident as in 2012, the ‘big six’ publishers accounted for 50% of book sales in the US (*Neofield 2015*). However, in 2021 the top five publishers (following the acquisition of Penguin by Random House) accounted for 85% of total US book sales as shown in the table above (*The Business Research Company 2022*). This dominance of large-scale publishers is expected to continue with Penguin Random House announcing its intentions to acquire Simon & Schuster in late 2021 (*Lee 2021*).

Several studies have identified that books with similar characteristics faced significant differences in sales volume primarily due to the scale of the publishing house distributing the book. Larger publishers were able to strongly influence title success due to their prior experience, greater negotiating powers with retailers and increased financial resources allowing them to attract renowned authors (*Throsby, Zwar and Morgan 2018*).

Relevance to analysis:

As the context concentrates on a small publisher, it is critical to differentiate between the scale of publishers in the data. The importance of accounting for publisher scale is identified by Michel (2016) who studied the sales outcomes of a series of short stories on Bookscan and observed larger publishers had approximately twice the sales volumes compared to small independent publishers despite the books sharing similar characteristics. Therefore, considering publisher scale in the analysis will ensure the predictions presented to the small publisher reflect the characteristics of their business.

Furthermore, it is important to consider the changing market dynamics as the continuing growth of large-scale publishers will further erode the sales power of small and mid-tier publishers. Therefore, the historical sales outcomes for small-tier publishers in the data may not be reflective of future outcomes.



Sources for Characteristic 1:

www.thebusinessresearchcompany.com. (n.d.). Book Publishers Market Analysis, Size And Trends Global Forecast To 2022-2030. [online] Available at:

<https://www.thebusinessresearchcompany.com/report/book-publishers-global-market-report> [Accessed 1st March 2022].

Neofield, S. (2015) How Many Publishers Are There, Really?. [online] Available at:

<https://www.sarahneofield.com/how-many-publishers-are-there-really/> [Accessed 1st March 2022].

Alter, A. and Lee, E. (2020). Penguin Random House to Buy Simon & Schuster. The New York Times. [online] 25 Nov. Available at: <https://www.nytimes.com/2020/11/25/books/simon-schuster-penguin-random-house.html>.

Michel, L. (2016). Everything You Wanted to Know about Book Sales (But Were Afraid to Ask). [online] Available at: <https://electricliterature.com/everything-you-wanted-to-know-about-book-sales-but-were-afraid-to-ask/>



Characteristic 2: Influence of Author Popularity on Sales

Another characteristic of the book publishing industry is the influence of author power in generating sales. Author power refers to the brand of a recognised author and is a strong signal of loyalty that is easily identifiable in the market. The author's power is approximated by the success of previously published titles. Successful titles generally develop a following with readers, thereby resulting in a higher probability of purchasing later titles as readers form a connection with the authors works leading to repeat success (*Otten, Clement & Stehr 2019*). The importance of prior success is evident from a study on the New-York Times Bestsellers List (NYTBL). The study identified that the 2,468 books on the list were from only 854 authors thereby indicating that the list is dominated by authors with multiple bestsellers (*Wang 2018*). Figure 1 (below) demonstrates this observation as a few high-profile authors such as James Patterson have many bestsellers. The influence of author power was observed to be more prevalent for Fiction authors as they commonly write in a serialised manner and hence once a following is generated, subsequent books already have an established fan base which leads to increased sales.

Relevance to analysis:

The evident influence of author power on sales means it is an important factor to capture in the analysis. Assuming the data is available, it may be useful to obtain a best-selling authors list from an index such as the NYTBL to identify authors with previous success. The brand of an author can also be assessed using sales data on their previous releases. Alternatively, the popularity of an author can be examined via other forms of data such as their social media following. These approaches will help distinguish more prolific authors from their less renowned counterparts and allow for a more accurate prediction of book sales.

