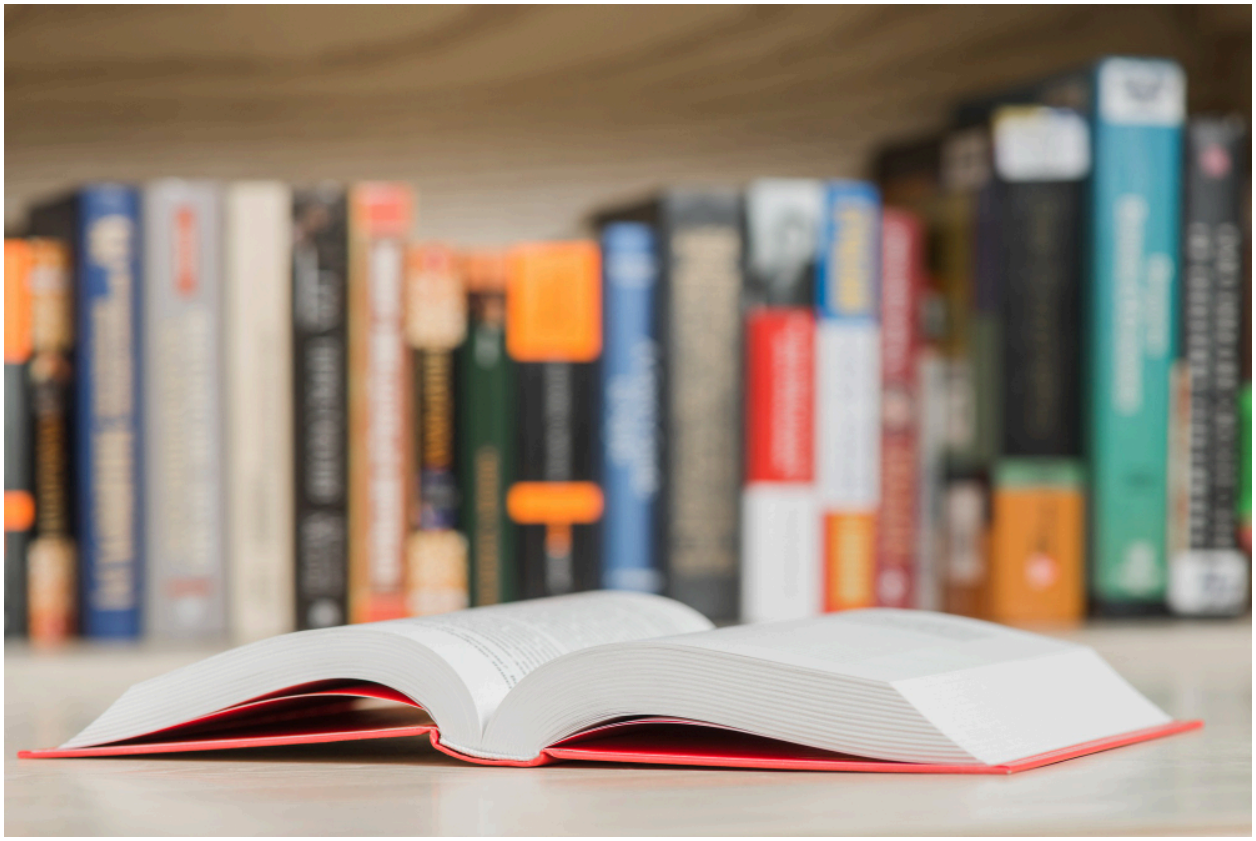
DTSA 5304 Final Project

Best Selling Books Dataset



Gabrielle Charlton

April 30th, 2024

DTSA 5304 - Fundamentals of Data visualization

Final Project

A brief recap of your data, goals, and tasks, focusing on those that most directly influence your design

---

The dataset I will be using is the **best-selling-books** dataset made by D Rahulsingh (Kaggle Source: *https://www.kaggle.com/datasets/drahulsingh/best-selling-books*).

