Data Analysis : The Cool Stuff

Sowrabha Horatti Gopal
sgopal3@wisc.edu

Pradeep Kashyap Ramaswamy
pradeep.kashyap@wisc.edu

Sharath Prabhudeva Hiremath
shiremath@wisc.edu

1 Details of Table used for Analysis

After the amazing laborious work of data acquisition, cleaning, blocking, matching, merging and mapping (the boring part as the Prof. calls it), we finally obtained our Table on books with clean and integrated data. Schema of the Table is shown below in 1 -

Attributes
Title
Author
ISBN13
Publication
Published_Date
Ratings
Genres
Reviews
Average Rating
Language

Table 1: Final Schema for Table E

Number of tuples present in final combined table E - **1268**

1.1 Sample

Table 2 gives 6 sample rows from the final combined table.

ID	Title	Author	ISBN13	Publication	Date	Ratings	Genres	Reviews	Avg_Rtg	Lang
1	Fool's Errand (Tawny Man Series #1)	Robin Hobb	9780553582444	Random House Publishing Group	2002	"52,825"	"Fantasy, Fiction, Fantasy, High Fantasy, Science Fiction Fantasy"	"1,157"	4.28	English
2	The Yiddish Policemen's Union	Michael Chabon	9780007149827	HarperCollins	2008	"50,351"	"Fiction, Mystery, Science Fiction, Science Fiction, Alternate History, Mystery, Crime"	"5,860"	3.69	English
3	Unless	Carol Shields	9780007154616	Fourth Estate (GB)	2006	"10,719"	"Fiction, Cultural, Canada"	877	3.63	English
4	Remarkable Creatures	Tracy Chevalier	9780007178377	HarperCollins Publishers Ltd	2010	"31,582"	"Historical Fiction, Fiction, Historical, European Literature, British Literature, Literature, 19th Century, Adult Fiction, Womens, Audiobook, Book Club, Adult"	"3,674"	3.8	English
5	The Winter Rose	Jennifer Donnelly	9780007191321	Harper Collins	2009	"14,609"	"Historical Fiction, Romance, Historical, Fiction, Romance, Historical Romance"	"1,197"	4.3	English
6	Genghis: Bones of the Hills (Khan Dynasty Series #3)	Conn Iggulden	9780385342803	Random House Publishing Group	2010	"12,327"	"Historical Fiction, Fiction, Historical, War"	363	4.32	English

Table 2: Sample data from Final Table E

2 Data Analysis

2.1 Things to Explore

We wanted to discover various statistics based on the dimensions - publisher name, genres, year, language from our data rich table. So we decided to perform OLAP - style analysis on our table. Here's a list of things we wanted to explore -

- Which genres are stuffed with the highest number of books? And the genres which no one knows that they exist (somewhat like Google Plus)?
- What is the average rating of books for each of the top genres?
- Who are the top publishing houses in each genre?
- The total number of ratings for all books based on year
- Average Rating by Language

We started out with the above questions and found some interesting results on our journey in the **cool** Data Analysis world.

2.2 Exploration

We used pandas - Python Data Analysis Library, to analyze the data in our table in OLAP style. Here's a brief description of the stages of our data analysis (which can also be see in our jupyter notebook)

- **Load Data :** First step was to load (*obviously!*) our integrated table from Stage 4 as a pandas dataframe.
- **Sanitize the Columns :** Some of the columns weren't in our desired data type. So we had to convert data types and perform some cleaning to remove NaN values. For example, the Genres column was represented as a string with multiple values separated by commas, which was not suitable for any data analysis. So we had to perform some transformation and bring it analyzable form. So we wrote a lambda function to split the string by commas and then strip each word of whitespaces. We applied this function on Genres column and stored it as a list.
- **Roll - Up :** Then we rolled up our data to get the number of Books present in each genre. This is how we found out the top 10 genres and the bottom 10 genres, which can be seen in **figure** 1
- **Slicing :** Once we had top genres, we sliced our table into 5 dataframes one each for the top 5 genres. On these sliced dataframes, we started looking into the data for any interesting insights.
- **Roll - Up Returns:** We grouped of each sliced dataframe by publisher and aggregated the number of books. This helped us answer who the top three publishers are in each genre. Also we rolled up to find out the average rating in each genre, as this might suggest the readers' interest and overall rating of the genre.
- **Roll - Up Reloaded :** Again we went back to our source Table E to perform some aggregation operations. We calculated the average rating based on the language this time. We grouped by the language and then found out the average rating for each language. We also calculated the total number of ratings given for all the books combined every year.
- **Drill - Down :** Upon calculating the total number of ratings for every year, we found out some unusual pattern for year 2006. So we drilled down for 2006 data and explored it further. This is discussed in detail in next section.

