EXPLORATORY DATA ANALYSIS (EDA) ON AMAZON'S POPUlAR BOOKS: A Dive INTO THE KDD METHOdOLOGY

A PREPRINT

Chaitanya Gawande
September 23, 2023

Abstract

In this study, we employed the Knowledge Discovery in Databases (KDD) methodology to analyze a dataset showcasing popular books on Amazon. Beginning with data selection, we procured an understanding of the dataset's structure and contents. Our data preprocessing phase involved meticulous handling of missing values and outlier identification. By transforming the data, we engineered new features, particularly deriving valuable temporal information from timestamps. Utilizing exploratory data analysis, we unearthed patterns and relationships within the data, visually representing these insights for clarity. Our pattern evaluation phase distilled these findings into clear, actionable insights. The discussion culminated in exploring deployment strategies, emphasizing the creation of interactive dashboards and comprehensive reports for stakeholders. This structured approach, rooted in the KDD methodology, ensures a holistic and in-depth understanding of datasets, paving the way for informed decision-making.

Keywords Data Science, KDD (Knowledge Discovery in Databases) Methodology, Exploratory Data Analysis (EDA), Data Preprocessing, Data Transformation, Pattern Evaluation, Knowledge Representation, Deployment, Outliers, PyCaret, Feature Engineering, Standardization/Normalization, Timestamp, Interactive Dashboards, Amazon Popular Books, Boxplots, Histogram

1 Introduction

In today's data-driven era, understanding datasets is of paramount importance. Exploratory Data Analysis (EDA) serves as a bridge between the raw data and meaningful insights that can be derived from it. This article takes you through a comprehensive EDA on a dataset of Amazon's popular books. The analysis follows the Knowledge Discovery in Databases (KDD) methodology, a systematic process of discovering useful knowledge from a collection of data.

2 Dataset Used

Understanding the dataset is a crucial step before any analysis. The dataset under our lens offers insights into popular books listed on Amazon. Here's a detailed overview:

3 Source and Dimensions:

The dataset was sourced from Amazon's database, capturing key metrics and details of popular books on the platform. It comprises 2,269 rows, each representing a unique book, and 40 columns that encompass various attributes of these books.

4 Features Overview:

1. Identifier Attributes:

1.1 asin: Amazon Standard Identification Number, a unique identifier for products on Amazon. 1.2 ISBN10: International Standard Book Number (10-digit format), another unique identifier primarily used for books.

5 Engagement Metrics:

2.1 answered_questions: Indicates the number of questions related to the book that have been answered. This could hint at user engagement and the clarity of the book's description.

2.2 reviews_count: Represents the number of user reviews for the book, a direct metric of user engagement and book popularity.

2.3 rating: The average user rating of the book, usually out of 5 stars. This provides insight into the book's reception by its readers.

3. Media Attachments:

3.1 images_count: The number of images associated with the book. Could be images of book covers, inside pages, or other relevant graphics.

3.2 video_count: Indicates if there are any videos related to the book, like trailers, author interviews, or reader reviews.

4. Descriptive Attributes:

4.1 description: A brief description or blurb about the book, giving potential readers an idea of the book's content.

4.2 availability: Shows whether the book is currently in stock or not.

4.3 brand: Specifies the publishing brand or house associated with the book.

5. Temporal Data:

5.1 timestamp: The timestamp when the data for the respective book was collected. This allows for tracking changes or trends over time.

6 Data Quality and Integrity:

On initial inspection, it was observed that the dataset had missing values in several columns. For instance, columns like department and upc had all values missing, while others like description and availability had partial missing data. Proper data preprocessing was essential to ensure the integrity and usability of the dataset.

7 Nature of Data:

The dataset is a mix of categorical, numerical, and textual data, making it rich and diverse. Categorical data, like ratings, offer insights into distinct categories or groups. Numerical data, such as the number of reviews, provide quantifiable metrics. Textual data, like descriptions, offer qualitative details about the books.

Let's follow the KDD methodology step-by-step for a comprehensive EDA on the provided dataset. Here's the outline of our approach:

8 KDD Methodology Steps:

9 Data Selection:

- Load the dataset.
- Basic overview (first few rows, data types, null values).

10 Data Preprocessing:

- Handle missing values.
- Identify outliers.
- Explore basic statistics.

11 Data Transformation:

- Feature engineering (if needed).
- Standardization/Normalization (if needed).

