Case Study:

# **Book Information, Publishers and best Route**

Asher Harman

## Business Task

We have considered looking at book sales and ratings, along with the book’s publishers. Based on sales information, ratings of both the books, and the degree of fame an author has, along with the publishers, one can deem it appropriate to decide what route to take when publishing one’s book.

To do that, we will look at this from the point of view of an author, using an intermediate to novice author rating to select the predicted average of the data sets for the average author. We will consider a few factors to make a definite judgment: gross sales, publisher revenue, sales price, and units sold.

There are three primary questions I want to ask and measure during this analysis.

1. How is Gross Sales affected by the number of copies sold?
2. Are the number of copies sold relevant to the book's sales price?
3. Which publisher would I suggest an author with intermediate writing ability publish with?

You will produce a report with the following deliverables:

1. A clear statement of the business task you have selected to investigate
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of your analysis
5. Supporting visualizations and critical findings
6. Based on what you discover, a list of additional deliverables would be helpful to include for further exploration.
7. Your top high-level insights based on your analysis

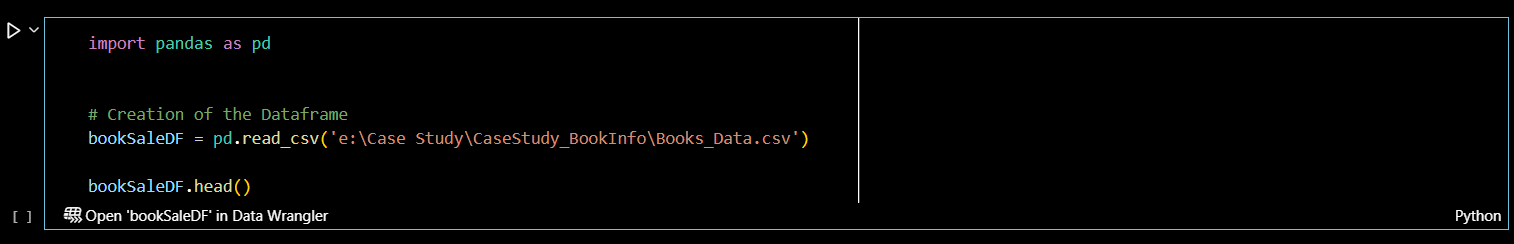
## Data Processing

Dataset <https://www.kaggle.com/datasets/thedevastator/books-sales-and-ratings>

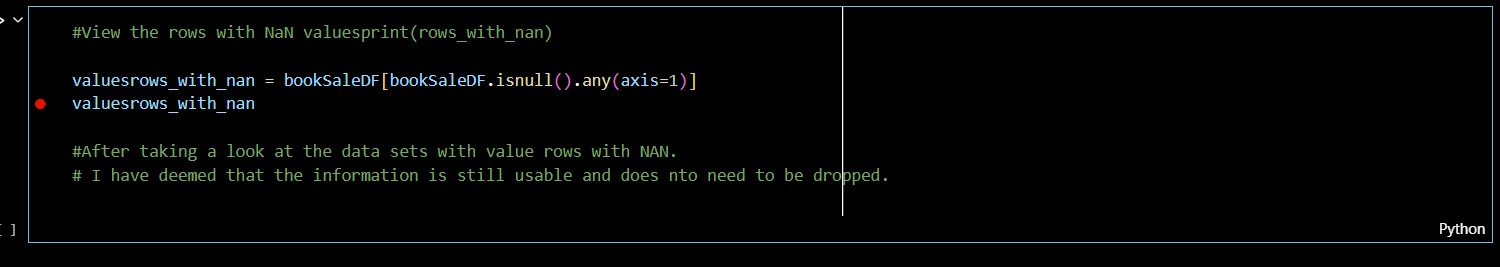
The dataset used in a data set from Kaggle consists of a previously cleaned dataset. It contains approximately 1070 books. The columns consist of index, publishing\_year, book\_name, author, language\_code, author\_rating, book\_average\_rating, book\_ratings\_count, genre, gross sales, publisher revenue, sale price, sales rank, publisher, and units sold. Josh Murrey originally compiled and cleaned the dataset.

In VStudio Code, I created a Juytper Notebook that runs and edits the data set using an instance of Python.

The first step was to import the Pandas, and then the CSV file with the .head() was used to ensure the dataset was imported correctly.