3 Conclusion

3.1 Learnings

3.1.1 Genre Analysis

We wanted to find out the interests of readers. Basically, what subjects pique most of the readers' interest. Is it Fantasy? or Fiction? or any other genre? We present our findings here.

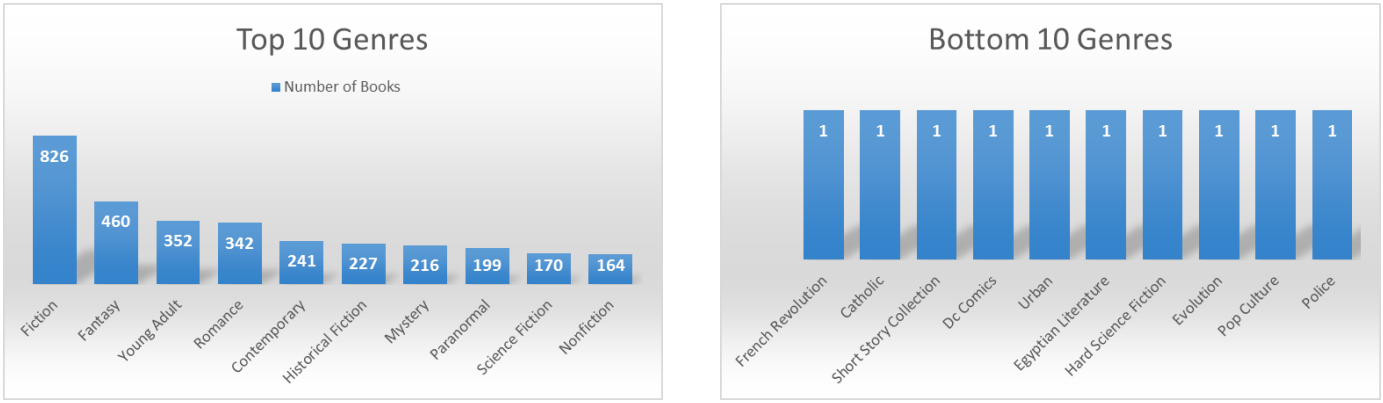

Figure 1: Top and Bottom 10 genres

In the Figure 1, we can clearly see who wins the competition. It comes as no surprise that ***fiction*** and ***fantasy*** take top 2 positions. Also the next three - **Young Adult, Romance, Contemporary** are among the usual suspects.

And then we can look at the bottom 10 genres. These ended up in the bottom mostly because of their unique naming. *Although an atheist might say Catholic Genre is at that right position.*