This dataset holds data regarding best selling books. The main columns of this dataset I will use include:

- Book
- Author(s)
- Original language
- First published
- Approximate sales in millions
- Genre

For this exercise, I'm imagining that I am at a point in my life when I am considering a drastic career change. A career I've always wanted was to open a bookstore. In preparation for opening my own book store, I am having a hard time finding classics for my books inventory. The main goal for me is to know what books are most guaranteed for sales and success. I would like to know:

What kinds of books are the most popular?

- What genre typically sells the most?
- Which author typically has sold the most ?
- What original language is the most popular for books sold?

What are the trends with popular book sales over the years?

- What has been the most popular genre over the years?
- Who has been the most popular author over the years?
- What has been the most popular original language over the years?

Problem: Wanting to streamline my focus on building inventory that will be successful

Task 1: What kinds of books are the most popular?

What genre typically sells the most? Which author typically has sold the most ? What original language is the most popular for books sold?

- **Goal**: Knowing what aspects of books sold that relates to popularity, makes it easier to understand what kind of inventory to lean towards for better success of the bookstore.
- **Means**: Looking at datasets for market analysis of books sold over time, and their overall popularity.
- **Characteristics**: Compare count and overall sales to the various aspects to learn about the popularity for each.
- **Target Data**:
    - Book
    - Author(s)
    - Original language
    - First published
    - Approximate sales in millions
    - Genre
- **Workflow**: This task is performed first to understand what the most popular books, authors and genres are
- **Roles**: Book Merchants, Librarians, Authors

Task 2: What are the trends with popular book sales over the years?

What has been the most popular genre over the years? Who has been the most popular author over the years? What has been the most popular original language over the years?

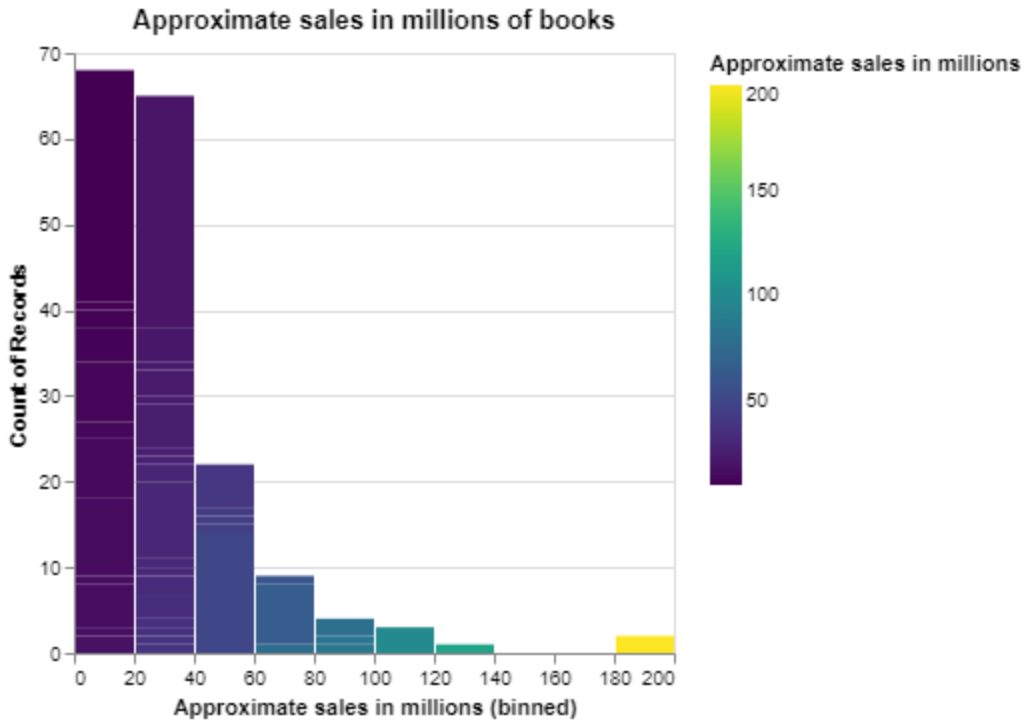
- **Goal**: Knowing what trends have been made over the years in the book industry can convey how the bookstore owner can strategize inventory building for greatest success
- **Means**: Looking at how the most popular authors, and genres relate to the overall sale of books over time via datasets and visualizations
- **Characteristics**: Compare authors' and genres' count to the overall sales and time eclipsed to learn about the popularity for each.
- **Target Data**:
    - Book
    - Author(s)

