SMM635: Data Visualisation Mid-term Project

Group 18: Bonanventura William, Fung Ka, Laubenthal Christoph, Low Ming, Vijitratanakit Pakatorn

November 2019

Plot 1: Trajectory Of Reviews Of The Top 100 Books In The First 6 Months

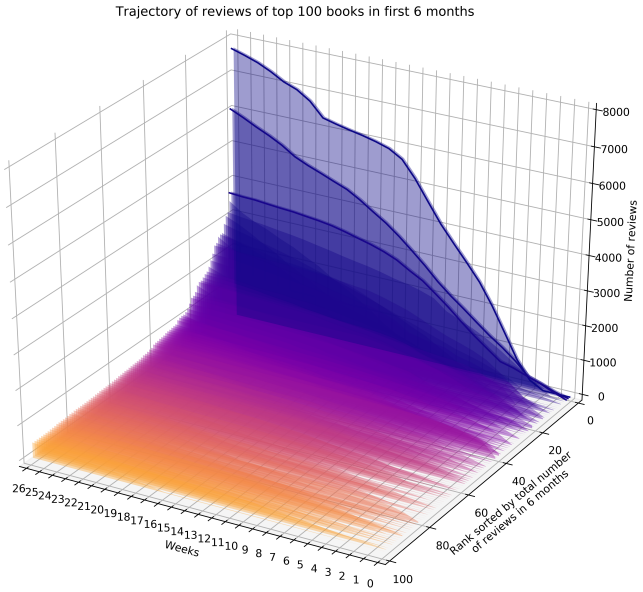

Figure 1: Plot 1

Summary and justification of design choices

The trajectory of the count of reviews for the top 100 books from the first review of that book to 26 weeks after, sorted ascendingly from books with the lowest review count to the highest review count.

The graph emphasises the nonlinear distribution of reviews when comparing the development of the most reviewed books against the rest over the timespan of half a year. This relationship holds true even when we only use the top 100 most reviewed books of the sample, as shown in the plot. The lines on top of the top three most reviewed books stress the divergence in the count of reviews even in between the very most reviewed books.

While the third most reviewed book has a total count of about 3,500 reviews six months after its first review, the respective review count of the most reviewed book yields more than twice that amount at

about 7,500 reviews. The bottom half of the top 100 most reviewed books received only up to about 500 reviews at the end of the six months.

Another insight is that even the top three most reviewed books started off more or less on par with the rest of the top 100 most reviewed books. Only after the third week do their trajectories show steeper inclines that start to differentiate them for the rest of the books. Our third analysis confirms that high total review counts of more than 5,000 reviews after six months don't require an equally strong start.

The analysis shows that it is very unlikely for a book to get reviewed more than 3,000 times in the first six months of its release. Marketing analysts working for publishers can use this information to contextualise the performance of their books on Amazon.

Dimensions of the visualisation wheel

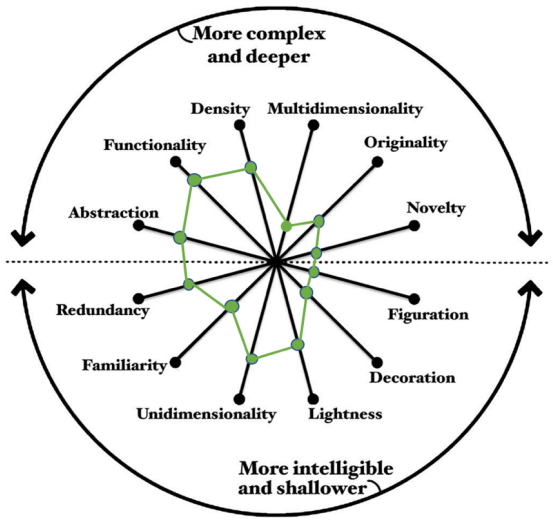

Figure 2: Dimension wheel for plot 1

Abstraction vs Figuration

The graph is more abstract than figurative because it does not include visual representations of the amount of reviews for each of the 100 most reviewed books. Since the information density of the graph is quite high, we decided to make it more abstract and thereby less visually complex to achieve a decent readability for the audience. Since the audience consists of statistically competent marketing analysts, we can expect it to have little problems deriving the core message out of the graph.