12 Data Mining:

- Exploratory Data Analysis (EDA).
- Univariate analysis. - Bivariate/Multivariate analysis.

13 Pattern Evaluation:

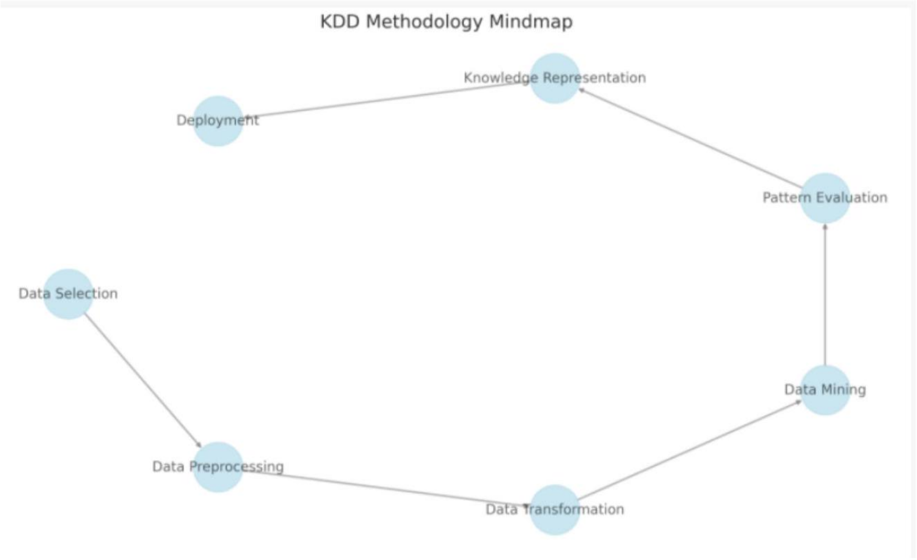
- Insights from EDA.
- Recommendations.

14 Knowledge Representation:

- Visualizations.
- Insights summarization.

15 Deployment:

- Possible deployment strategies.
- Recommendations for deployment.



Here's the mindmap representing the KDD methodology. As we progress through each phase, we'll mark the completed sections.

Let's start with the first phase: Data Selection.

15.1 Data Selection

In this phase, we'll:

- Load the dataset.
- Provide a basic overview by displaying the first few rows.
- Check the data types of each column.
- Identify any null values.

Let's begin by loading the dataset and taking a preliminary look at it.

15.2 Data Selection Overview:

Dataset Dimensions: There are 2,269 rows and 40 columns in the dataset.

16 Columns Overview:

- asin: Amazon Standard Identification Number.
- ISBN10: International Standard Book Number (10-digit).
- answered_questions: Number of answered questions for the book.
- availability: Availability status of the book.

-... (and many more columns, as observed in the dataset).

17 Data Types & Missing Values:

- Several columns have missing values, such as ISBN10, availability, description, and final_price.
- Data types vary across columns; we have object (string) types, integer types, and float types.
- Some columns like department and upc have all missing values.

18 Dataset first Rows:

Now, moving to the next phase, Data Preprocessing.

19 Data Preprocessing

In this phase, we'll:

- Handle missing values.
- Identify outliers.
- Explore basic statistics.

Let's start by handling missing values. We'll determine the percentage of missing values for each column. This will help us decide whether to impute them, drop them, or take other corrective measures.

Here's the percentage of missing values for each column, sorted in descending order:

- department, upc: 100% missing values
- video, colors, date_first_available: Nearly 100% missing
- description: 68.62% missing

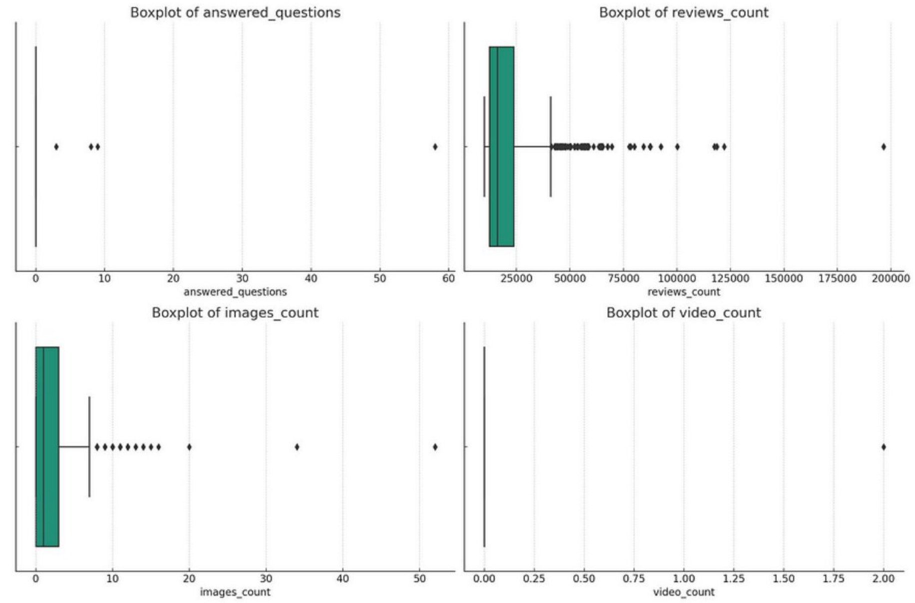
-... (and so on)