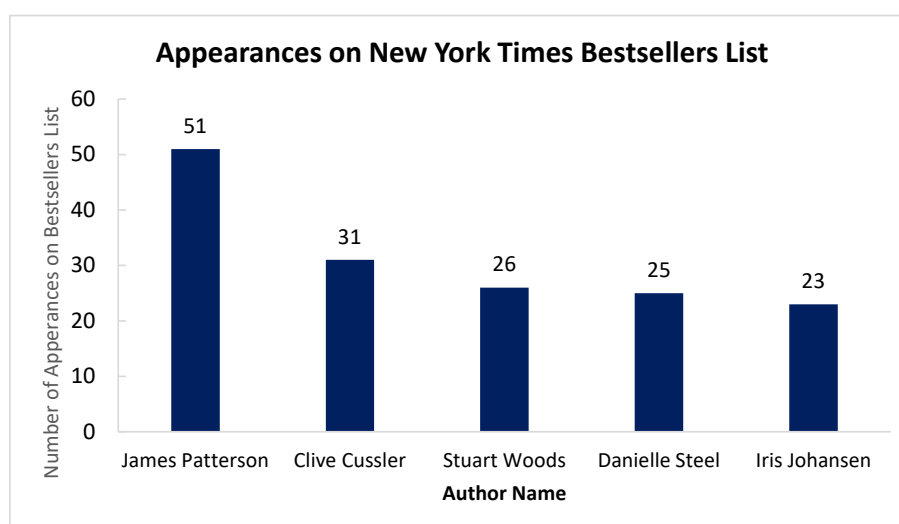


Figure 1 – represents the appearances of popular authors on the NYTBL between 2008 – 2016.

Source: Yucesoy, B., Wang, X., Huang, J. and Barabási, A.-L. (2018)



Sources for characteristic 2:

Otten, C., Clement, M. and Stehr, D. (2019). Sales estimations in the book industry – comparing management predictions with market response models in the children's book market. Journal of Media Business Studies
<https://www.bwl.uni-hamburg.de/mm/news/190612-neueraufsatz/otten-clement-stehr.pdf>

Yucesoy, B., Wang, X., Huang, J. and Barabási, A.-L. (2018). Success in books: a big data approach to bestsellers. EPJ Data Science, 7(1).
<https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0135-y>



Characteristic 3: Digital disruption from alternative book formats

Another key characteristic of the book publishing industry has been the digital disruption of publication formats and the influence that a book's format can have on sales volume. Traditionally, books have been sold in either hardcover or paperback formats. However, these formats have now been challenged by digital books in the form of ebooks and audiobooks (Meghan 2021). These digital formats have seen substantial growth with the Audio Publishers Association (APA) reporting audiobook sales grew by 35% in 2018 and 20% in 2019. The APA forecast a 20% - 25% annual growth in audiobook sales over the next decade. A study into the impact of ebook and audiobook sales identified that releasing a book via multiple formats could increase sales by up to 43.8% (Chen & Smith 2017). This indicates that the increasing digital disruption in the book publishing industry will lead to lower sales for publishers who fail to incorporate digital books within their distribution strategy.

Relevance to Analysis:

This characteristic is relevant in our analysis as sales volumes are heavily influenced by the formats that a book is released in. However, as the context focuses on a small publisher it is important to identify whether the publisher is currently incorporating digital formats in their distribution strategy. Therefore by considering the formats of the books published, we can ensure the results of the analysis align with the sales characteristics likely to be encountered by this publisher.

Furthermore, the dynamic nature of digital publishing means that assumptions made in the analysis may not be robust as the industry shifts towards a larger proportion of digital publications instead of hardcopy books. This is an important consideration as the audiobook space is expected to see strong growth in the next decade as shown in figure 2. Hence, future analysis into book sales will need to account for extra factors such as the number of digital formats.