Functionality vs Decoration

The graph is rather functional and does not include a lot of decorative elements. The colour scheme in the graph is a bit redundant but at the same time, it emphasises the differences in count of reviews between different parts of the top 100 books. We therefore argue that the colour is necessary redundancy. The reason why we made the graph more functional is because its visual appeal is not our primary concern; instead, the information that is contained in the graph should be easily readable to the audience, which is interested in new insights for its marketing strategies.

Density vs Lightness

The graph is information dense since it includes three different variables: Time in weeks, ranking of the top 100 books, and count of reviews. We argue that this amount of density is appropriate to get as much insight as possible. If, for example, we had averaged the amount of reviews over six months, thereby

leaving out the ranking of the top 100 books, the insights for the audience would have been reduced to an inaccurate picture of the trajectory in the count of reviews. Again, since we present to marketing analysts, the audience will expect as many new insights as possible and therefore a high information density.

Multidimensionality vs Unidimensionality

The plot is more unidimensional than multidimensional. Even though it contains three axes, each of those axes represents only one variable. We consciously decided to not include more variables in this graph since it is already quite information dense and more variables would have contributed to a higher complexity and lower readability. The price, for instance, is brought into the analysis in our second plot.

Originality vs Familiarity

We visualised this plot to be in the middle of originality and familiarity. On the one hand, traditional visualisation elements for statistics like line graphs make the plot easily readable and functional. On the other hand, the three-dimensional visualisation is not conventional and combined with the visually pleasing colour scheme, we aimed to increase its originality. The goal was to make it as enjoyable and easy as possible for the audience to read it and understand its insights.

Novelty vs Redundancy

Due to its information density, we decided to include some redundant elements in this graph that don't add distinctly new information but increase readability and emphasise its main points. These include the colour scheme of the trajectories, the lines on top of the three most reviewed books, and the grid of the three-dimensional plot. Thus, the degree of redundancy is quite high but we argue that this is necessary redundancy that does not add complexity but helps the audience capture the main insights.

Plot 2: Price Distribution of Books across Different Categories

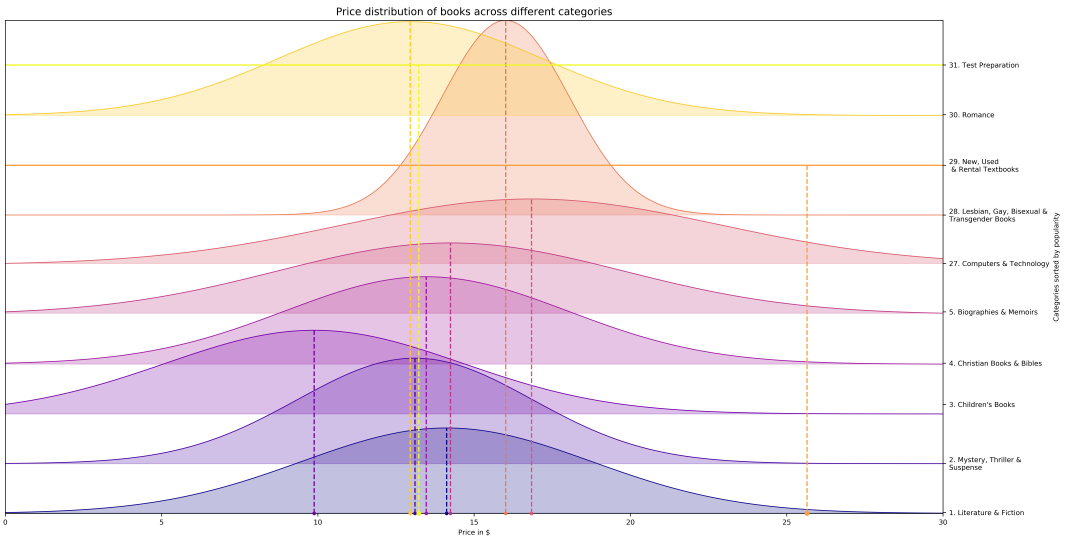

Figure 3: Plot 2

Summary and justification of design choices

This distribution graph gives insight on the most optimal pricing of popular books for the top 5 and bottom 5 categories. The measure of the optimum level of pricing is given by a distribution curve that

shows where price range reviews are mostly posted. The mean of each distribution curve is also marked on the graph to see where the highest amount of the reviews is. We specifically used the dataset for the top 10,000 books in terms of highest count of reviews for a six month period to see the price tendency of the most popular books. Then we specify the books based on the categories and chose the top five and bottom five categories to see the difference in distribution among the top and the bottom. The categories are then sorted from the most popular to the least popular on the graph.

