HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

PROJECT REPORT

BEST-SELLING BOOKS ANALYSIS

Group 2

Nguyen Duy Hung - 20194436

Dang Thanh Lam - 20194442

Le Hai Son – 20194449

Nguyen Hoang Nhat Quang - 20194448

Table of Contents

I. Introduction

Books play an important role in our life, introducing us to new knowledge, broadening our horizons as well as giving perspective to the world around you. A good book has the power to change the way we think, talk and analyze things. It is said that around 2.2 million books are published every year, that's such a huge number that somehow emphasizes the importance of books and success of book industry. As book lovers, we do this project to discover the variety of books nowadays based on data we scrape from two big e-commerce site: tiki and amazon. Our data has some book attributes such as category, rating, price, book sales order and we will analyze these attributes to find information about book buyers' habits, the correlation among these attributes and what factors make a book popular and best-seller.

II. Data Scraping

What is *Data scraping*: *Data scraping* is a technique extracts data from human-readable output coming from another program. Everything we can see on the screen, we can scrape it

We choose 2 e-commerce sites for data scraping: Amazon and Tiki best-sellers. They are the most reputation e-commerce website and they have the same original: they are both selling book when establish. Amazon for the international user, Tiki for Vietnamese user(mostly)

With the using of scraping tool: Scrapy. Scrapy is a Python framework for large scale web scraping. It gives us the tools need to efficiently extract data from websites, process them as we want, and store them in structure and format.

Pros:

- Fast, simple, extensible
- Scrapy provides an asynchronous mechanism which processes multiple requests in parallel.

Cons:

- Difficult when scraping pages generated via JavaScript (solved by using scrapy-splash or scrapy-selenium)
- Installation is different for different operating systems, Scrapy is only for Python 2.7. +

Some difficult when scraping in our project:

- Blocked because of sending too many requests: Amazon is strict with scraping data. When using the default setting, our bot got banned after each 100-200 requests. To bypass this, we first slow down the scraping process by adding a small delay between page loads, then disable the cookie. We rotate the user agents to make it looks like the requests are made by different device using the same IP address. Finally, we change the IP address manually by connecting to different networks. After multiple attempts, we successfully collected the detail information from all 2690 books
- Data selectors when extracted from the browser is different when using selector query of scrapy to queries these selectors: Use scrapy shell to check before setting selectors of the spiders, delete cookies of the desired sites before queries.

Result: 2000 items (Tiki), 2690 items (Amazon)

(*) Data description: Tiki: top-100-best-seller book over year from 2011-2020, Amazon: top-100-best-seller book over year from 1995-2021.

Data attribute:

9 variables: 4 categorical – 4 numerical (only the *Year* attribute can be treated as categorical or numerical depends on our purpose)

- Category: the category that set to the book (when user want to sell this book). For tiki, we can scrape by the category so, we have the current_category attribute, where the book on. *(categorical)*
- Year: the year of book bestseller *(categorical or numerical)*
- Order: rank of the book when on a particular category or particular year *(numerical or ordinal categorical)*
- Name: book's name *(categorical)*
- Price: the book's price *(numerical)*
- Author: the book's author *(categorical)*
- Rating: the book's rating *(numerical)*
- Review: how many reviews of the book *(numerical)*
- Type: the book cover's type (only amazon) *(categorical)*

The missing data:

0.3% of Amazon (not much), quite clean

6.8% of Tiki (very much), figure 1 shows the missing value of Tiki

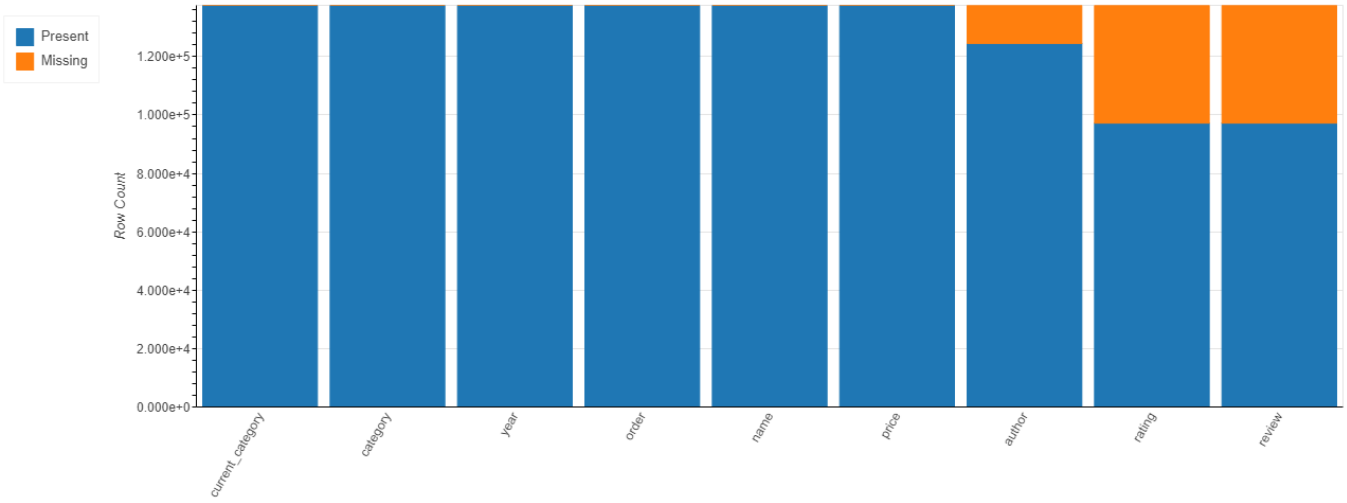


Figure 1: Missing value of Tiki

Almost missing data is *rating* and *review* (each of them are missed up to 30% cells), besides that, we can see *author* are also have a large number of missing cells.

III. Data Cleaning

Data cleaning is very important jobs when doing EDA. *'Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.'*(Wikipedia).

First, we need to ensure the consistency of data value after scraping and having the data, like remove the '$' currency of *price*, remove ',' on numeric value, set the data type for each column…We change the currency unit used in Tiki from VND to USD, using the exchange rate of 23030 VND/1 USD. We also roughly translate the categories of Tiki from Vietnamese to English.

For missing data, for categorical variables, we fill in 'Unknown'; For the numerical variables, we ignore them.

For noisy data, the causing of it is happened when the data is old, for a long time and not updated (Database or the fault of programmer), Like different categories, products which are not book but still have index on the table. We do

some statistic methods like getting the mean value of year and then remove those noisy rows.

We also faced the table level: Heterogeneous data representations: different ways of representing the same real-world entity, Existence of approximate duplicates:

Ex:

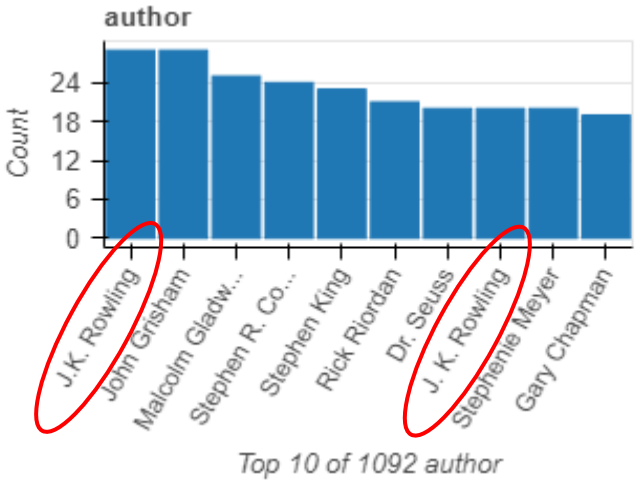


Figure 2: Instead of: 'J.K. Rowling', mistyping 'J.K. Rowling '

For those types of error, we are using OpenRefine, a friendly, easy-to-use Desktop Application

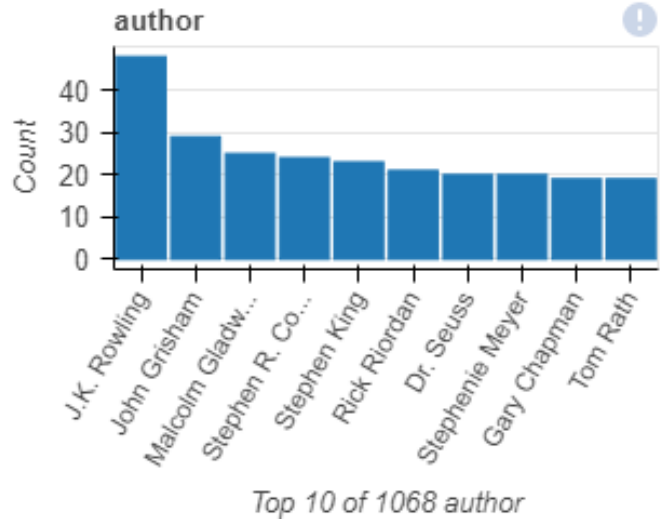


Figure 3: After using OpenRefine to correct error

IV. EDA

1. Basic analysis

Univariate EDA:

Amazon:

Overview

Dataset Statistics

Number of Variables	9
Number of Rows	2689
Missing Cells	68
Missing Cells (%)	0.3%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	1.1 MB
Average Row Size in Memory	434.3 B
Variable Types	Categorical: 5 Numerical: 4

Dataset Insights

`order` is uniformly distributed	Uniform
`order` is skewed	Skewed
`rating` is skewed	Skewed
`review` is skewed	Skewed
`price` is skewed	Skewed
`name` has a high cardinality: 1665 distinct values	High Cardinality
`author` has a high cardinality: 1068 distinct values	High Cardinality
`category` has a high cardinality: 189 distinct values	High Cardinality
`year` has constant length 4	Constant Length

Tiki:

Overview

Dataset Statistics

Number of Variables	9
Number of Rows	137553
Missing Cells	93681
Missing Cells (%)	7.6%
Duplicate Rows	11
Duplicate Rows (%)	0.0%
Total Size in Memory	100.8 MB
Average Row Size in Memory	768.7 B
Variable Types	Categorical: 4 Numerical: 5

Dataset Insights

`author` has 13027 (9.47%) missing values	Missing
`rating` has 40327 (29.32%) missing values	Missing
`review` has 40327 (29.32%) missing values	Missing
`year` is skewed	Skewed
`price` is skewed	Skewed
`rating` is skewed	Skewed
`review` is skewed	Skewed
`current_category` has a high cardinality: 379 distinct values	High Cardinality
`category` has a high cardinality: 450 distinct values	High Cardinality
`name` has a high cardinality: 57499 distinct values	High Cardinality
`author` has a high cardinality: 13073 distinct values	High Cardinality

For more detail of each column, we also have a basic report html file for each site.

In conclude, we see some similarity between two sites: the categorical variables have high cardinality; the numerical variables tend to be skewed except that the *order* of Amazon is uniform distribution while the *order* of Tiki is skewed. This happened because of the data. Tiki doesn't set every categorical or every year full of rank. Some rare categories or the years those far from current have the ranks stop very soon ()

2. Pearson & Spearman Correlation

Multivariate EDA between two numerical variables:

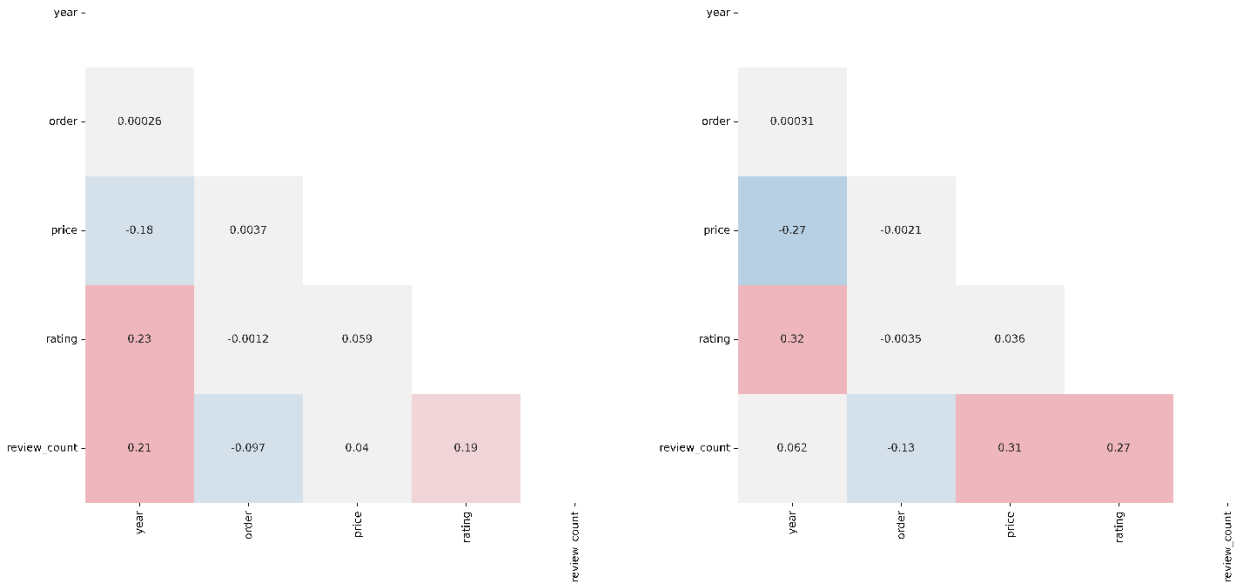


Figure 4. Pearson's Correlation (left) and Spearman's correlation (right)

The heat map show that our correlation between two numerical variables is not quite 'related' to each other: no linear correlation, no clear monotonic function between two numerical variables

3. Chi square

Multivariate EDA between categorical variables:

We have 4 categorical variables, which is: *name, author, category, type,* and *year* which we treated as a categorical variable.

We apply chi-square for *year* and *type* because we can be easily to follow what happened (*year* has about 20 values, *type* has 13 values while the others have so many values).

To use the chi-square test, we need to perform the following steps:

1. Define null hypothesis and alternate hypothesis. They are:

H_o (Null Hypothesis) — that the 2 categorical variables being compared are independent of each other.

H_1 (Alternate Hypothesis) — that the 2 categorical variables being compared are dependent on each other.

2. Decide on the α value. This is the risk that willing to take in drawing the wrong conclusion. Setting α=0.05 when testing for independence. This means we are undertaking a 5% risk of concluding that two variables are independent when in reality they are not.

3. Calculate the chi-square score using the two categorical variables and use it to calculate the p-value. A low p-value means there is a high correlation between two categorical variables (they are dependent on each other). The p-value is calculated from the chi-square score. The p-value will tell you if your tests results are significant or not.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Figure 5. calculate the chi-square value for each cell using the formula for χ2

Results:

chi-square score	Degree of Freedom	p-value
550.612675662608	338	2.117512805492058e-12

With p-value = $2{,}12 \times 10^{-12} << 0.05(\alpha)$ - this means the two categorical variables (*type, year*) are correlated.

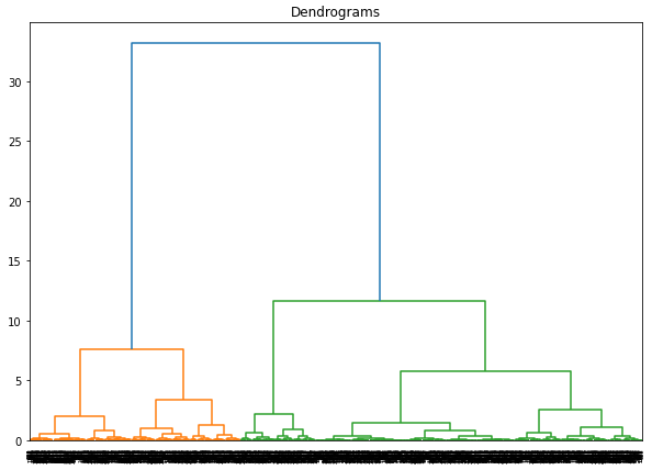
4. Clustering

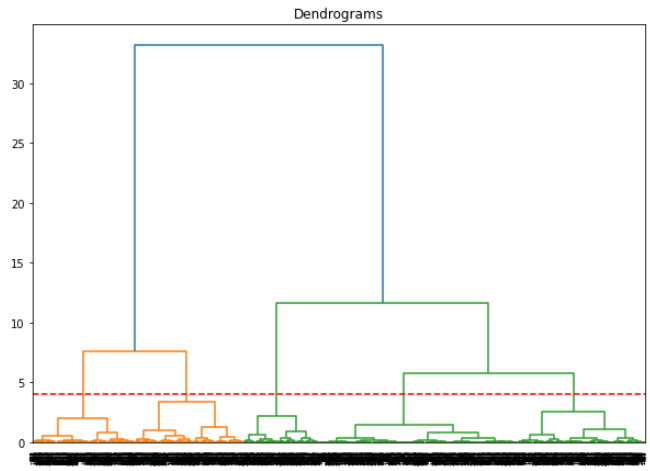
For clustering, if there are just two dimensions (variables), we could simply use a scatterplot to look for the clumps. But when there are many variables, the task becomes more challenging and thus it necessitates algorithms. There are two major types of clustering algorithms: 1) Hierarchical clustering, 2) non-hierarchical clustering (k-mean clustering).