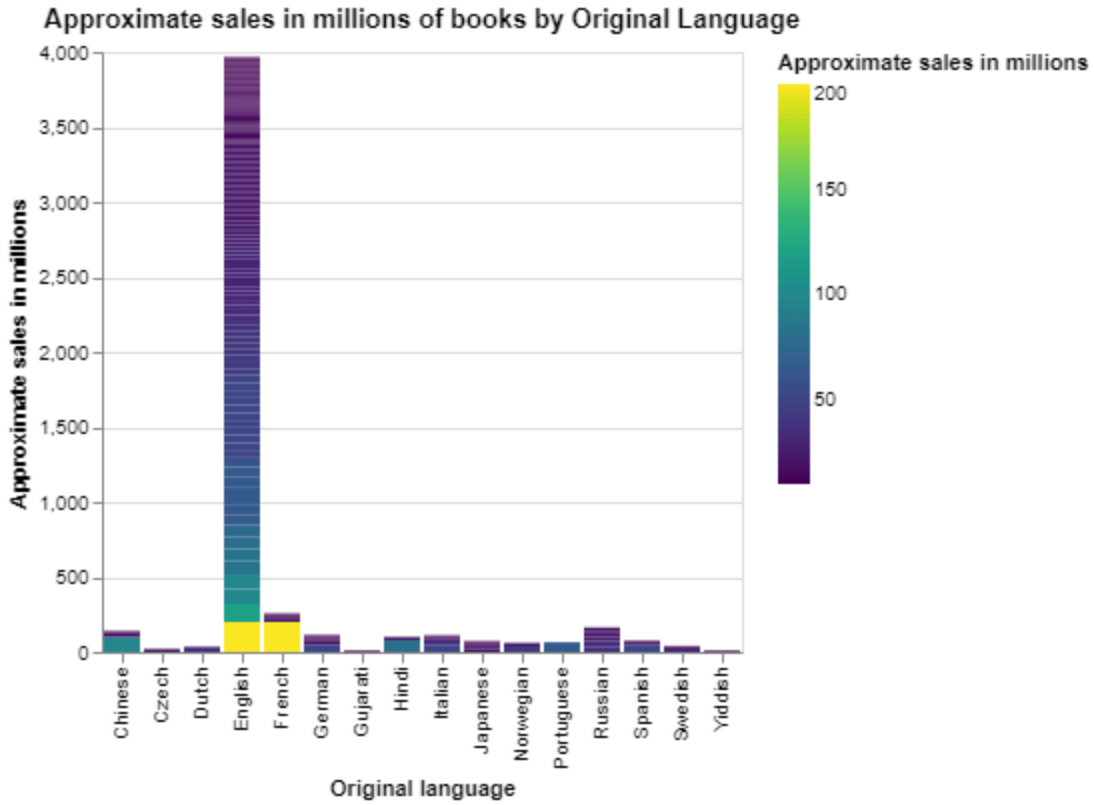
    - Original language
    - First published
    - Approximate sales in millions
    - Genre
- **Workflow**: This task is performed second to get a more indepth look at what has been happening in the industry and what could come from it in the future
- **Roles**: Book Merchants, Librarians, Authors

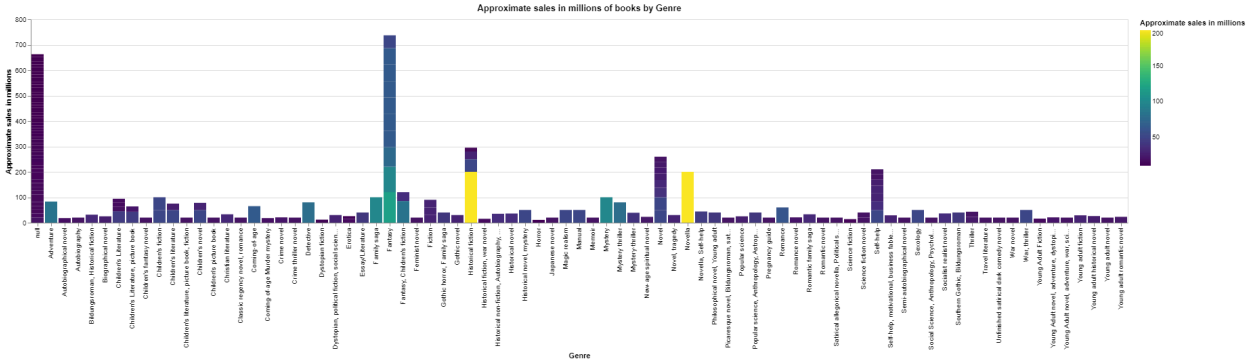
Screenshots of and/or a link to your visualization implementation (see below for additional guidance)

