Bookstore

Using A dataset from Amazon

Second Term 1445 H

1- Bias and Fairness

1.1 Summary of our research on data bias and fairness

Our research delves into investigating the issues of bias and fairness within the realm of data science, aiming to understand their complexities, identify mitigation strategies, and assess their implications. Starting by exploring common sources of bias, encompassing sampling, selection, measurement, and algorithmic biases. Then, exploring established frameworks and tools, such as the Fairness, Accountability, and Transparency (FAT) framework, AI Fairness 360 (AIF360), and Google's What-If Tool (WIT), used for evaluating and addressing bias and fairness in datasets and models. Through a systematic review of literature and case studies, the research reveals the intricate challenges faced by data scientists in navigating bias and fairness concerns.

Identification of Sources of Bias

After we did the research, we managed to identify and categorizes the various sources of bias that can permeate the machine learning pipeline. These sources encompass biases arising from data collection methods, labeling processes, feature selection criteria, algorithmic design, and evaluation methodologies.

Measurement and Evaluation

We delve into the methodologies and metrics employed for measuring and evaluating bias in machine learning models. This includes an exploration of quantitative measures, qualitative assessments, and interpretability techniques aimed at uncovering and quantifying the extent of bias present in model predictions and decisions.

Overview of Toolkits and Frameworks

We provide an overview of notable toolkits and frameworks specifically designed to address bias and fairness considerations in machine learning. This includes an examination of open-source initiatives such as AI Fairness 360 (AIF360), Google's What-If Tool (WIT), Fairness, Accountability, and Transparency (FAT) framework, and others that offer comprehensive sets of metrics, algorithms, and visualization tools for assessing and mitigating bias in machine learning models.

Mitigation Strategies

Finally, we explore a range of mitigation strategies and best practices aimed at mitigating bias and promoting fairness in machine learning models. These strategies encompass algorithmic

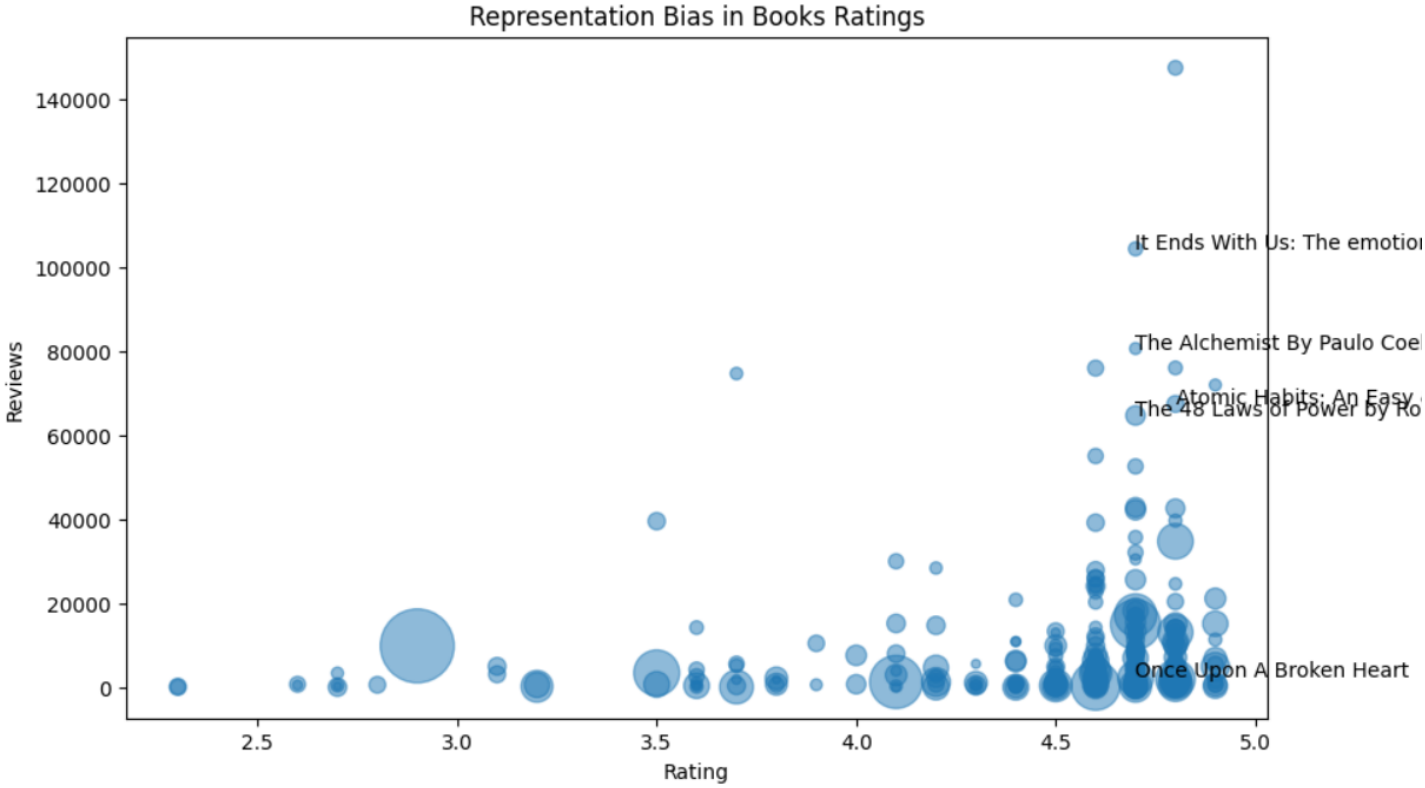
interventions, data preprocessing techniques, transparency and accountability measures, and interdisciplinary collaborations to ensure a holistic approach to fairness-aware machine learning.