If we just looking at our raw data, especially so many variables, the way we perform plotting and from that, we give decision how many clusters that fit to our data is tricky, and the initial number of clusters for k-mean is challenging. That's how we choose ***Hierarchical clustering***.

First, Hierarchical clustering work only for continuous numeric variables, so we need to drop the categorical variables (assume that customer only pay their attention on the number), then fill the null value by the mean value. Then, normalize the data, so that all the variables to be the same scale.

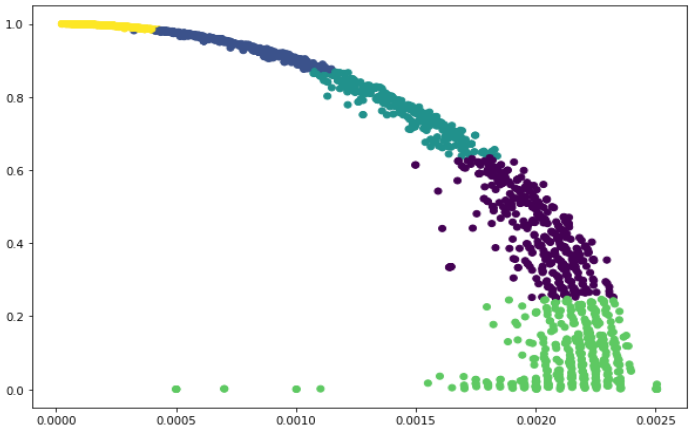
Second, we draw the dendrogram to help us decide the number of clusters for this problem (the number of clusters is not important depends on our chosen, it only has meaning if we have already determined particular clusters)





We have 5 clusters after choosing threshold. We merge the most similar points or clusters in hierarchical clustering. We decide which points are similar and which are not by taking the distance between the centroids of these clusters. The points having the least distance are referred to as similar points and we can merge them. We can refer to this as a distance-based algorithm as well (since we are calculating the distances (here we use *Euclidean* distance) between the clusters).

Results:



V. Visualization

1. Name

Book name plays an important role in grabbing the reader's attention, especially for online shopping, when it's harder for the customers to read a few pages inside the book compared to shopping in the traditional bookstores. In figure 2,3,4, we used Python's Wordcloud library to respectively represent the most common words and phrases in top-selling book names from Amazon and Tiki to see if there is any naming trend in the book market.

Figure 6. Amazon's Book

As we filtered out the words that are common in English and Vietnamese in general, the most common word from Amazon's book names is 'life', followed by 'guide'. This is reasonable since a lot of people read books to get advice in life or to get the experiences from the life of someone else. It's not a surprise that 'secret' is on top, as the word is intriguing, making the book sounds rarer and more valuable. 'Cookbook' and 'Recipe' are quite popular, which may infer that American customer have a great interest in cooking and cuisine. "Harry Potter" is a very special case, it is a novel series from J. K. Rowling, not a general word that many books share like the others. The fact that it is one of the most popular phrases just shows how successful the series is.

Figure 7. Tiki's English book

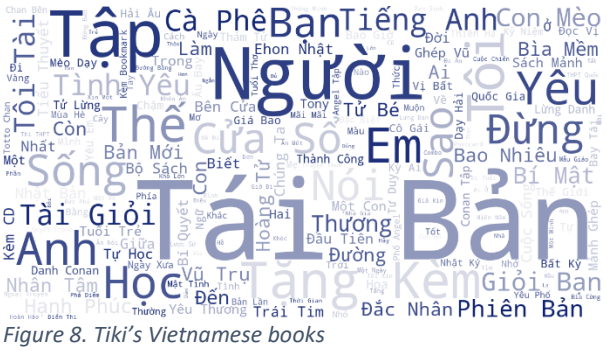


Figure 8. Tiki's Vietnamese books

Harry Potter is not only successful in the US but also very famous in Vietnam too, considering it the most popular phrase in Tiki's English book names. Other than that, most of the other words come from the books for English learners, since learning English with books written in English is a good practice to speed up the learning process.