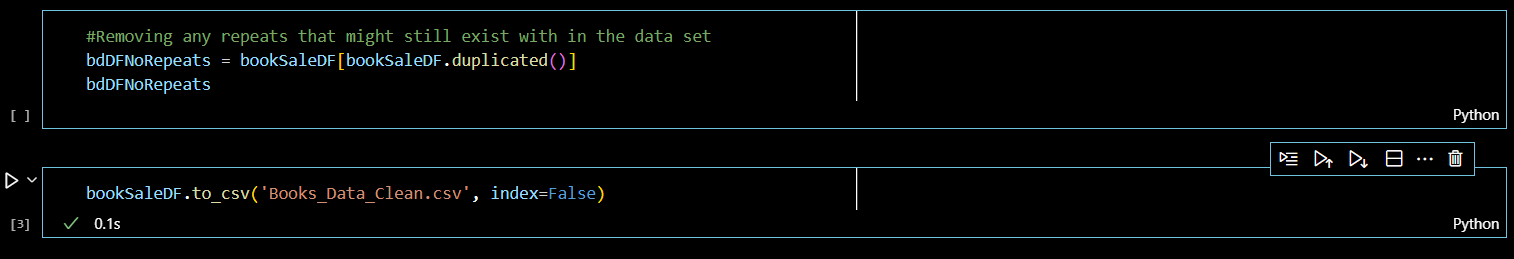


We then want to double-check out all the rows with nothing inside. When we have done so, we have viewed them all. The empty cells were those within names, book Names, Authors, and language\_code. When looking at these, we believed that the empty cells would not affect the analysis, so we did not remove any rows.



The next step was to verify that there were no duplicate rows in the data sets. A quick search guaranteed that there were no duplicates in the initial dataset.

We exported the dataset to CSV to enhance its versatility and empower you to work with it in your preferred tools. This gives you the freedom to analyze the data as you see fit.



We took the Book\_Data\_Clean dataset into Google Sheets work and standardized the datasets. We followed this by standardizing the names, lowercase all information, and replacing all spaces with underscores. Our following standardization is to change the gross\_sales, publisher\_revenue, and sales\_price to a dollar amount to prevent any strange rounding errors that could happen. Another standardization we did in the Google Sheets suite is standardizing the Publishers, specifically HarperCollins Publishing, into correct and more widely used HarperCollins Publishers.