Conclusion

In summary, our research on data bias and fairness provides a comprehensive overview of the key concepts, challenges, and strategies associated with addressing bias and promoting fairness in machine learning and data science. By synthesizing insights from diverse perspectives and methodologies.

1.2 Evaluation of our dataset's potential biases

A- Representation Bias: To evaluate this, we plot a scatter plot representing books based on their ratings and the number of reviews they have received, this may show us if there any representation bias.



- **X-axis (Rating):** This axis represents the rating of the books. Each point on the x-axis corresponds to the rating of a particular book.
- **Y-axis (Reviews):** This axis represents the number of reviews each book has received. Each point on the y-axis corresponds to the number of reviews a particular book has.
- **Bubble Size (Price):** The size of each bubble (or point) in the plot represents the price of the book. Larger bubbles indicate higher prices, while smaller bubbles indicate lower prices.
- **Annotations (Book Names):** Each bubble in the plot is annotated with the name of the corresponding book.

In the plot, Y-axis represents the number of reviews each book has received, there appear of points at higher values on the Y-axis indicating a bias towards books that have accumulated a larger number of reviews. This bias might suggest a preference for promoting or featuring books that are already popular or well-known.

B- Sampling Bias:

The dataset may suffer from sampling bias if the books included are not representative of the entire population of books available on Amazon. This could occur if only bestselling or highly rated books are included, which is not in our case, we include a random sample of books available on Amazon.

C- Labeling Bias:

In our books dataset, we may have human bias, this could include personal preferences, cultural biases, or subjective interpretations influencing the labeling process.

D- Feature Selection Bias:

In this analysis, feature selection was not explicitly performed due to the limited number of available columns. so, feature selection bias isn't an issue in our project.

E- Algorithmic Bias:

There's potential for bias in the algorithmic processes used for data collection, feature engineering, or any subsequent analysis. The algorithm we used for data collection (web scraping), may prioritize certain types of content or sources over others, leading to selection bias.

1.3 Implications of biases and fairness of our project's conclusions

- The biases identified could lead to skewed conclusions, especially if the dataset does not accurately represent the broader population of books.
- Biased conclusions may lead to biased recommendations. For example, if certain books are disproportionately favored due to biases in the dataset, it could influence recommendations made to readers, publishers, or retailers.