In Tiki's Vietnamese books, however, there is no clear trend in book names. The most common word is 'Tái Bản' (republish) because in Vietnam, many publishing houses tend to reprint the popular books in the past with new cover to boost the sale. The rest contains multiple word spans from the title of some prominent books: "Tôi tài giỏi, bạn cũng thế" (I am gifted, so are you), "Hoàng tử bé" (Le Petit Prince),.. The word "Học" (Learning) and "Tiếng Anh" (English) again show that many Vietnamese buy book for learning purpose, especially about English language. The reason maybe that the online book market in Vietnam is new and most of the customers are the young people who are more familiar with new technologies.

2. Category

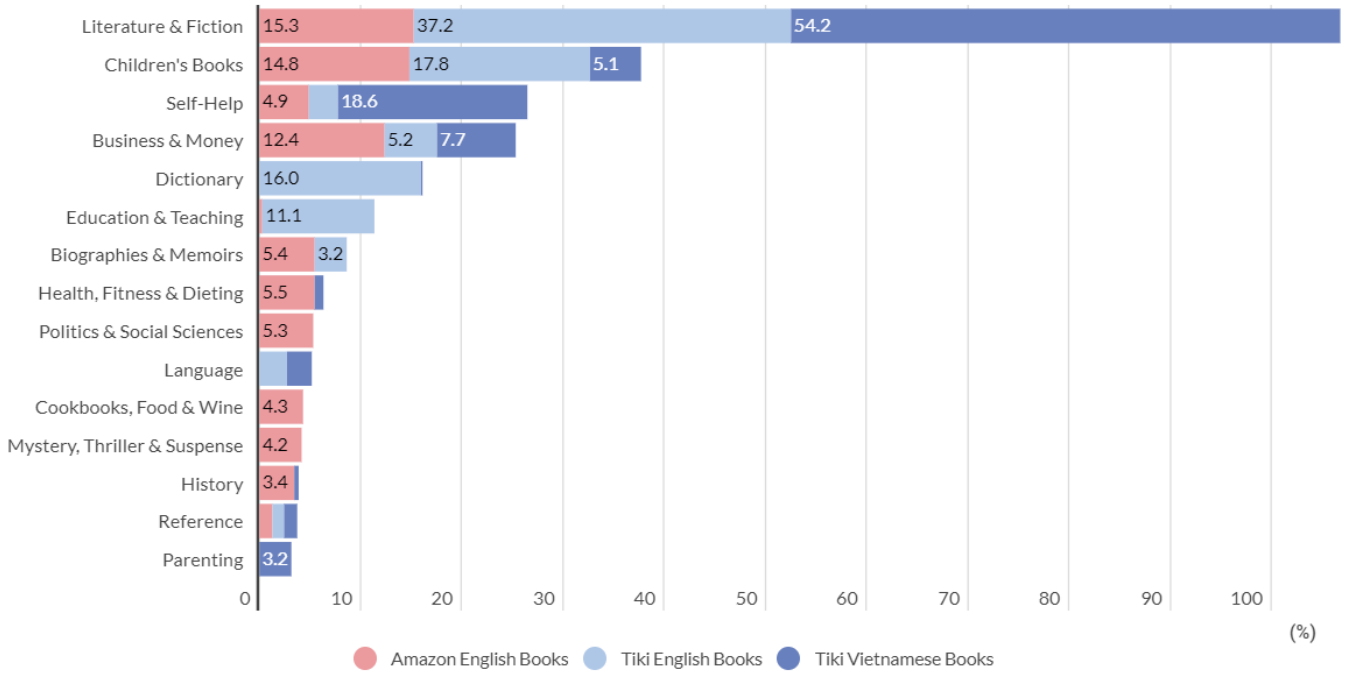


Figure 9. Top 15 popular categories

(See the interactive version at: Inforgram)

Figure 5 represents the percentage of the top 15 popular categories overall, ranked by the sum percentage of each category. Overall, Amazon's books are more evenly distributed than Tiki's books which are more concentrated in 3-4 top categories. Literature & Fiction is dominating, it leads in popularity in all 3 sections, although having significant difference in the proportion. For Tiki Vietnamese book, 54.2% of top-selling books are Literature & Fiction, while the figure for Amazon's book is only 15.3%. Self-help books sell well in Vietnamese with the proportion of 18.6%, but not for the other two groups. As discussed before, Dictionary and Education & Teaching are important categories in Tiki's English Book, they take up a more than a quarter of this group.