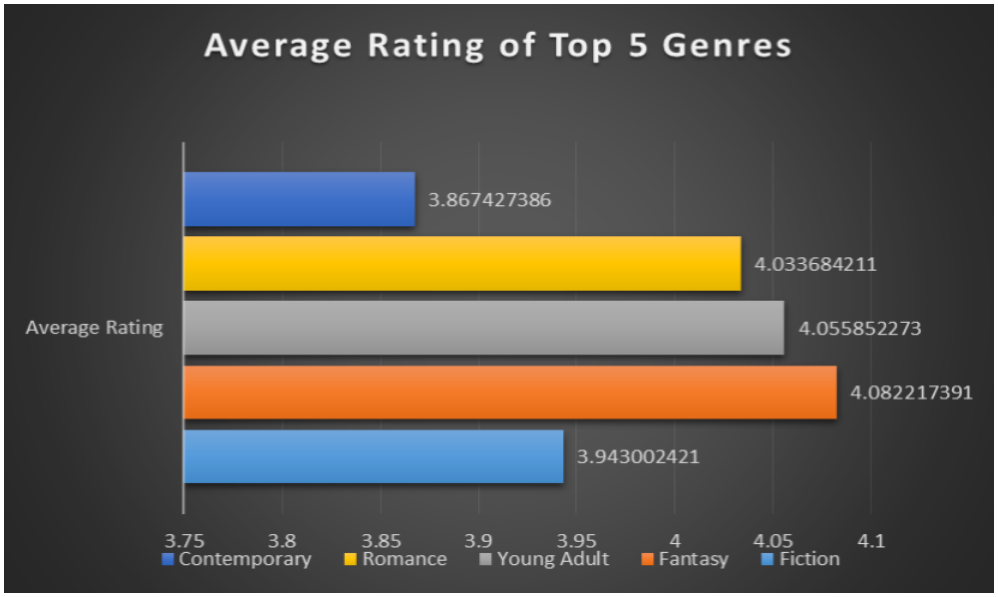

Figure 2: Average Rating by Genre

Then we plotted a bar graph representing the average rating of each genre. Again, Fantasy and Fiction are in top 5 occupied the first and fourth positions, though by a very close margin.

3.1.2 Publisher Analysis based on Genre

Then we set out to analyze top publishers in each genre. The following figures represent the top three publishers in each genre.