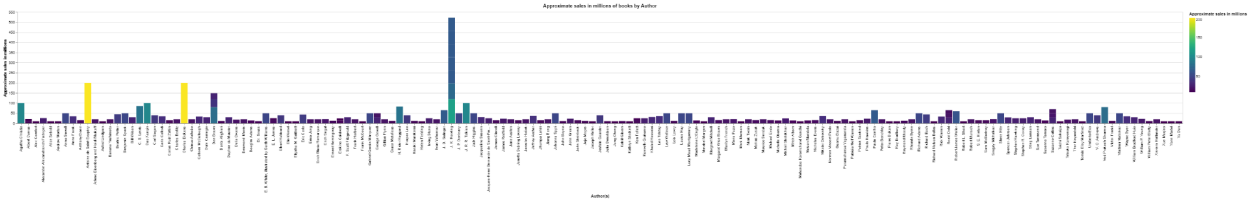
---

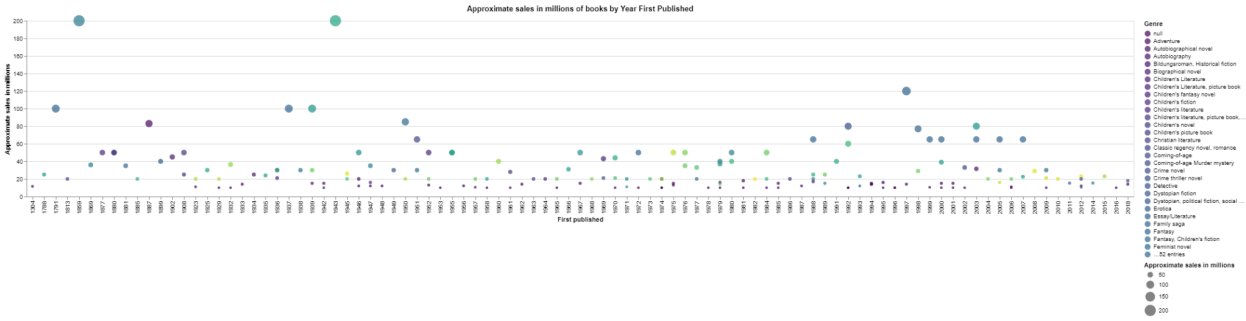
https://www.kaggle.com/code/gabriellecharlton/fundamentals-of-data-vis-final-project




Approximate sales in millions of books by Original Language
Approximate sales in millions
200
150
100
50
Approximate sales in millions
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Chinese
Czech
Dutch
English
French
German
Gujarati
Hindi
Italian
Japanese
Norwegian
Portuguese
Russian
Spanish
Swedish
Yiddish
Original language


Approximate sales in millions of books by Genre
Approximate sales in millions
Genre


Approximate sales in millions of books by Author



A summary of the key elements of your design and accompanying justification

---

I decided to go with stacked bar charts for some of my visualizations because I wanted to get a really good overview of the total amount of sales per category, rather than using a truncated approach (i.e. count(*insert category*)). As for my longer visualizations, they are a bit cumbersome when shrunken down, but I worked hard on assigning the appropriate size and color attributes to best showcase the differences between the entities. I used only Aletair for my visualizations and I feel proud that I was able to get as much insight using it.

A discussion of your final evaluation approach, including the procedure, people recruited, and results. Note that, due to the difficulty of recruiting experts, you can use colleagues, friends, classmates, or family to evaluate your designs if experts or others from your target population are unavailable.