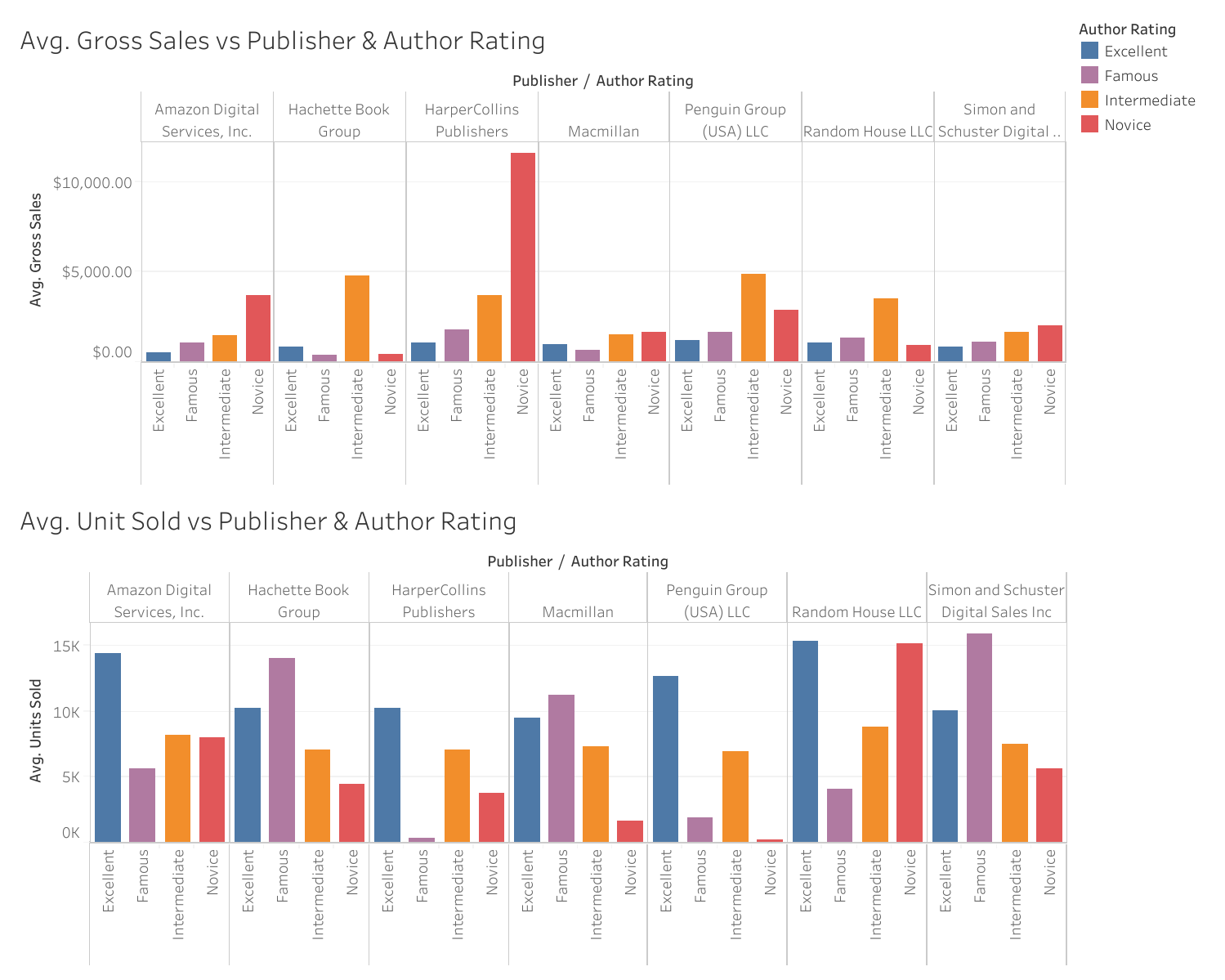
## Analysis

To get a straightforward look at some vital statistics from the spreadsheet. I created a pivot table sorted by the Publishers. I took the average of all ratings, gross sales, publisher revenue, and average sale price, along with how many books each publisher published.

|  | AVG book rating | AVG gross sales | AVG publisher revenue | AVG sale price | COUNT |
| --- | --- | --- | --- | --- | --- |
| Amazon Digital Services, Inc. | 4.00 | $1,216.35 | $247.07 | $3.23 | 600 |
| Hachette Book Group | 3.96 | $3,481.68 | $2,089.01 | $6.36 | 66 |
| HarperCollins Publishers | 4.04 | $2,815.29 | $1,689.17 | $6.20 | 79 |
| Macmillan | 4.00 | $1,315.29 | $789.18 | $5.40 | 41 |
| Penguin Group (USA) LLC | 4.05 | $3,299.65 | $1,979.79 | $8.65 | 108 |
| Random House LLC | 4.00 | $2,633.13 | $1,579.88 | $7.34 | 120 |
| Simon and Schuster Digital Sales Inc | 4.04 | $1,398.46 | $839.08 | $5.87 | 56 |
| **Grand Total** | **4.01** | **$1,856.62** | **$843.28** | **$4.87** | **1070** |