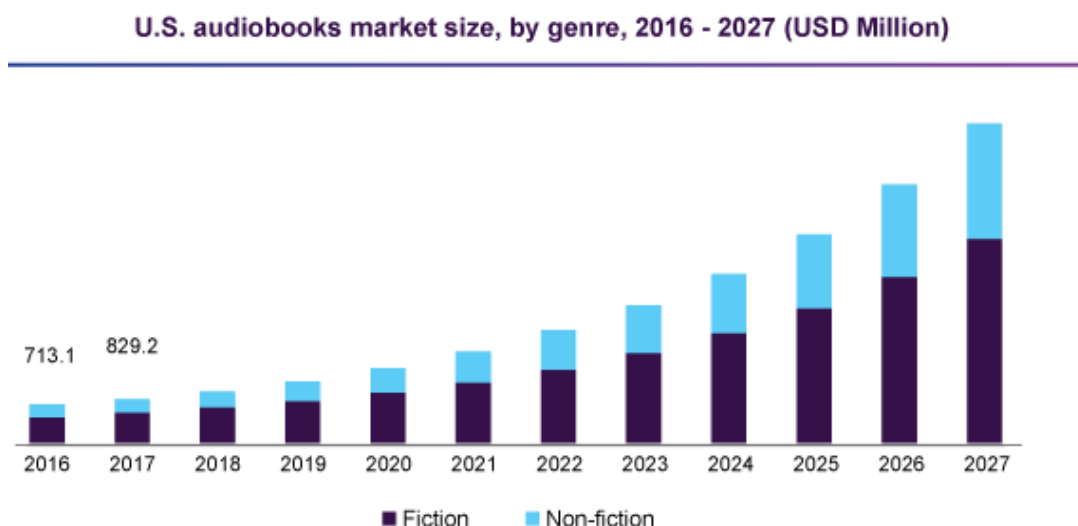


Figure 2 – represents the increasing market size of audiobooks in the US which is dominated by the Fiction genre.

Source: Grandview Research 2020.



Sources for Characteristic 3:

Meghan J. and 2021 (2020). Why Are Books Published in Hardcover First? [online] Reader's Digest. Available at: <https://www.rd.com/article/why-hardcover-books-are-published-first/> [Accessed 10 March 2022]

Chen, H and Smith, M. (2017). The Impact of Ebook Distribution on Print Sales: Analysis of a Natural Experiment. [online] Available at: <https://www.cmu.edu/entertainment-analytics/documents/technology-impact-on-entertainment/the-impact-of-ebook-distribution.pdf>.

Grandview Research. Audiobooks Market Size, Share | Industry Report, 2020-2027. [online] Available at: <https://www.grandviewresearch.com/industry-analysis/audiobooks-market>.

Dermott, K.M. Council Post: Changing Trends In The Publishing Industry. [online] Forbes. Available at: <https://www.forbes.com/sites/forbesbusinesscouncil/2021/02/11/changing-trends-in-the-publishing-industry/?sh=49e7dbcc3a5d>. [Accessed 2 March. 2022].



Characteristic 4: Specific Genres dominating the industry

Another key characteristic of the book publishing industry is the influence that a book's genre can have on sales. There are 5 primary genres which include fiction, drama, nonfiction, folklore, and poetry (*Encyclopedia of Communication and Information 2018*). Each of these genres have multiple subgenres. For example, within Fiction exists Science Fiction and Fantasy whilst within Non-Fiction exists biography subcategories. The genre can have a material impact on sales and profitability due to the demographics that certain genres target (*Kidder 2022*). This is evidenced by the top five most profitable book genres (both hardcopy and ebooks) sold on amazon during 2021 (*Smailes 2021*):

1. Romance (\$1.4 billion)
2. Crime and Mystery (\$728.2 million)
3. Religious and inspirational books (\$720 million)
4. Science Fiction and Fantasy (\$590 million)
5. Horror (\$79.6 million)

This statistic is supported by Wang (2019) who identified that Crime and Mystery alongside Horror dominated bestsellers lists with 50% of the US book sales between 2008 – 2016 falling within these genres. This demonstrates that a few genres have dominated sales across the book publishing industry over an extended period. The consistent dominance of books sales by certain genres over the past decade demonstrates that genre is an important characteristic when predicting sales.

Relevance to analysis:

The consideration of genre is important in the analysis as we are using rating counts from the Goodreads platform to determine sales. These ratings counts may be skewed towards genres which target a younger audience as these individuals are more likely to leave reviews on the Goodreads platform. Therefore, during the analysis it will be important to consider the types of genres present in the data to identify whether the rating counts appear to be influenced by genre.

Furthermore, small publishers are known to focus or specialise on a smaller range of genres compared to larger publishers. Therefore, when presenting the analysis results to the small publisher it will be important to identify whether they focus on certain genres as this would further tailor the analysis to their business.