The usage of the distribution curve rather gives us insight on how the reviews are scattered among different price ranges and see how slight increase and decrease in price for books from the mean within that category leads to a decrease in the count of reviews posted. We decided to plot separate distribution curves for each category as different categories have different audiences that have different consumer behaviours in regards to the price points of the books. Thus, we can see the variability of reviews for different categories.

By understanding this, publishers can understand what prices to set for products that they are set to release. This is critical for publishers as if they set the price too high or too low they may risk the book losing popularity.

It can be seen that the top five categories have distribution curves that show that there is a high variability in reviews. When distribution curve has high variability and deviation, it means that there is more flexibility for publishers and authors to alter the price from the mean as it leads to slower change in the drop of reviews. The books published in the LGBT category have a lower standard deviation. Therefore, the pricing flexibility appears to be more limited.

With this information marketing analysts can determine the ideal price range of books they want to publish in different categories.

Categories	**Optimal Price ($)**
Literature & Fiction	14.12
Mystery, Thriller & Suspense	13.11
Children's Books	9.88
Christian Books & Bibles	13.47
Biographies & Memoirs	14.24
Computers & Technology	16.83
LGBT Books	16.008
New, Used & Rental Textbooks	25.65
Romance	12.95
Test Preparation	13.23

The table above displays the optimal price of each book category and where most reviews are scattered among in the distribution curve. Marketing analyst could retrieve this information to deliver to publishers for the pricing strategy of the books.

Dimensions of the visualisation wheel

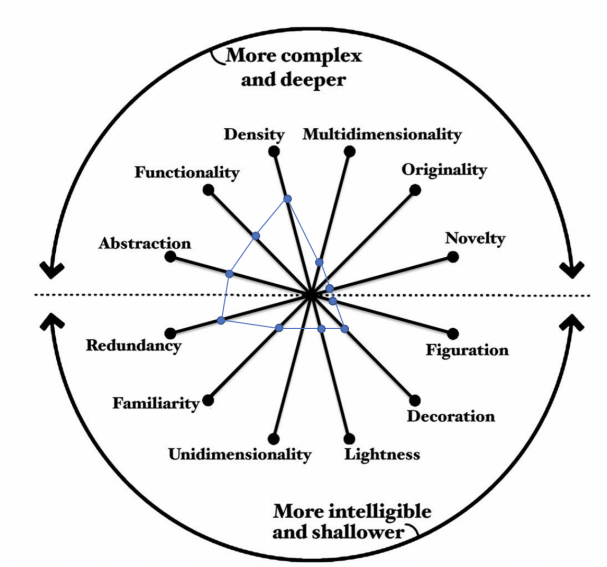

Figure 4: Dimension wheel for plot 2

Cairos Wheel

In terms of Kairos Wheel of Visualization, the graph is more abstract than figurative as it does not include visual representation of books or the different categories that could add value to the graph. The graph is direct in its attempt to portray the variability of reviews among different price range of books. To ensure this, we used a distribution curve where viewers would be familiar, thus enhancing the readability of the graph.

It has a reasonably high degree of redundancy as it explains the variability of reviews for books in different categories. It also includes extra features to the graph that would help the reader understand the graph. This includes features such as marks of the means of each distribution curve so readers can find out the exact optimal price for each category.

In terms of being in the density end of the spectrum, the graph tries to fit the optimal pricing for a range of different categories in the graph, so readers can understand the different relationship of price and ratings rather than just one curve for the whole of books. The graph is more functional as we emphasize the delivery of the variability of reviews among books across different categories. However, the graph also has some decoration values as we added colour gradients to each distribution so readers could easily differentiate different distribution curves when the curves overlap each other.