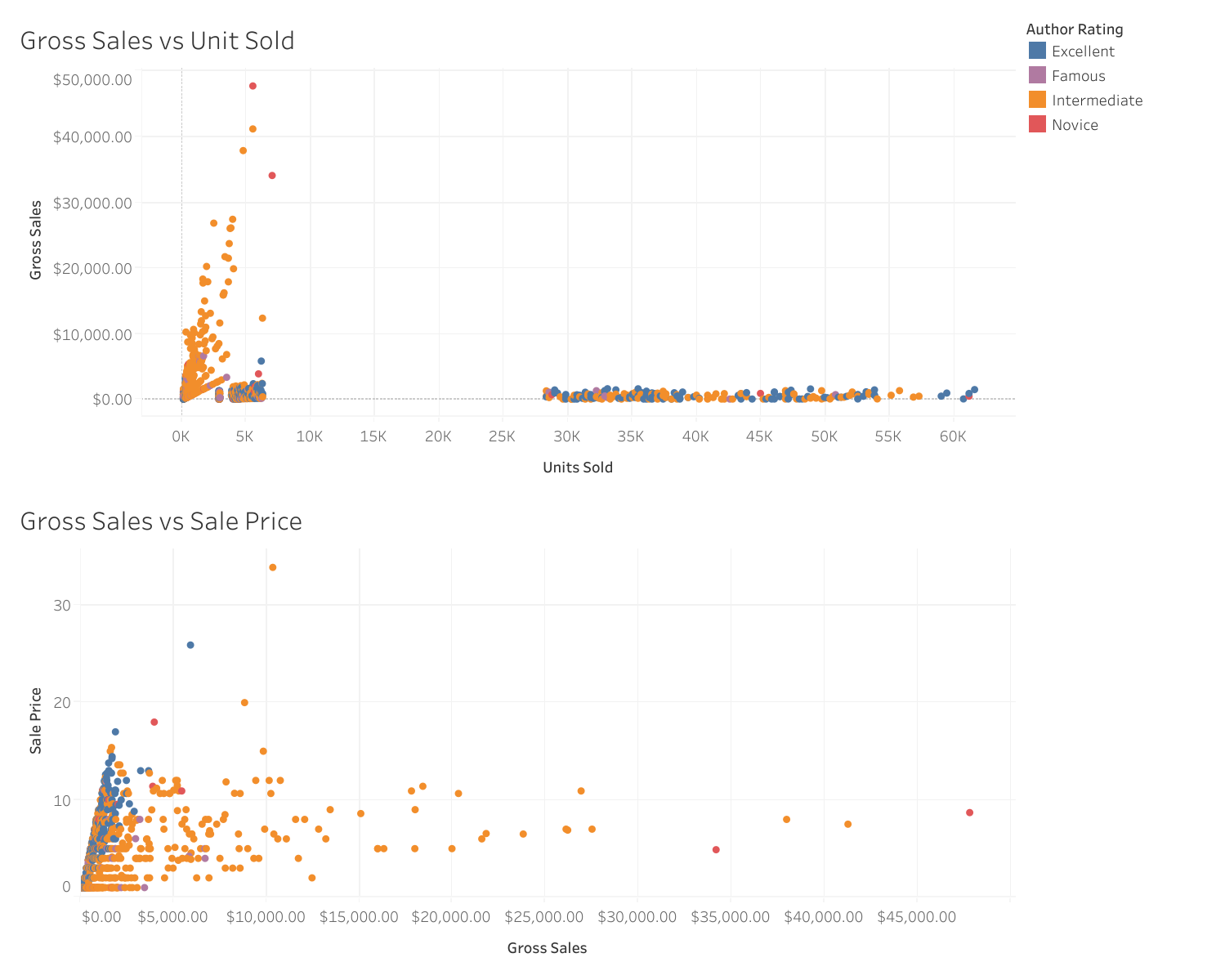
The first noticeable feature is that all the average book ratings are about the same. A rating of ~4.00 implies the average quality is about the same for each publisher. This could be from the impact of averaging everything that shows different quality of work after already averaging the rating of each book from multiple other sources.

The next feature to examine is the AVG gross sales, which creates an exciting observation when comparing it to AVG publisher revenue and COUNT. Immediately, it became clear that since Amazon Digital Services has the most significant number of books we have measured, this heavily reduced their average gross sales and average publisher revenues. This implies that there is quite a lot of content on Amazon that might need to be more profitable, but they are willing to host since other books they publish make up for the lack of income to a greater degree. This is obvious, with Amazon having ~56% of the total books. While not explicit, Amazon Digital services most likely put less work into each work, allowing authors to publish under their licenses than the traditional publishing methods of bringing in editors and making the book more presentable compared to other publishers. This also accounts for the relatively minor revenue that Amazon Digital Services takes in comparison to the rest of the publishers.