My final evaluation approach was using the Think Aloud strategy. I had a family member go through my visualization and I asked them how easy it was to use, what can be improved, and how easy they were able to answer any of the questions. I found that my visualizations were researched positively, yet had a learning curve for my participant. The Learning curve was me having to point out the legend and reiterate the question so that they could answer effectively.

This was very useful as it helped me understand how ease of use with visualizations and how clear your tools are leads to less questions users have to ask in order to reach their

conclusion. In the future, I think I will try to implement animated visualizations so that it can be more fun and interesting to users. The tooltip I used was received very positively however, so having that I think did help.

A synthesis of your findings, including what elements of your approach worked well and what elements you would refine in future iterations.

My findings from my analysis and visualizations are that:

1. The highest earning best selling books are typically individual books that were originally published many years ago. Where as, the current highest earnings books are mostly apart of a series (i.e. J. K. Rowling's Harry Potter series). For this book store merchant, investing in book series may be a great avenue to tap into a greater fan base and overall more loyal customers.
2. The most popular original language for books seems to be English from this dataset, but the next few popular languages are: French, Chinese, and Hindi. Probably investing in international language books or even English translated popular books could be useful for this book merchant.
3. The most popular series seems to be Fantasy, based on approximate sales in the millions.
    a. I believe this is skewed due to the Harry Potter series, however, that doesn't take away the cultural impact and shift in fantasy literature over the years from my personal observations. I would improve this by possibly looking at a larger dataset for a wider range of books and diving into the future trends of fantasy literature.

My think aloud approach was very successful and it helped me find small flaws in my work and ways to improve in the future. I think segmenting the top 10-20 against the bottom 10-20 books in this dataset would be a good approach next time, as it would help denote what actually happened in the book industry over time. It was easy to find the top of the dataset, but if I had analyzed the lower values, maybe I would've found more nuance and subtle shifts in how literature is created and consumed.

Module 1

Locate a dataset that you are interested in working with. The data should be sufficiently complex that you can ask lots of questions about it and engage in creative design techniques, but not so complex that you need specialized hardware or algorithmic approaches to analyze. While you are welcome to use any data you'd like, I recommend that your datasets are tabular (e.g., CSV, TSV, SQL, etc.), contain 5,000 or fewer datapoints (on the order of one hundred or so tends to be sufficiently interesting without causing lag in Altair), and is data that you're comfortable discussing as part of the course (e.g., avoid data that is overly private or classified). Discuss your dataset, including the data's source, key attributes/dimensions of the data, and your goals for working with that data (i.e., what are the key questions you want to answer). Identify existing relevant visualizations for working with that data (either using the same data, showing the same concepts, or just that might provide some inspiration) and critique those visualizations based on the practices from this module. What works well? What might need improvement or to change to answer your target questions?

The dataset I will be using is the **best-selling-books** dataset made by D Rahulsingh (Kaggle Source: *https://www.kaggle.com/datasets/drahulsingh/best-selling-books*).

This dataset holds data regarding best selling books. The main columns of this dataset I will use include:

- Book
- Author(s)
- Original language
- First published
- Approximate sales in millions
- Genre

For this exercise, I'm imagining that I am at a point in my life when I am considering a drastic career change. A career I've always wanted was to open a bookstore. In preparation for opening my own book store, I am having a hard time finding classics for my books inventory. The main goal for me is to know what books are most guaranteed for sales and success. I would like to know:

What kinds of books are the most popular?

1. What genre typically sells the most?
2. Which author typically has sold the most ?

3. What original language is the most popular for books sold?

What are the trends with popular book sales over the years?

4. What has been the most popular genre over the years?
5. Who has been the most popular author over the years?
6. What has been the most popular original language over the years?

A great example of visualizations are in a notebook made on Kaggle based on this dataset that focused on the best selling books dataset: https://www.kaggle.com/code/dhruvch121/data-analysis-of-best-selling-books . There they make very good analysis of what we can find in the dataset. I would only critique that the questions aren't directly focused on answering any specific questions, but that can also be a virtue as it has sparked some inspiration for me to go beyond what they have done. In terms of visualizations, I would say they had a great use of charts, but some visualizations were hard to understand and gain insight. I think I want to expand on these and really hone in on what my questions are asking of me.