Plot 3: Distribution of product reviews across time

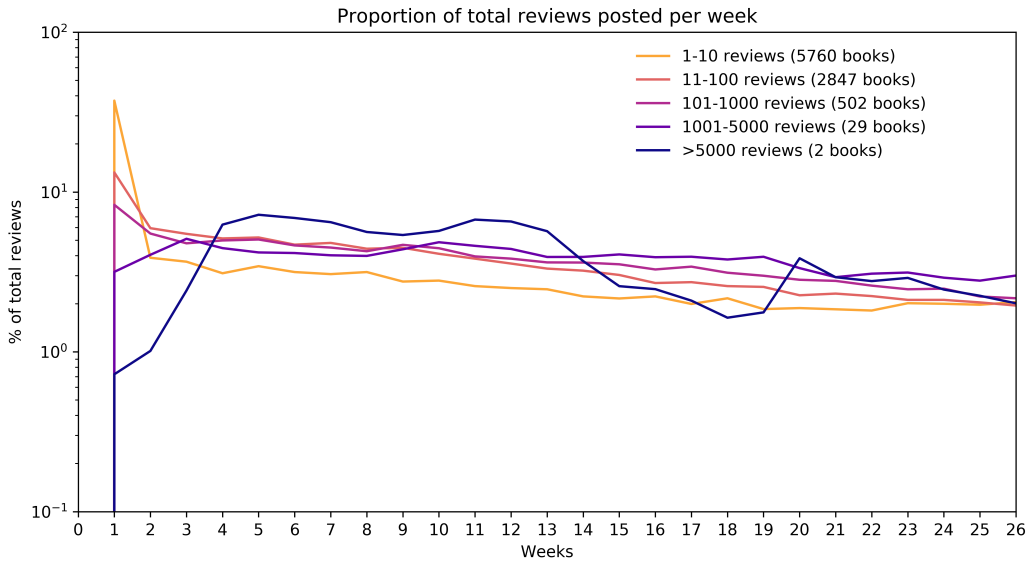

Figure 5: Plot 3

Statistical Analysis

The proportion of total reviews is taken as the number of reviews for between $t-1$ and t divided by the total number of reviews across the 26 weeks time period.

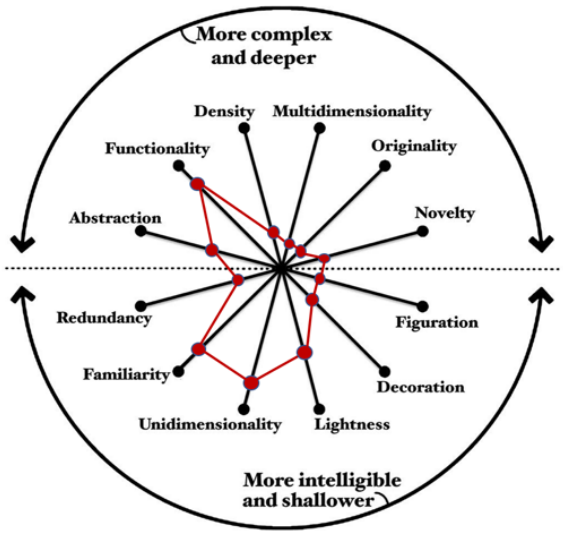

Figure 6: Dimension wheel for plot 3

Summary and justification of design choices

This graph helps understand how attention develops around new books when considering different categories of popularity. We again consider a timespan of six months from the first review until 26 weeks later. The lines represent the change of each week's amount of reviews that were received by each category of books, measured as a fraction of the category's total reviews over the six months.

Overall, the design of this plot has more of an emphasis on intelligible visuals with high familiarity, unidimensionality and lightness. There are only three kinds of variables being visualised; these being different ranges of total book reviews, the proportion of total reviews (%) and time in weeks. The graph is also highly functional, with a direct presentation of different proportions of total reviews across time through a log-scaled line graph.