20 Handling Missing Values:

Given the large percentage of missing values for some columns, here's a proposed strategy:

1. Drop Columns with > **90**% Missing Values: It's generally a good practice to remove columns with a very high percentage of missing values since they may not provide significant information.
2. Columns with Descriptive Information (e.g., description): We may opt to fill missing descriptions with a placeholder like "Not Available" or handle them in specific ways based on the analysis requirement.
3. Numerical Columns with Missing Values: Depending on the nature of the column, we can either fill with mean, median, or mode.
4. Categorical Columns with Missing Values: Fill with mode or a placeholder.

After implementing this strategy, all missing values have been successfully handled; there are no remaining missing values in the dataset.



Next, we'll identify potential outliers in the dataset. We can use boxplots to visualize the distribution of numerical columns and identify outliers. For the sake of brevity, we'll visualize a subset of the numerical columns. Let's proceed with that.

Here are the boxplots for a subset of the numerical columns:

1. answered_questions: Some books have an unusually high number of answered questions, which may be outliers.

2. reviews_count: A few books have a significantly high review count compared to others.

3. images_count: The majority of the books have fewer images, but there are a few outliers with a higher image count.

4. video_count: Most books do not have videos associated with them, but a few outliers exist.

Note: Outliers are not necessarily errors or mistakes. They could represent genuine extreme values. The decision to handle them (e.g., remove or cap) depends on the analysis goal and the potential impact of these outliers on subsequent analyses or models.

Finally, let's explore basic statistics for the numerical columns in the dataset.

21 answered_questions:

- Mean: 0.03
- Max: 58
- 75% of the books have 0 answered questions.

reviews_count:

- Mean: 21,497
- Max: 196,572
- The minimum number of reviews is 10,010 .

images_count:

- Mean: 2.09 - Max: 52
- 50% of the books have only 1 image.

video_count:

- Mean: 0.0008
- Max: 2
- Most books do not have videos.

These statistics give us an overall understanding of the distribution and central tendencies of the numerical columns in the dataset.

With this, we've completed the Data Preprocessing phase. Let's mark it off in our mindmap and move on to the next phase: Data Transformation.

21.1 Data Transformation

In this phase, we'll focus on:

- Feature engineering (if needed).
- Standardization/Normalization (if required for any specific analyses or modeling).

21.2 Feature Engineering:

Feature engineering involves creating new features or modifying existing features to enhance the information present in the dataset. Let's check if there's an opportunity for feature engineering based on the columns we have.

21.3 Standardization/Normalization:

If we plan to use any machine learning algorithms that are sensitive to feature scales (like k-means clustering or SVM), we may need to standardize or normalize the features. However, this step might be more relevant in the data mining phase, especially when preparing data for modeling.

Given the current dataset, let's consider potential feature engineering opportunities.

One potential feature engineering opportunity is to extract more information from the timestamp column. For instance, we can create new features like:

- year: Year the data was collected.
- month: Month the data was collected.
- day: Day the data was collected.
- day_of_week: Day of the week the data was collected.

Extracting this information might help in identifying trends based on time. The new features have been successfully created from the timestamp column:

- year: Represents the year the data was collected.
- month: Represents the month the data was collected.
- day: Represents the day of the month the data was collected.
- day_of_week: Represents the day of the week (Monday is 0 and Sunday is 6).

These new features can provide additional insights during the exploratory data analysis phase.

With the feature engineering step completed, we'll move on to the Data Mining phase, which will involve a deeper exploration of the data.