Sources for characteristic 4:

Encyclopedia.com. (2020). Publishing Industry | Encyclopedia.com. [online] Available at: <https://www.encyclopedia.com/literature-and-arts/journalism-and-publishing/journalism-and-publishing/publishing-industry>.

Wang, X., Yucesoy, B., Varol, O., Eliassi-Rad, T. and Barabási, A.-L. (2019). Success in books: predicting book sales before publication. EPJ Data Science

Smailes, G. (2020). Which book genre is the most popular? [online] proactivewriter.com. Available at: <https://proactivewriter.com/blog/how-to-pick-a-genre-for-your-book-what-is-the-most-popular-best-selling-book-genre>.

Kidder, H.L. (2020). Target Audience for Your Book: How To & Examples. [online] Self Publishing School. Available at: <https://self-publishingschool.com/target-audience-for-your-book/>



Characteristic 5: Sales variation due to month of book release.

A key characteristic of the book publishing industry is the seasonal nature of book sales. A study on sales performance within the book publishing industry identified higher sales volume across both Fiction and Nonfiction titles between the months of September and December (Wang, Yucesoy, Varol & Barabasi 2019). This seasonal fluctuation is evident from the graph below and can be credited to higher sales during the holiday period for the purpose of gift giving or to occupy time during the holiday season. The increase in sales volume can also be attributed to the start of the United States school year which generally requires a higher proportion of Nonfiction titles thereby explaining the larger increase in Nonfiction sales exhibited in the figure below. In addition, studies show that the book publishing industry exhibits an ‘Early Peak, Slow Decay’ sales pattern meaning most sales occur within the first 10 – 25 weeks following publication (Throsby, Zwar and Morgan 2018). As a result of the seasonal sales patterns mentioned above, publishers are strongly influenced to release titles during this high sales period.

Relevance to analysis:

The goal of our analysis is to predict the sales volume of a book prior to publishing. The analysis above indicates the presence of a seasonal sales pattern in the book publishing industry and thus should be investigated in our modelling. This information would be useful to the client as it may influence their decision on when to release titles to maximise potential sales. The seasonal nature of sales can be investigated by extracting the month a book was published and then analysing the relationship between month and historical sales outcomes in the data.

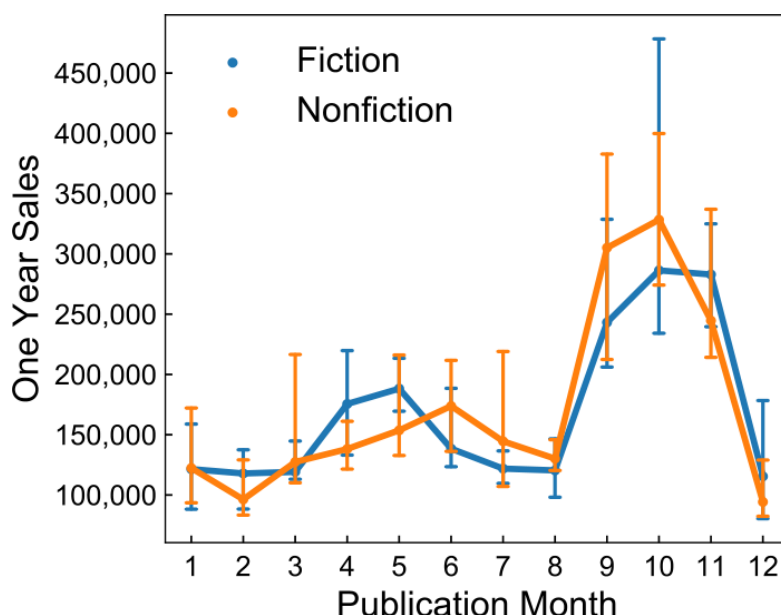


Figure 3 – represents the fluctuation in book sales across the year and highlights the need to account for seasonal fluctuations when modelling (Wang, Yucesoy, Varol & Barabasi 2019)



Sources for characteristic 5:

Wang, X., Yucesoy, B., Varol, O., Eliassi-Rad, T. and Barabási, A.-L. (2019). Success in books: predicting book sales before publication. EPJ Data Science

Throsby, D., Zwar, J. and Morgan, C. (n.d.). Australian Book Publishers in the Global Industry: Survey Method and Results Macquarie Economics Research Papers. [online]
Available at:

http://www.businessandeconomics.mq.edu.au/_data/assets/pdf_file/0004/600997/MacquarieEconomicsPublishersReport-final.pdf.