3. Rating distribution

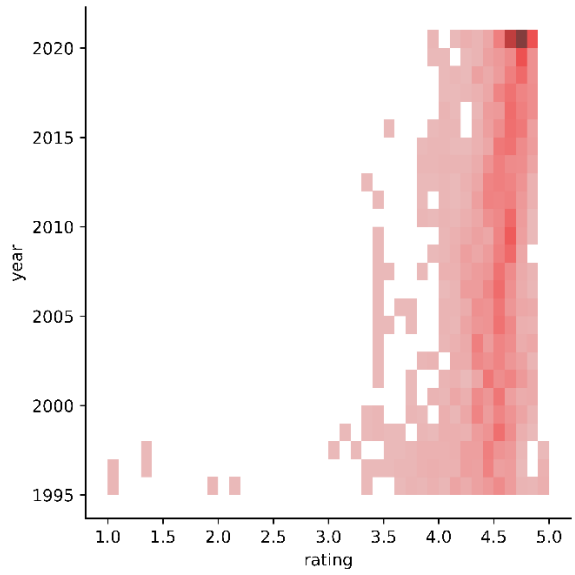


Figure 10. Amazon's rating distribution

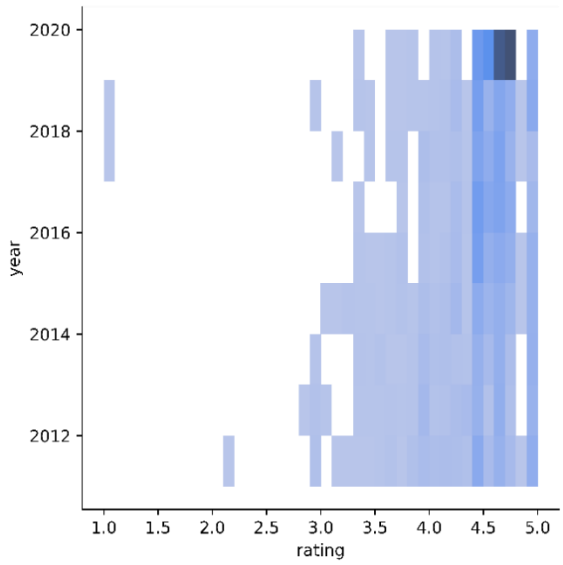


Figure 11. Tiki's rating distribution

Figures 6 and 7 respectively illustrate the distribution of user rating on books from Amazon and Tiki over years. The overall trend is similar between the two sources: More evenly distributed, with the old books, more concentrated on the higher rating for newer ones. However, while the highest rating, 5 out of 5 stars, has disappeared in Amazon books for many years, it is not a really rare thing in Tiki. This happens because Amazon is a much larger market, every book from the top 100 in recent years has at least hundreds of reviews, which makes achieving the perfect rating nearly impossible. For Tiki, in contrast, many books have less than a hundred reviews.

Another interesting thing is that in the most recent year data from both sources, the proportion of books that have rating at 4.7 and 4.8 is significantly higher than others, and the most common rating is 4.8. This may indicate that the books which are trending tend to receive more positive reviews and get a little bit overrated when they first came out. However, the numbers will go down and be stabilized as time passes.