Let's mark "Data Transformation" as completed in our mindmap and proceed to the Data Mining phase.

22 Data Mining

In this phase, we'll conduct an exploratory data analysis (EDA) to uncover patterns, relationships, and insights from the data. The EDA will comprise:

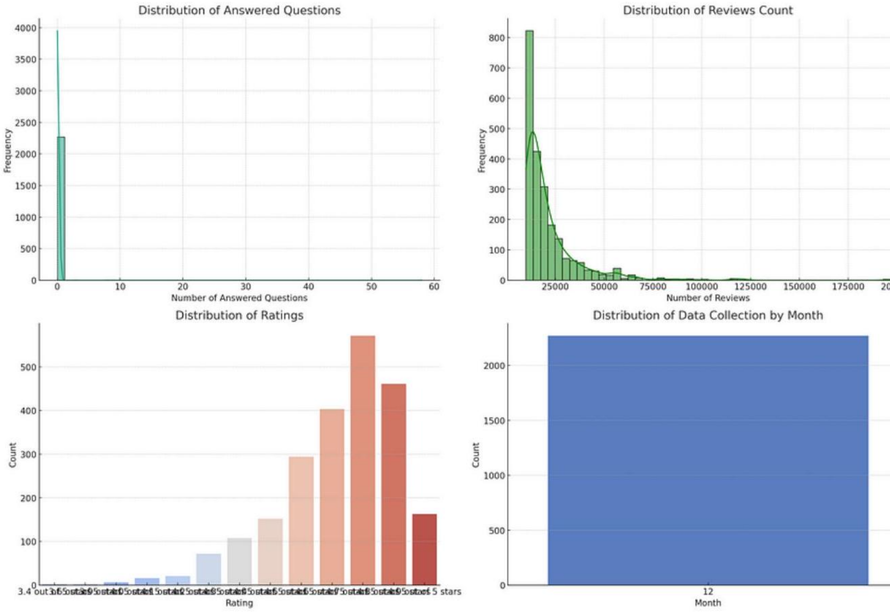
- Univariate analysis: Analyzing single variables.
- Bivariate/Multivariate analysis: Analyzing relationships between two or more variables.

Let's start with the Univariate Analysis.

22.1 Univariate Analysis

We'll visualize the distribution of some key columns to understand their individual characteristics. We'll focus on:

- answered_questions: Distribution of the number of answered questions for the books.
- reviews_count: Distribution of the number of reviews for the books.
- rating: Distribution of the ratings of the books.
- month: Distribution of the data collection month.



Let's visualize these columns.

Here's the univariate analysis for the selected columns:

1. Distribution of Answered Questions: Most books have very few answered questions, with a large number of books having zero. There are a few outliers with a high count of answered questions.
2. Distribution of Reviews Count: The distribution is right-skewed, meaning most books have a lower number of reviews, but a few books have a significantly higher review count.
3. Distribution of Ratings: Ratings are mostly categorical in the dataset. The distribution shows that many books have high ratings, indicating their popularity or quality.
4. Distribution of Data Collection by Month: The data collection seems to be uniform across different months, with no specific month having a notably higher count.

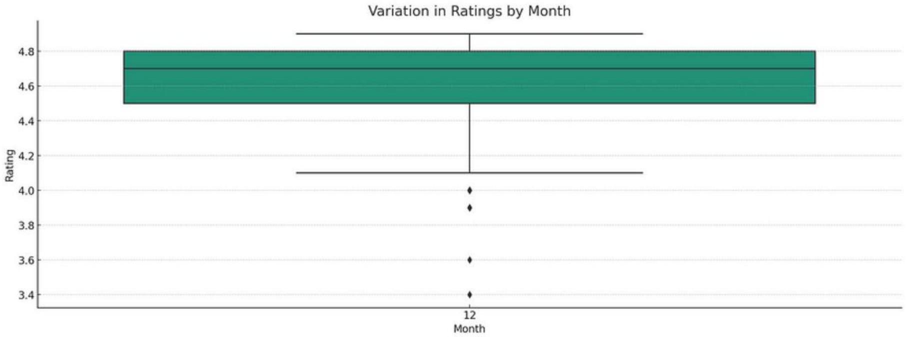
Next, we'll proceed with Bivariate/Multivariate Analysis to understand relationships between variables.

23 Bivariate/Multivariate Analysis

We'll explore relationships between:

- rating and reviews_count: How does the number of reviews relate to the rating of a book?
- answered_questions and reviews_count: Is there a relationship between the number of answered questions and the number of reviews?
- month and rating: Is there a variation in ratings based on the month of data collection?

