**CS839 Project Stage 4 Report**  
Sreejita Dutta, Deepanshu Gera, Rahul Jayan

1. **Dataset and Data Merging**

**<Deepanshu Fill THIS>**

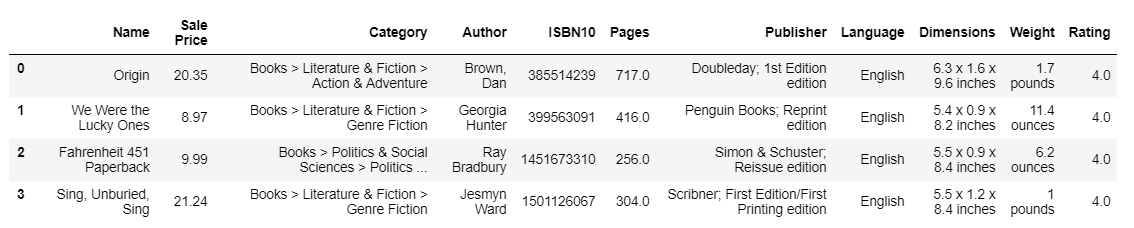
1. **Statistics of Merged Table**

The final dataset contains 485 tuples.

The schema of this table includes the following attributes:

*'Name', 'Sale Price', 'Category', 'Author', 'ISBN10', 'Pages', 'Publisher', 'Language', 'Dimensions', 'Weight', 'Rating'*

Here are four representative tuples in the table.



1. **Data analysis: tasks, steps and conclusions**

We perform **OLAP-style** exploration for analyzing our merged dataset. This involves the four most common operations used in OLAP: roll-up, drill-down, slice and dice.

Based on our dataset, we aim to answer the following questions:

1. **What is the average price of books in each category?**
2. **How many books are available in each category?**
3. **What are the names of reasonably priced books available in the category with highest number of books?**
4. **Publisher vs Price**
5. **What is the average rating of books in each price range?**

We group the Sale Price of books in 10$ intervals between 0 – 99 $ since all our books fall into this price range. For each price range, we compute the average rating of the books that are sold in that price range.

Here, we can observe that most books are sold at a price below 50$ and the average rating of books is highest in the 40-49$ price range and lowest in the 20-29$ price range.



1. **Which authors receive high ratings for their books consistently?**

We aim to find a list of favorite authors from the given dataset. For answering this question, we decided to obtain the names of authors who have received a rating of at least 4 stars out of 5 for more than 7 books.

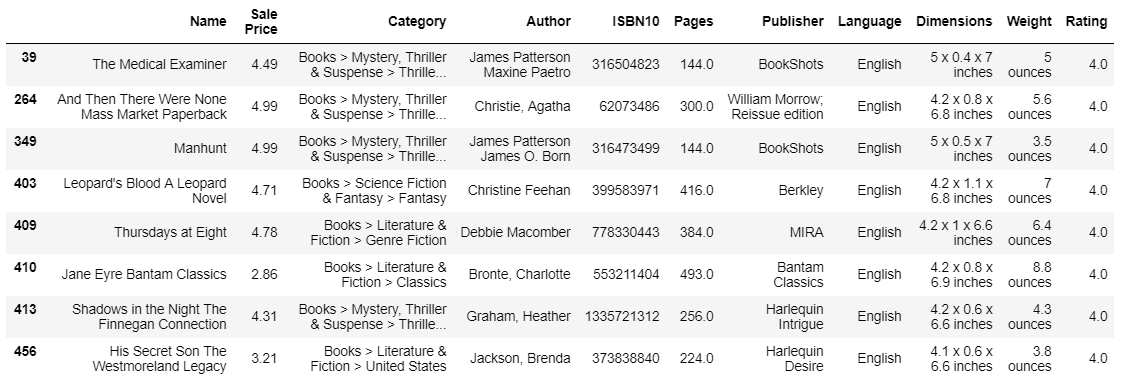
First, we normalize the author names (as they are in no specific format). Then, we find all the authors that satisfy the above conditions. Here is the list in decreasing order of popularity.

*Stephen King, Nora Roberts, Lee Child, Louise Penny, Patricia Briggs, Janet Evanovich, Brandon Sanderson,* *James Patterson*

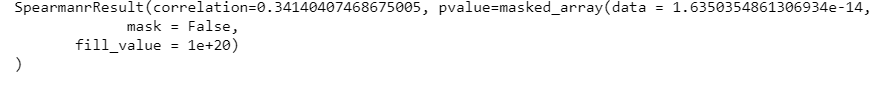
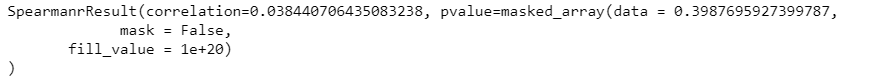
1. **Which books have the potential to offer greatest value for money?**

This information could potentially be useful to customers who want to buy a high rating book (at least 4 starts out of 5) at a reasonable price of less than 5$.

From the obtained list of 8 books qualifying this condition, it is interesting to note that there are two books written by *James Patterson* (who is one of the popular authors from 6.) and majority of the books fall under the *Thriller & Suspense* category.



We also compute Spearman Rank correlation coefficient to find if any monotonic **correlation** exists between:

* ‘Sale Price’ and ‘Rating’: This is to figure out if highly rated books are generally priced higher. The result obtained shows that there exists a correlation between the two attributes and the low p-value indicates it is statistically significant. 
* ‘Sale Price’ and ‘Pages’: This is to understand if there is any relation between larger books and higher prices. The result shows that there doesn’t seem to be any correlation between the two variables. 

Some problems in our data analysis is listed as follows.

* **Author names are not entered in any specific order or format.**

The names of authors are not entered in any specific format like *<First Name, Last Name>*. For example, Dan Brown is entered in two ways: i) Dan Brown ii) Brown, Dan. There is also no clear way in which multiple authors are separated. Due to these issues, extra steps of analysis and transformation had to be performed before utilizing author data.

* **Preprocessing is required for certain fields to perform any analysis.**

Attributes like “Category” and “Publisher” have extraneous data. So, they needed to be preprocessed to avoid interference with the analysis tasks. For example, some publisher names were surrounded by texts indicating the print type or the print edition.

1. **Future work**

We can perform various other analysis task on this dataset like classification, clustering, association rule mining, anomaly detection, and many other OLAP queries, provided the size of data is sufficient.

Examples of some of these are:

1. **Supervised learning:**  In our dataset, we have certain tuples where the Rating of the book is not present. Given sufficient training data, we could train a classifier to predict the rating of a book. This could also be helpful in predicting the ratings of new books.
2. **Clustering:** We could cluster books based on genre, author, price and rating. This might be useful in certain scenarios. For example, books within the same cluster might appeal to a certain demographic.
3. **Code for merging the tables**

**<Deepanshu FILL THIS>**