Question 6: Limitations of analysis

Please note: The limitations are targeted at an audience with basic data analytics technical knowledge as we had the option to choose to target it to the publisher or a technical audience.

Word Count: 1479 excluding headings and references

Limitation 1: Data Quality - Lack of Date Published Information

A limitation of the analysis was the data quality of certain features such as date published. The date published feature was provided in varying formats with a high proportion of datapoints not including the 'month of publication'. This impacted the analysis by preventing a 'month published' predictor being incorporated in the model. As a result, the model does not capture potential seasonal effects of book sales and likely reduces the accuracy of the model. This is a material impact as many studies engaged in predicting book sales have historically found 'publication month' to be an important factor when predicting book sales prior to publication (Wang, Yucesoy & Barabasi 2019).

Steps to Overcome Limitation 1:

To overcome this limitation in future we could obtain more accurate data relating to the publication date. This would allow us to obtain the missing published months and days. Collecting this extra information would reduce the significant imbalance of 'month published' data shown in the graph below. Therefore, allowing us to use the month a book was published as a predictor in the model.

An alternative approach to overcome this limitation is to randomly impute (fill) the missing data. However, this approach is generally preferred for trivial amounts of missing data whereas our dataset indicated a significant volume of missing data. Imputation approaches may also distort the relationship between the predictors and book sales (Gelman, A. and Hill, J. 2007). Therefore, the preferred approach would be to obtain the actual dates from online sources.



Figure 1 – demonstrates the high volume of datapoints indicated as being published in January. However, this is caused by missing entries in the data being defaulted to the 1st month. The imbalanced data caused by missing data prevented a seasonality study from being undertaken.



Limitation 2: Inability to determine Author Popularity

A limitation of the data and modelling methodology was the lack of information on an author's previous success. This limitation impacted the study as we were unable to accurately capture the authors 'brand power' as a predictor in the model. The authors brand power from previous successes is known to have a material impact in influencing sales (*Wang, Yucesoy, Varol & Barabasi 2019*). Therefore, a goal in the model building process was to include a feature that identified authors who had previously published a best-selling title as this would increase the likelihood of future titles reaching the 'High Sales' category. An alternative method was used which identified authors that appeared more than 10 times in the data. These authors were deemed as 'popular authors' but when assessing the relationship between these authors and their average book rating counts it was apparent that some authors did not have strong historical sales. Therefore, the inability to identify previous bestselling authors was likely an important factor which led to the models inability to accurately predict books in the 'High Sales' categories. This limitation is particularly important for the smaller publisher as they may want to understand the sales benefits of working with high profile authors and whether the additional money spent to engage these writers will lead to a material improvement in sales.

Steps to Overcome Limitation 2:

There are multiple steps to overcome this limitation and capture the authors 'brand power' thereby improving the models prediction accuracy amongst the 'High Sales' category. For example, 'Bookscan' presents weekly sales data which can be used to calculate the previous sales of all books written by a particular author. Thereby, identifying bestselling authors through strong historical sales. This would improve the models ability to differentiate the 'Moderate Sales' and 'High Sales' categories which was seen as an issue in the developed model with many 'High Sales' books being predicted as 'Moderate Sales'.

Wang, Yucesoy, Varol & Barabasi (2019) present another approach to allow for author visibility which was incorporated in their model to predict book sales. The proposed approach used an author's Wikipedia page as a measure of public interest in the author, thereby capturing their 'brand' power which is likely to influence sales. Implementing either of these steps would likely lead to more accurate identification of high-profile authors with an establish profile and improve model performance in predicting 'High Sales' books.