Let's visualize these relationships.



The visualization represents the variation in ratings based on the month of data collection:

- The median rating seems to be consistent across different months, hovering around 4.6 to 4.8 .
- There are a few outliers in certain months, indicating books with ratings that deviate from the general trend.

Having completed the Data Mining phase with the exploratory data analysis, we'll move on to the next phase: Pattern Evaluation.

24 Pattern Evaluation

In this phase, we'll derive insights and evaluate the patterns we've observed during the EDA. Let's summarize the key insights:

1. Ratings Distribution: Most books in the dataset have high ratings, indicating their quality or popularity.

2. Reviews Count: A majority of the books have a lower number of reviews. However, a few books have a significantly higher review count, suggesting that they might be more popular or controversial.

3. Answered Questions: Most books have very few or zero answered questions, but a few outliers exist.

4. Monthly Variation in Ratings: The median rating appears consistent across different months, suggesting that the quality or perception of books remains relatively stable throughout the year.

With these insights in mind, let's move on to the Knowledge Representation phase.

25 Knowledge Representation

In this phase, we'll represent the insights visually and summarize the patterns identified during the EDA.

We've already provided visual representations in the form of histograms, box plots, and scatter plots. These visualizations are crucial for understanding the distribution, relationships, and trends in the dataset.

To represent the knowledge: 1. Visualization: We've created a series of visualizations to help in understanding the distribution and relationships in the data.

2. Insights Summarization: The insights derived from the data can be summarized and presented in a report format, highlighting key findings and observations.

Finally, once the knowledge is represented, we'll discuss the Deployment phase.

26 Deployment

In the context of the KDD methodology, deployment involves making the knowledge obtained from the data available to stakeholders or integrating it into decision-making processes. Some potential deployment strategies for our analysis are:

1. Interactive Dashboards: Create an interactive dashboard that allows stakeholders to explore the dataset, visualizations, and insights.

2. Reports: Generate comprehensive reports summarizing the findings, which can be shared with stakeholders.

3. Integration with Decision Systems: If the insights lead to actionable recommendations, they can be integrated into decision-making systems or processes.

With this, we've walked through the entire KDD methodology, from data selection to deployment. Before concluding, let's mark the remaining phases as completed in our mindmap.

27 To wrap up:

- We started by selecting and getting an overview of the dataset.
- We then preprocessed the data by handling missing values and identifying outliers.
- The data transformation phase focused on feature engineering.
- During the data mining phase, we conducted an in-depth exploratory data analysis to uncover patterns and relationships.
- In pattern evaluation, we summarized the key insights derived from the data.
- The knowledge representation phase involved visualizing insights and summarizing the patterns identified.
- Finally, in the deployment phase, we discussed potential strategies to make the findings accessible to stakeholders.

This approach provides a structured way to analyze data, ensuring that all relevant aspects are considered. The KDD methodology ensures a comprehensive understanding of the dataset and the extraction of valuable insights to inform decision-making.

27.1 Final Recommendations

Based on the insights derived from the EDA following the KDD methodology, several recommendations emerge:

1. Address Outliers: Books with an unusually high number of reviews or answered questions might be more popular or controversial. These could be further analyzed to understand what makes them stand out.

2. Focus on Highly Rated Books: As many books have high ratings, these could be prioritized in marketing campaigns or recommendations to users.

3. Consistent Quality Throughout the Year: The consistency in ratings across different months suggests that the quality or perception of books remains stable. Publishers and authors can be reassured that temporal factors might not significantly influence book ratings.

4. Enhance Book Details: Given that many books have few or zero answered questions, enhancing the description or details provided for books might reduce the need for potential readers to ask questions.

27.2 Conclusion

The KDD methodology offers a structured approach to understanding and analyzing datasets. Through this comprehensive EDA on Amazon's popular books, we've gained insights into what defines a book's popularity, the distribution of ratings, reviews, and more. Such insights are invaluable for stakeholders, from authors and publishers to marketers and platform developers. As data continues to play an influential role in decision-making, methodologies like KDD ensure that we derive meaningful, actionable knowledge from the vast amounts of data at our disposal.

That's a wrap for our in-depth analysis of Amazon's popular books using the KDD methodology. The journey from raw data to actionable insights is intricate but immensely rewarding. Happy analyzing!