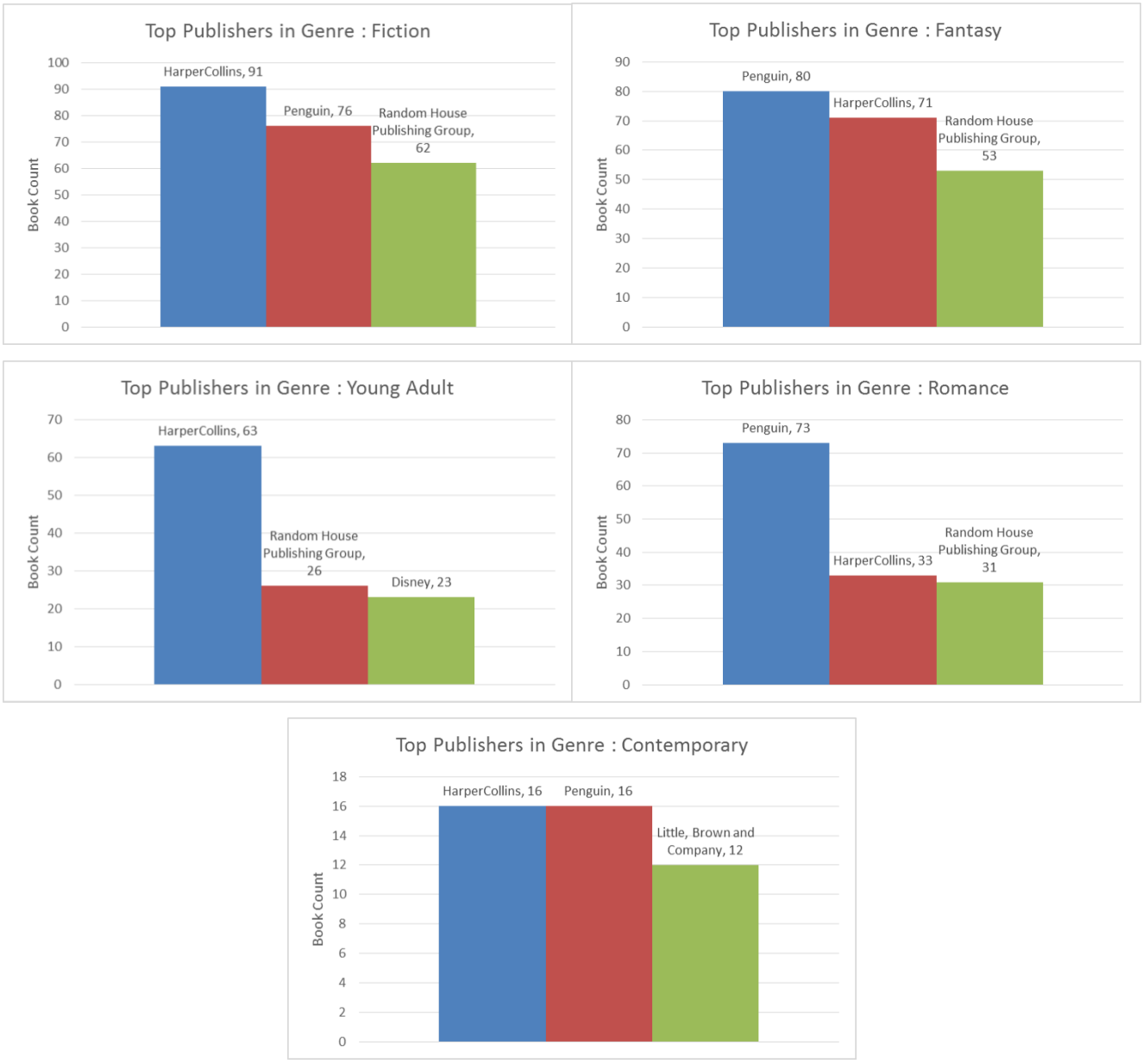

Figure 3: Top Publishers in Top 5 Genres

We can see that the top positions in each genre is always occupied by HarperCollins or Penguin. We can infer from this that these are the top publishing houses, publishing most books in the most popular genres.

3.1.3 Stats based on Year (And the curious case of 2006!)

Interesting Fact Alert! We wanted to study user involvement, in terms of total number of ratings given to books, grouped by their published year. This is a good metric of reader engagement. First we obtained the total number of ratings by published year, which we can see in figure 4

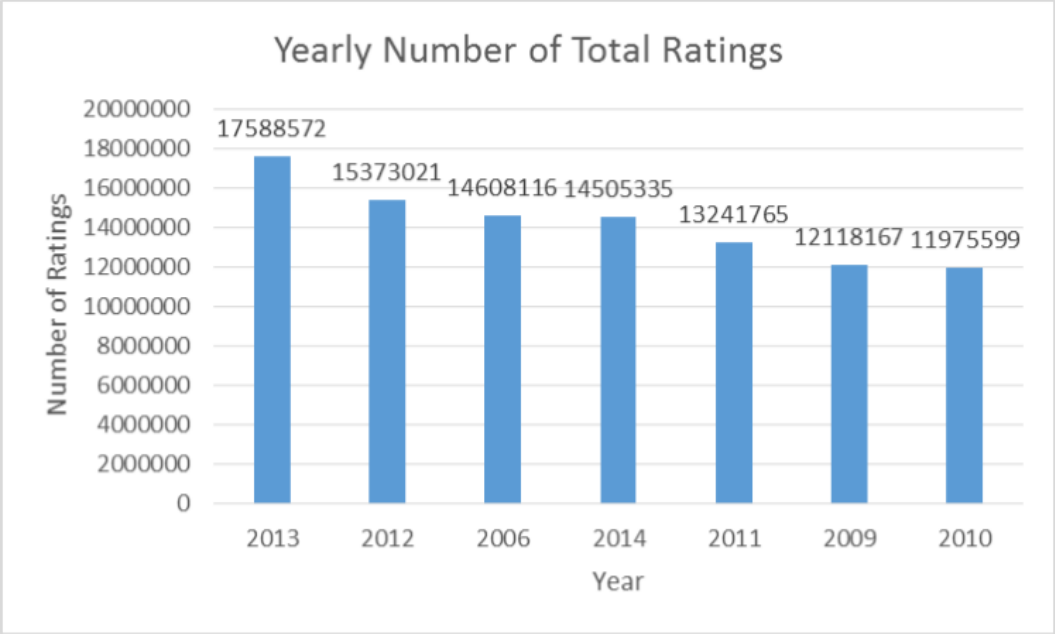

Figure 4: Total Number of Book Ratings by Year

We wanted to know why books published in these particular years got more number of ratings when compared to others. Our first suspicion was that our table has more number of books published in these years. In other words, our table has more number of rows for these years. So we plotted a graph representing number of books against published year. This can be seen in figure 5