1.4 Recommendations for mitigating biases in future data collection

To mitigate biases and promote fairness in the analysis, several recommendations can be implemented. Firstly, improving data collection practices is paramount. This involves broadening the scope of data collection beyond e.g. bestsellers or highly rated books to ensure a more diverse and representative sample. Implementing random sampling techniques can further reduce bias in the selection of data.

Addressing representation bias is crucial. Efforts should be made to achieve a more balanced representation e.g. (authors and kinds of books in the dataset), preventing the favoring of certain groups over others. Transparency and documentation play a key role in enhancing the integrity of the analysis. Conduct thorough data cleaning to identify and rectify errors, inconsistencies, and outliers that may introduce bias. Documenting the dataset's limitations, biases, and preprocessing steps enables users to interpret findings accurately and understand the context of the data.

Additionally, algorithmic auditing is crucial. Conducting thorough audits of algorithms used in data collection and analysis helps identify and mitigate algorithmic biases. Continuous monitoring and evaluation are essential to maintain fairness and reliability in subsequent analyses. Regular assessments of the dataset for biases allow for the timely identification and updating of mitigation strategies. Diversifying data sources helps mitigate reliance on a single platform or source, reducing the risk of platform-specific biases.

2- Data Processing and Cleaning

The Data Processing and Cleaning section encompasses a series of tasks aimed at preparing and refining the dataset for analysis:

1- Missing values were addressed by identifying them with **"isnull().sum()"** and subsequently removing rows containing missing data via **"dropna()".**

2- Filtering operations were conducted to exclude rows with specific values; rows with 'Format' labeled as 'Prime Video' or 'None', and rows with 'Author' listed as 'None' were removed.

3- Data transformation steps involved standardizing the 'Rating' column by removing the "out of 5" substring and extracting decimal ratings using regular expressions.

4- The 'Rating', 'Price', and 'Reviews' columns were converted to numeric data types using **"pd.to_numeric()"**.

5- String manipulation methods such as **"str.replace()"** were applied to clean the 'Author' column by removing "by" substring.

6- Duplicate entries in the 'Book Name' column were eliminated with **"drop_duplicates()".**

7- Categorical features like 'Format' and 'Other Format' were encoded numerically using **LabelEncoder** from **"sklearn.preprocessing".**

8- The most frequently occurring author was identified with **"mode()"** to gain insights into the dataset. And helping us to answer our third question in phase 1.

9- Summary statistics were generated using **"describe()"** to understand numerical variables better in the further analysis.

10- Calculates the average rating of books for specified price ranges by splitting the price into ranges to help us answer our fifth question in phase 1

(The full Code is in the notebook)

These comprehensive steps ensure the dataset is thoroughly cleaned, transformed, and primed for further analysis.

3- Exploratory Data Analysis (EDA)

Overview:

We conducted an exploratory data analysis (EDA) on the dataset using Python. The analysis aimed to gain insights into the dataset's structure, distributions, and relationships between variables. We utilized various statistical summaries and visualizations to explore the dataset effectively.

3.1 Types of Analysis Performed:

3.1.1- Data Types Identification:

Identified two types of variables in the dataset:

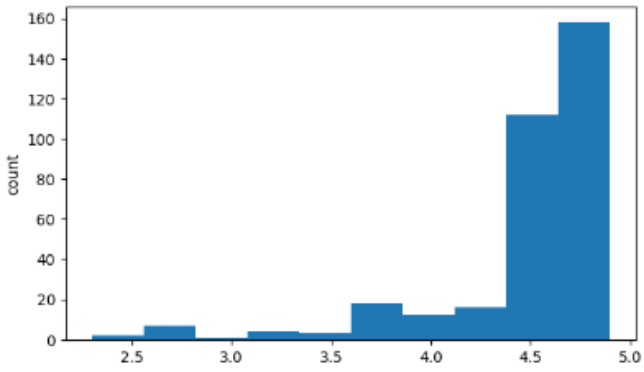
- Categorical Variables: Including 'Book Name', 'Author', 'Format', and 'Other Format'.
- Numerical Variables: Comprising 'Rating', 'Reviews', and 'Price'.