Limitation 3: Lack of Publisher Cost Information

A limitation of the analysis was the lack of information relating to the cost of publication. This limitation materially impacted the analysis as we were not able to establish the volume of sales the publisher requires to breakeven. Based on industry information it was assumed that the publisher would only want to consider profitable book proposals. Therefore, when developing the sales prediction categories an understanding into the publishers costs would allow us to more accurately differentiate between loss making and profitable books. The lack of details on publication costs forced us to employ a data driven approach alongside industry benchmarking to group books into low, moderate or high sales categories with low sales defined as up to 4,000 sales, moderate from 4,000 to 40,000 and high above 40,000. However, as this is a small publisher the costs can vary significantly and industry benchmarks are unreliable. Therefore, the established categories may differ significantly from what the publisher deems as loss making or profitable thus having a material impact on the results. For example, if the publishers actual breakeven threshold is 2,000 instead of 4,000 this would imply that all the books currently labelled as 'Low Sales' (between 2,000 and 4,000) are profitable and adjusting these sales categories can have a material impact on the models sales predictions.

Steps to Overcome Limitation 3:

This limitation can be overcome by communicating with the publisher to gain insight into their fixed costs and expected margins to identify a breakeven sales threshold. This client specific cost information would assist in accurately categorising 'Low', 'Moderate' and 'High' sales thresholds and make them more relevant to the small publisher. For example, if the 'Moderate' sales threshold is set at a breakeven point then any books predicted within the 'Low' category can be rejected by the small publisher.



Limitation 4: Arbitrary conversion between rating count & sales

Another limitation in the modelling methodology was the approach used to convert rating count to book sales. As the data did not provide sales figures, the approach involved predicting a books rating count before multiplying this figure by 4 to obtain sales volume (*Lawrence 2015*). This assumption impacted the study as it was directly used to group books into 'low', 'moderate' or 'high' sales categories. For example, books with a rating count above 1,000 were classified as moderate since the threshold was identified as 4,000. The conversion factor is genre specific as Lawrence (2015) mentions the assumption was based purely on fantasy genre books. This is problematic as a book's genre determines the age of its readers which impacts their likelihood to leave a rating. Hence, books targeting a younger demographic may have a significantly higher proportion of ratings compared to books targeting older age groups. Therefore, this assumption can lead to significantly inaccurate sales estimates across different genres. This has a material impact as the sales categories of each book will depend on the genre and its target demographics likelihood to leave a review as opposed to actual sales figures. This is particularly important for smaller publishers as their limited reach will generally result in lower rating counts compared to books from larger publishers.

Steps to Overcome Limitation 4:

This limitation can be overcome by using direct sales figures instead of relying on the conversion factor which is influenced by external factors such as genre and reader demographics. A potential solution to the problem would be to use the sales data from Bookscan which compiles a large database of sales information. Using credible data would reduce the uncertainties in the sales category labels and improve accuracy which would make it easier for the publisher to implement the model.



Limitation 5: Inaccurate approach used to group publishers by size

Another limitation was the approach used to allocate publishers between small-tier, mid-tier and large-tier groupings. From domain research, we identified the importance that publisher scale can have on sales volume. Therefore, we attempted to aggregate publishers into small-tier, mid-tier and large-tier groupings to assess the impact of scale on sales. The approach used to group publishers by size was to choose the top 35 most frequently occurring publishers and match their publishing imprints (subsidiaries) to the parent publishing organisations. Publishers not in the 'big 5' but within the top 35 were considered 'mid-tier'. Any publishers outside the top 35 were referred to as 'small-tier' publishers. The limitation of this approach is that we have assumed that any infrequently occurring publishers are 'small-tier'. The limitation is material as the goal is to predict sales for a small publisher, therefore we would want to ensure the small publisher category accurately reflects publishers with similar characteristics. The implication is that there are likely several 'mid-tier' publishers in the 'small-tier' grouping which may artificially inflate the sales outcomes of the 'small-tier' publishers and hence inflate sales predictions for books the client is considering.

Steps to overcome Limitation 5:

To improve on this limitation in future analysis, the process to categorise 'small-tier' and 'mid-tier' publishers should be defined by a more relevant figure such as the number of books published or revenue. Furthermore, the relevance of the analysis can be improved by identifying similarly sized publishers to develop a separate grouping of publishers who share common characteristics with the client.

References for Limitations:

<https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-019-0208-6#Tab1>

<http://www.stat.columbia.edu/~gelman/arm/missing.pdf>



Question 7: Video Executive Summary

Redacted to de-identify the student.