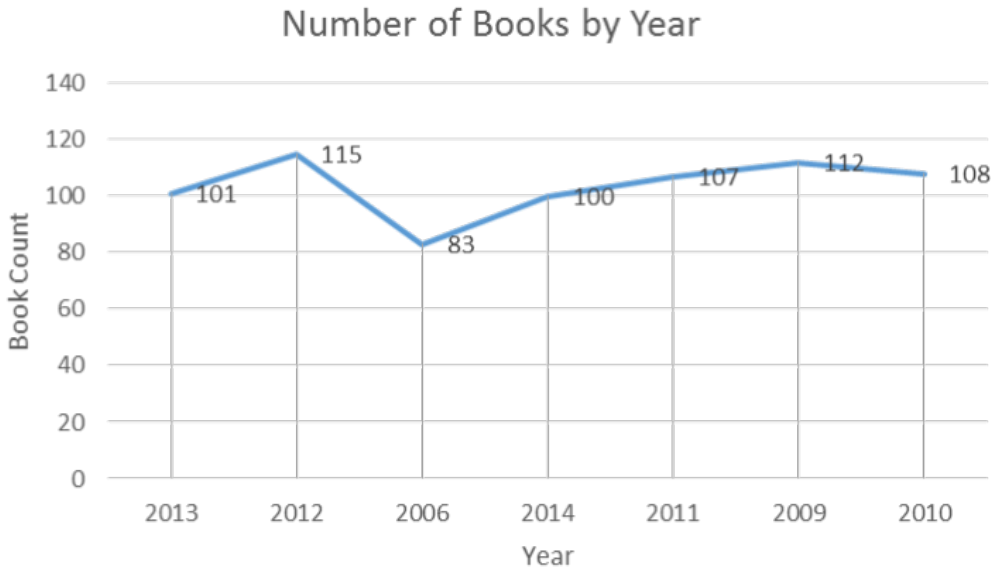

Figure 5: Book Count by Year

Now, it got even more interesting. Though we saw a trend that all these years had book numbers more than 100, **year 2006** had way less number of books when compared to others. There must be some reason for this anomaly in user engagement. So we decided to drill down into particulars of books published in the year 2006. We picked picked the top 10 books based on total number of ratings for the year 2006, which is shown in the figure 6.

Title	Author	Published	Ratings	Avg_Rating
Twilight	Stephenie Meyer	2006	3734750	3.57
Harry Potter and the Half-Blood Prince (Harry Potter Series #6)	J. K. Rowling	2006	1692717	4.53
The Lightning Thief (Percy Jackson and the Olympians Series #1)	Rick Riordan	2006	1318508	4.22
The Glass Castle	Jeannette Walls	2006	610198	4.23
Deception Point	Dan Brown	2006	456653	3.66
Never Let Me Go	Kazuo Ishiguro	2006	306014	3.8
Extremely Loud and Incredibly Close	Jonathan Safran Foer	2006	305581	3.97
The Truth About Forever	Sarah Dessen	2006	176304	4.13
Dead as a Doornail (Sookie Stackhouse/Southern Vampire Series #5)	Charlaine Harris	2006	176007	4.01
Cell	Stephen King	2006	145610	3.64

Figure 6: Top 10 Books by No. of Ratings in 2006

Now we have our answer!
As we can see from the top books, this was the year in which Harry Potter and the Half-Blood Prince, Twilight [1], percy Jackson and The Lightning Thief, Deception Point, Never Let Me Go were published. These are some of the best-selling novels of all time. Because of their popularity, the reader engagement is more for these books, which resulted in high number of total ratings for 2006!

[1] *We still believe that Twilight is an injustice towards mankind, just as much as Justin Bieber*

3.1.4 Language based Rating

We wanted to see which language books had the highest average rating. We found out the following result -