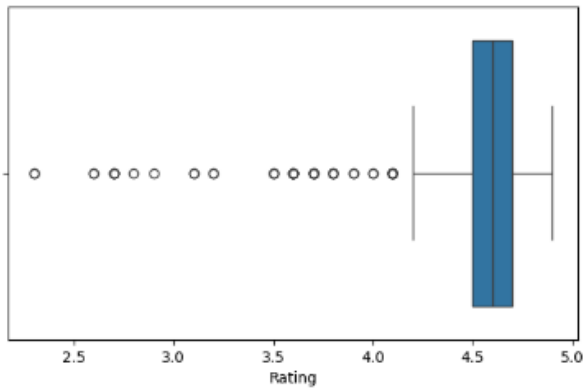
3.1.2- Statistical Summaries:

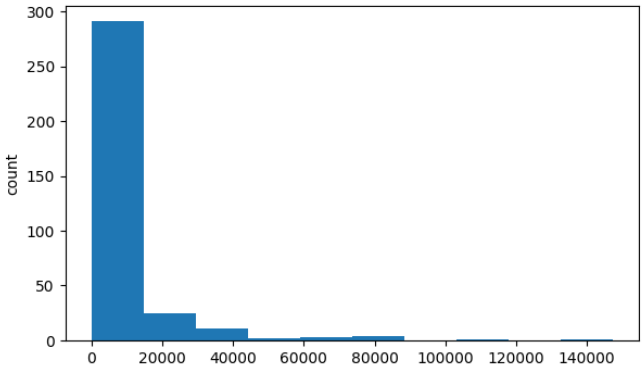
- Utilized descriptive statistics to summarize numerical variables, including mean, median, standard deviation, min, and max values.
- Calculated skewness to understand the distribution shape and detect potential outliers.

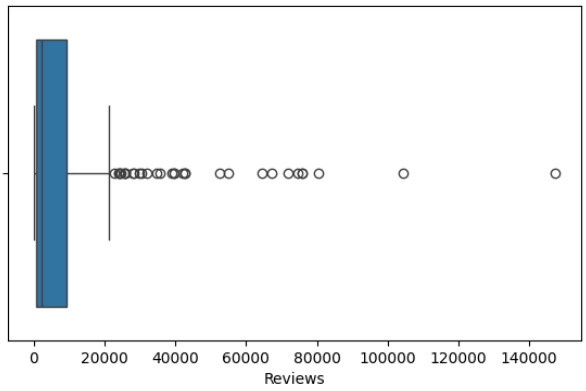
3.1.3- Visualizations:

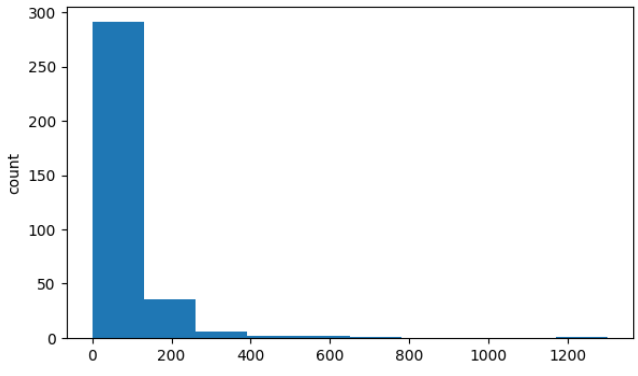
- Employed histograms and boxplots to visualize the distributions and spreads of the numerical variables(Rating, Reviews, Price).