This is done to illustrate clearly and simply the timings at which different proportions of total reviews for respective categories are posted, which then enables a comparison with the overall count reviews received. This would then help pinpoint the exact period within the 6 months, where most reviews are posted to then provide relevant suggestions.

Insights

From the graph, we can deduce that:
A majority of the books (around 60%) received 1-10 reviews in the first six months (26 weeks). 37.3% of their total reviews were posted in the first week with a sharp fall (-33%) from week 1 to 2 and with the weeks following week 1, the % of total reviews posted each week fluctuating between 1.8 – 3.9%.

To add, 30% of books received 11-100 reviews with ¡ 10% of the books falling into the 100-1000 and 1001 – 5000 reviews received categories. The % of total reviews across the 3 reviews received categories all experience a similar downward trend with the largest proportion of reviews received in Week 1-3 and the smallest proportion received in the last two weeks (week 25 and 26).

Although most books are off to a good start in the first week, this does not determine how well it will perform in the rest of the 6 months as the proportion of reviews decreases after the first few weeks. Different marketing efforts could therefore be introduced during the first few weeks following the book launch to ensure the sustenance of attention (which should be reflected through similar or increased levels of reviews). These efforts could include outreach to book bloggers for influencer marketing, digital marketing agencies for online advertisement and potential cross promotion with relevant parties, for example, toy companies for children's books and supermarkets for cookbooks.

Moreover, the count of reviews received in the first 3 weeks is shown to be a good indicator of the general popularity of the book. Therefore, one could identify books that fall into the 1-10 review received category in the first week and put less emphasis on those books when looking at the design and monitoring of product launch campaigns.

The ¿5000 reviews received category stands out with a different trajectory over the 6 months' period, with the smallest proportion of reviews received (0.7%) in Week 1. However, there are only two books that fall into the ¿5000 review category, both being John Grisham's fiction novels Sycamore Row and Gray Mountain. Since the sample for this category is too small for meaningful inference, more observations are needed in future studies for better interpretation of the trend.

1. The labels consists of sequential numbers.
2. The numbers starts at 1 with every call to the enumerate environment.

Plot 4: Impact of the financial crisis on performance of different book categories