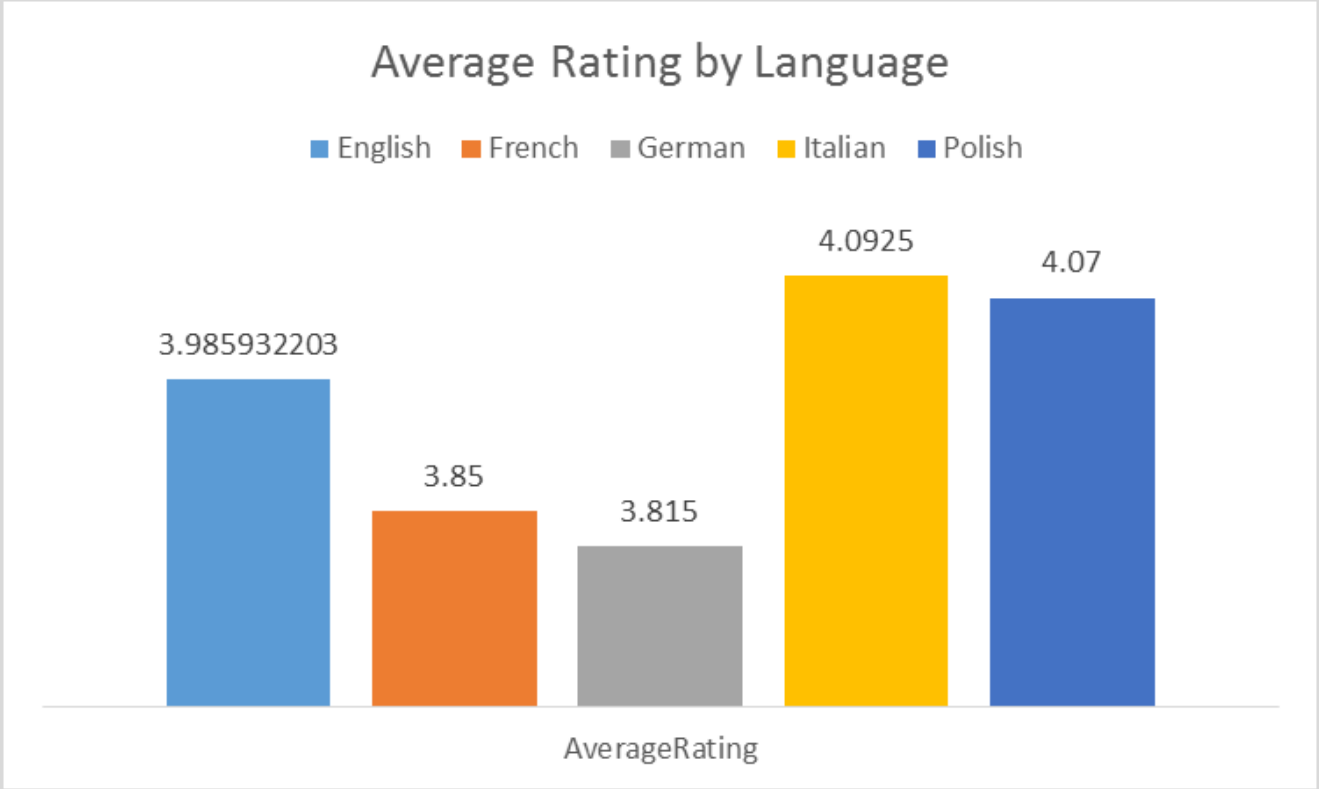

Figure 7: Average Rating by Language

It was interesting that Italian and Polish beat English in Average Rating. Then we drilled down to see what's so great about these two languages. We discovered that there were only 4 Italian books in the table and only one Polish book, whereas we had 1239 English books. So, we found out that this average rating is unjustified.

3.2 Problems

Here's a brief summary of the problems we faced during this phase of project -

- The first problem we discovered was that there were some rows with null value in Genres field. We had to fliter these out.
- Genres column gave us the most trouble. As already mentioned, the genres were represented as a string with values separated by commas. So we had to process this and perform a transformation to convert the string into a list.
- Again, the Genres column had duplicated values, so we had to remove the duplicated genre name in each cell of Genres column.
- Since each entry had multiple genres, it was a tricky analysis procedure for us. We couldn't directly perform simple groupby and aggregations. This wouldn't have yielded any results based on genres. We had to take special routes to get our results.
- Other problems included, too much whitespaces in some columns like Publishers. We had to clean that out as well.
- We also had to normailize the publisher names. For example, 'Harper Collins' and 'HarperCollins' are the same, but won't group together during analysis because of that space in between, in the former case.

4 Future Work

The Project description page says "If you have more time, what would you propose you can do next?" Well, if we had more time, we would have studied well for Adv.OS exam.

Jokes apart, here's a list of things we planned on doing if we had more time

- Perform analysis based on the authors
- Find out the correlation between authors and publishing house.
- Prediction of rating based on author and publishing house
- Correlation of book's title with it's popularityrating
- Clustering of books based on genres.
- Prediction of price of book based on year, author and publishing house (and other details if we could get more)