Module 2

Your Module 1 discussion post identified some high-level goals for working with a dataset of interest to you. In this post, you will expand on those goals to characterize your target problem and develop some low-fidelity prototypes for working with that data. First, identify two to three tasks you would wish to complete with your data, identifying:

- *Why is a task pursued? (goal)*
- *How is a task conducted? (means)*
- *What does a task seek to learn about the data? (characteristics)*
- *Where does the task operate? (target data)*
- *When is the task performed? (workflow)*
- *Who is executing the task? (roles)*

Then, sketch a set of preliminary low-fidelity prototypes for addressing these tasks with the given data. You may either sketch freeform or use the Five Design Sheets approach to generate these prototypes (hand-sketched on paper is fine). Upload a copy of your sketches as part of your post.

For this exercise, I'm imagining that I am at a point in my life when I am considering a drastic career change. A career I've always wanted was to open a bookstore. In preparation for opening my own book store, I am having a hard time finding classics for my books inventory. The main goal for me is to know what books are most guaranteed for sales and success.

Problem: Wanting to streamline my focus on building inventory that will be successful

Task 1: What kinds of books are the most popular?

What genre typically sells the most? Which author typically has sold the most ? What original language is the most popular for books sold?

- **Goal**: Knowing what aspects of books sold that relates to popularity, makes it easier to understand what kind of inventory to lean towards for better success of the bookstore.
- **Means**: Looking at datasets for market analysis of books sold over time, and their overall popularity.
- **Characteristics**: Compare count and overall sales to the various aspects to learn

about the popularity for each.

- **Target Data**:
    - Book
    - Author(s)
    - Original language
    - First published
    - Approximate sales in millions
    - Genre
- **Workflow**: This task is performed first to understand what the most popular books, authors and genres are
- **Roles**: Book Merchants, Librarians, Authors

Task 2: What are the trends with popular book sales over the years?

What has been the most popular genre over the years? Who has been the most popular author over the years? What has been the most popular original language over the years?

- **Goal**: Knowing what trends have been made over the years in the book industry can convey how the bookstore owner can strategize inventory building for greatest success
- **Means**: Looking at how the most popular authors, and genres relate to the overall sale of books over time via datasets and visualizations
- **Characteristics**: Compare authors' and genres' count to the overall sales and time eclipsed to learn about the popularity for each.
- **Target Data**:
    - Book
    - Author(s)
    - Original language
    - First published
    - Approximate sales in millions
    - Genre
- **Workflow**: This task is performed second to get a more indepth look at what has been happening in the industry and what could come from it in the future
- **Roles**: Book Merchants, Librarians, Authors

Module 3

The target question you want to answer:

What kinds of books are the most popular?

- What genre typically sells the most?
- Which author typically has sold the most ?
- What original language is the most popular for books sold?

What are the trends with popular book sales over the years?

- What has been the most popular genre over the years?
- Who has been the most popular author over the years?
- What has been the most popular original language over the years?

The people you would recruit to answer that question:
The people I would recruit to answer this question would be people who are interested in learning about the book industry either professionally or recreationally. The best people for insight would be professionals such as authors, book merchants and librarians, since they can hopefully give more nuance

The kinds of measures you would use to answer your data (e.g., insight depth, use cases, accuracy) and what these measures would tell you about the core question:

Since this dataset is both numerical and categorical, and deals with trends, what I would do is use a qualitative evaluation approach with systematic surveys to gain user insight on what has been observed in the visualization and how well it responds to personal/ professional experiences.

The approach you will use to answer that question (e.g., a journaling study, a formal experiment, etc.):

I would use either a systematic survey (as stated above) or Think Alouds so see if my tool is useful to users.

How you would instantiate those methods (i.e., what would your participants do?):
What I would do is show the visualization tool, and ask the users what they thought of it

as they go through. If they found anything that was interesting or not to their expectations, I would note it and add it to consideration for improving my tool. This is why think-alouds would be the most useful.

What criteria would you use to indicate that your visualization was successful:
If the users can understand the visualizations to answer the two questions I laid out effectively.