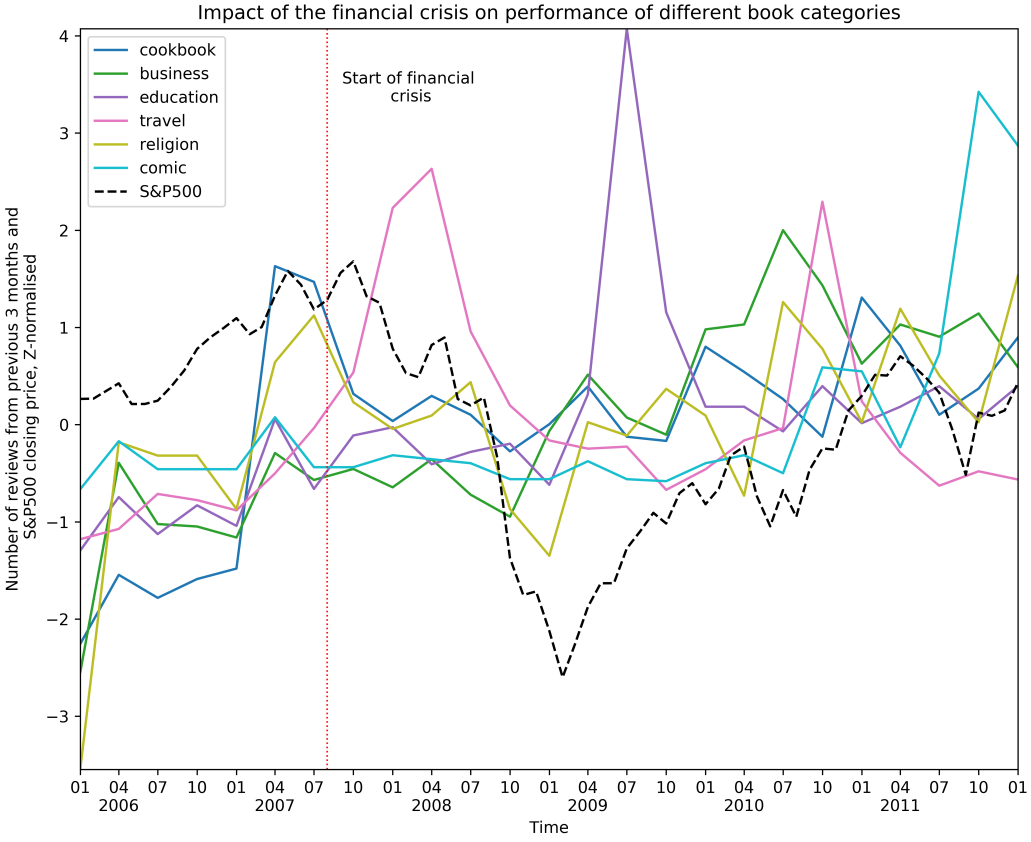

Figure 7: Plot 4

Summary and justification of design choices

The purpose of the last plot is to demonstrate how different book categories perform based on the count of reviews during an economic crisis. In the context of this plot, we have specifically used the period from 2006 to 2011 to include the financial crisis of 2007-2008 as well as preceding and succeeding years to observe how certain book categories perform in the lead up to the highest point of the SP 500 before the financial crisis and the subsequent recovery after reaching the bottom of the SP 500 during the financial crisis.

Dimensions of the visualisation wheel

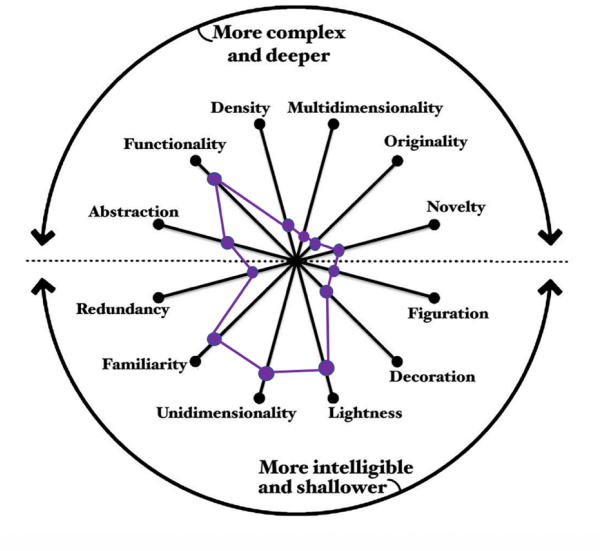

Figure 8: Dimension wheel for plot 4

With reference to the Cairo's Wheel of Visualization, the plot demonstrates more abstraction as compared to figuration as the plot does not include visual representation of books or the different categories that could add value to the plot.

The plot is highly functional as it uses a line chart to represent the different book categories and the SP 500 index. Embellishments have not been used to decorate the graph. Furthermore, this enhances the 'familiarity' attribute as it conforms to the most common visualisation patterns such as bar charts and scatter plots.

The plot also scores highly on the 'lightness' attribute as it provides an overview of information conveying few details such as the count of reviews posted and the SP 500 index against time in order to be easily understood.

Insights

For the period before drastic declines in the SP 500 (Jan 2006 to July 2008), categories such as 'cookbook', 'travel' and 'religion' achieved a peak in the number of reviews posted close to the peak of the SP 500 index. Similarly, they experienced a marked decrease in the number of reviews posted coinciding with declines in the SP 500 index during the financial crisis. Reviews posted in the 'business' category continued to display a stable, upward trajectory with the number of reviews posted increasing above its mean number of reviews posted during the entire period (2006 to 2011) for the first time when the SP 500 reached its all time low (Feb 2009). A dramatic spike in the number of reviews posted in the education category is observed in July 2009 following the SP 500 reaching its bottom. The comics category posted a consistent number of reviews during the majority of the time period with a sudden increase in reviews posted during October 2011.

With these insights, marketing analysts could hypothesis on how attention develops around specific book categories during times of economic distress so as to tailor their marketing efforts towards certain book categories based on financial market performance. Furthermore, they can also consider the possible reasons around sudden increases in the number of reviews and hence why attention develops around certain products at an instance in time.