Final Project Summary

# Divina M. Arpino

23FA15 Data Science Programming

CMP-262-86511

## Project Concept Summary

For my final project I selected to do data analysis for a fictitious client that is opening a small bookstore. The client would like recommendations on books he must absolutely invest in for his initial inventory. Questions being considered:

1. What books make the rankings over several years?
2. Authors with the highest counts of books in top 50 over the years and whether they are fiction or non-fiction books
3. What are the top 5 books in fiction and non-fiction genre based on highest number or rating and user reviews for each year?
4. Does cost of book influence their popularity?
5. Does the number of pages influence book popularity?

## Tools Used

Two Datasets used – Book reviews/ratings from Kaggle:

* [Amazon Top 50 Bestselling Books 2009 - 2019](https://www.kaggle.com/datasets/sootersaalu/amazon-top-50-bestselling-books-2009-2019/code)
* [Goodreads-books](https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks)

Other tools used:

* Excel – Initial peek at the data and some column shift adjustments
* Python 3 through Jupyter Notebooks
  + Libraries used: pandas, seaborn, matplotlib
* ChatGPT’s quick guidance / Google searches
* Other MS-Office applications such as Word and PowerPoint

## Insights

In my analysis I found that the data is not ideal for the situation, some additional insight is needed in order to provide a better analysis that more accurately answers the questions. Additional information that would help include:

* Categorization for age applicability (eg. Little Blue Truck)
* Further breakdown of genre (eg. Diagnostic and Statistical Manual of Mental Disorders)
* Box sets vs. individual included titles should be noted

Reviews and Ratings counts are just that, counts. A book could have 1M reviews, but a 2.0 rating. However, a highly reviewed book or one with high counts in ratings is likely a good candidate for being read, regardless of rating – they cannot be discounted.

## Recommendations

My recommendations for the client are as follows:

* Purchase all of the series of books that are in the top 50, most reviewed or most rated results, ie. If three Harry Potter books are listed, purchase more of the three making the list, but have some of all in hand of all others in that series.
* Purchase books from the most reviewed / rated authors, starting with those that are higher rated
* Fiction is a bit more popular than non-fiction, but non-fiction titles that are highly reviewed or rated should be kept on hand.
* Purchase titles that received top 50% or review and rating counts
* Purchase **all** titles that are in the top 50 of previews 10 years

## Future work

* Identify additional datasets to perform same analysis
* Refine my current analysis to exclude additional data, less focus on counts without factoring in ratings
* Further explore Pages to User Rating relationship
* Call you the least popular books
* Figure out why Harry Potter is only popular in late years in this dataset (data integrity?)