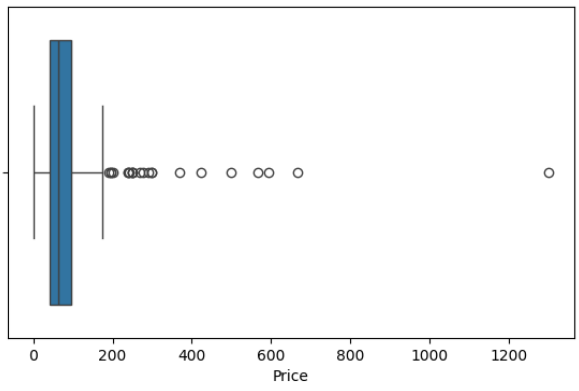






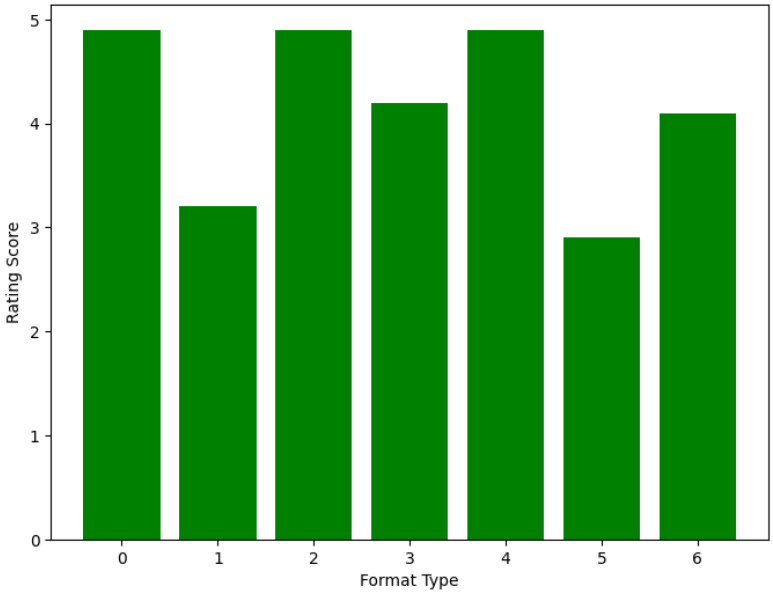






- Utilized a heatmap to visualize the correlation matrix between numerical variables, identifying relationships and dependencies.

- Used bar plots to display rating scores by book format type, uncovering potential trends or preferences.



3.2 Tools and Libraries Used:

- Programming Language: Python
- Libraries: Pandas, Matplotlib, Seaborn
- IDE: Jupyter Notebook

3.3 key Findings and Insights:

3.3.1 Skewness Analysis:

- Rating: -2.49 Negatively skewed distribution suggests that most books tend to have high ratings.
- Reviews: 4.28 Positively skewed distribution indicates that a few books receive a disproportionately high number of reviews.
- Price: 6.43 Highly positively skewed distribution suggests that only a few books are priced significantly higher than others.

3.3.2 Correlation Analysis:

- We detected the correlations between numerical variables, indicating potential dependencies. and by this heatmap we can answer one of our second questions in phase 1 which is the correlation between the price and the rating (-0.1).
- Further investigation can be done in the future to understand these relationships better.

3.3.3 Rating by Format Type:

- Identified variations in rating scores across different book format types, suggesting potential, preferences, or trends among readers. And by the bar chart we can see that "Paperback", "hardcover" and "Board Book" are the most rated formats in the dataset helping us to answer out first question in phase 1.

4- Future Steps

In the future steps of the project, there are several key areas to focus on. Firstly, we will try to make a deeper investigation to understand the factors affecting book ratings by trying to collect more data to find more information. Secondly, we will use more visualizations models between the "Price", "Format" and "Rating" to support our main goal which is optimizing our business pricing strategy for different book formats based on customer ratings and the number of customers who rated each book. Lastly, predictive modeling can be developed to forecast book ratings or sales, providing valuable insights for decision-making.