4. Mean price

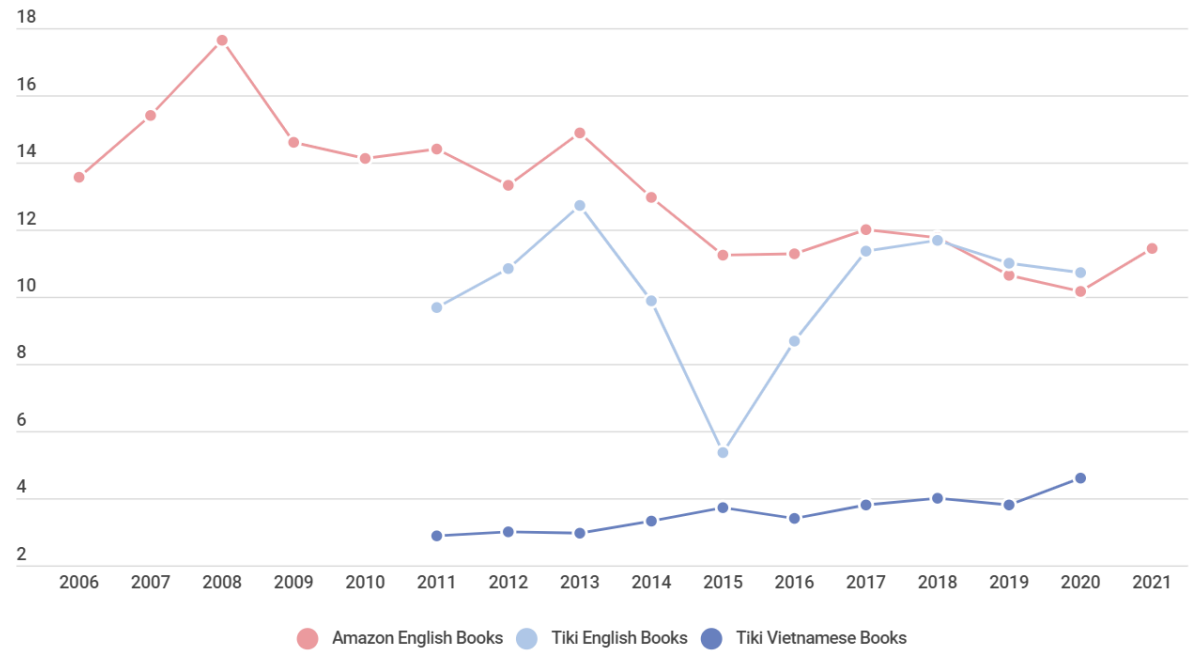


Figure 12. Mean price of books from 2006- 2021

(See the interactive version at: Inforgram)

Figure 8 shows the change in mean price of books from 3 groups between 2006 and 2021. There is a significant difference in the mean price of English and Vietnamese books. Vietnamese books, in general, are much cheaper. This could be caused by factors: cheaper cost to make books in Vietnam (almost English books on Tiki are imported), lower average income, and most books are paperback. However, the mean price of Vietnamese books steadily rises over the years, in contrast with the overall falling trend of English books. It's a surprise that the rise and fall in the mean price of English books on Amazon and Tiki are similar, considering that the main categories are much different.

VI. Conclusion

Every step of EDA is very important, from nothing, we find data, see what data we have, cleaning them, decided what type of variables they should be and then determined what EDA technique would apply to each type of variables, combine with visualization to find the data insight.

The project top book sale analysis is very interesting, but the data is hard to adapt our expectation. Almost variables are beautiful and great except the *price*, which is particular book should be had different price at a particular time and then we can do time series analysis but unfortunately each book or each product has the unique identify value, and the price always must be updated hourly with this id value.

Contribution

Task		Nguyen Duy Hung	Dang Thanh Lam	Le Hai Son	Nguyen Hoang Nhat Quang
Programing task	Scrape from Amazon	50%	50%		
	Scrape from Tiki	75%	25%		
	Data cleaning	25%	75%		
Analytic task	Subject proposal			50%	50%
	Basic analysis			50%	50%
	Pearson's and Spearman's correlation			50%	50%
	Chi square	50%			50%
	Clustering	50%	50%		
	Visualization		50%	50%	
	Writing report	25%	25%	25%	25%

Reference

- Data source:
  - Amazon Best Sellers in Books
  - Tiki Bookstore Best Sellers
- Scraping Data:
  - Scrapy Tutorial — Scrapy 2.5.1 documentation
  - scrapy-plugins/scrapy-splash: Scrapy+Splash for JavaScript integration (github.com)
- Data cleaning: OpenRefine
- Chi Square: Statistics in Python — Using Chi-Square for Feature Selection | by Wei-Meng Lee | Towards Data Science
- Clustering: Hierarchical Clustering | Hierarchical Clustering Python (analyticsvidhya.com)
- Data Visualization:
  - From data to Viz | Find the graphic you need (data-to-viz.com)
  - Infogram