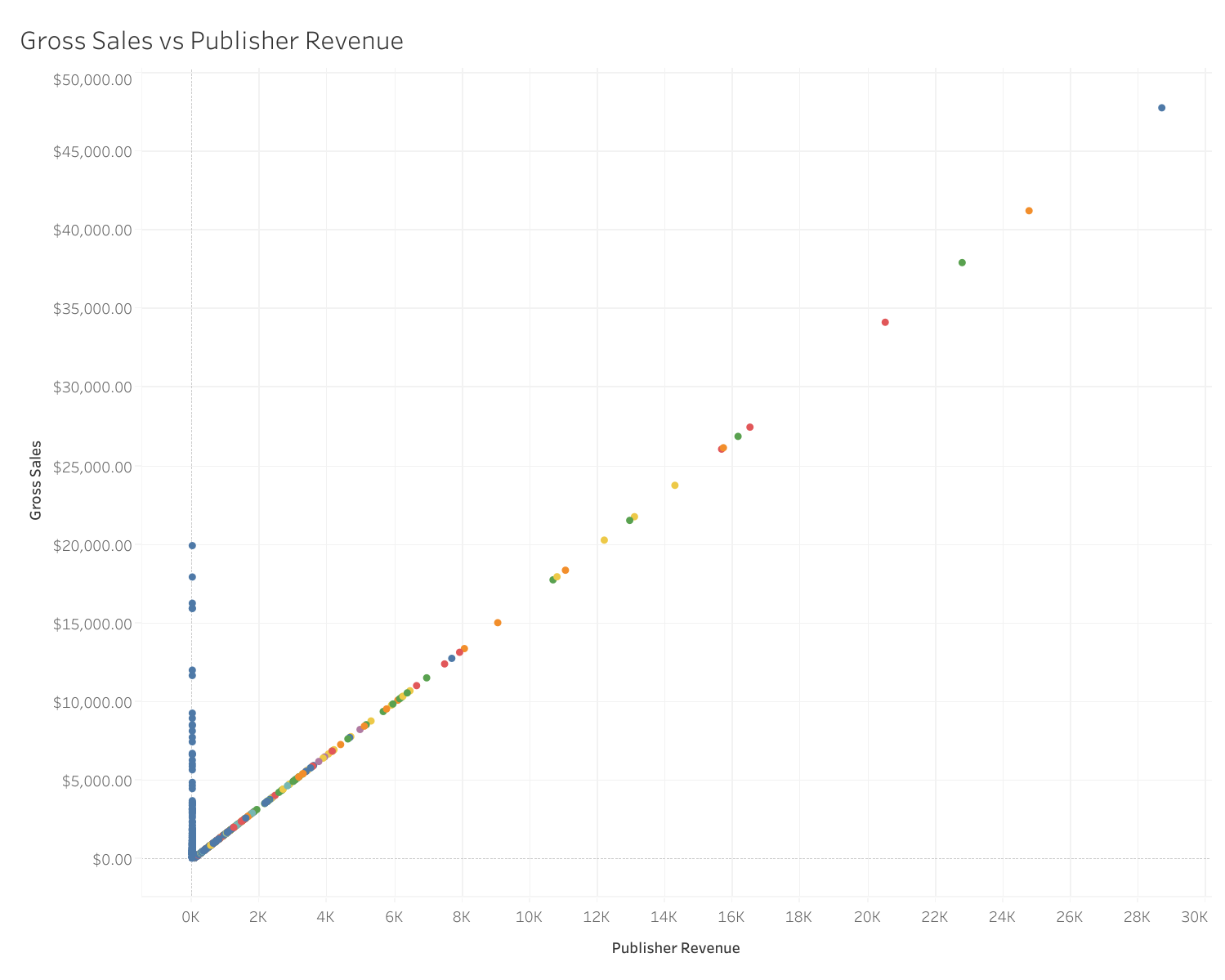
I am using the Avg. Gross Sales vs. Publishers & Author Rating and Avg. Units Sold vs. Publishers & Author Rating graphs, we can gain insights from the immediate data. Immediately, the highest average sales are for novice-level books from HarperCollins Publishers. The rest of the averages barely topped out at $5,000, though this could be due to the amount and need for more information. So, we will use the average units Sold to understand better which publisher supports authors of different talents.

My first insight is that, as expected, the Excellent and Famous Author levels are generally consistently high regarding the number of units told. This is not a strict rule, as Random House LLC appears to support Noice's talents quite a lot, as they match their famous category. For Intermediate Authors, they are all about the same, so there can not be an actual suggestion based on the table given. However, as a novice, going to HarperCollins Publishers or Random House LLC might be best.



Vital information regarding the number of units sold for the Gross Sales vs. Units Sold must be included, yet there are no Gross Sales. Though it makes sense, a quick scan underlines that most publishers are of intermediate skill level or fame. Interestingly, a novice created the top gross sales level author yet still gained wide popularity. However, due to the lack of conclusive information, there does appear to be a trend of more units sold higher than the gross sales. While expected, there are still enough questions that we need to do further research into.

It's important to emphasize that there is no clear correlation between book sale prices and gross sales. While there may appear to be a linear relationship initially, the data does not support this. The book market is complex, and a book's price does not necessarily dictate its sales. For instance, a book priced at approximately 10 dollars or less may sell more, depending on the situation and the book type.



The Gross Sales vs. Publisher Revenue, while expected, there is a general linear relation between the gross sales and the amount of revenue the publishers make. The most exciting piece of information is how Amazon Digital Services looks to have various options. Wither taking no money from sales.

## Recommendations

1. For further research, finding the specific cuts publishers request along with potential services they provide for the cut could help explain in greater detail why or how gross sales and publisher revenue are related.
2. Though most of the books published this way were fiction/ genre fiction, I would like to look further to see what genres of fiction perform the best and if there is a general trend related to genre and what the common user expects.
3. When it comes to publishing, this is a complicated and detailed subject. If you are an intermediate or higher level author, Amazon Digital Services is the best choice of publisher. This is due to flexible options, and they allow a more significant number of individuals to post their work while also taking a minimal amount of revenue from sales. Suppose you are a Novice Author, on average. In that case, you have a better chance of getting discovered and making money from HarperCollins Publishers or Random House LLC if you value discovery and sales more on average.