

**Data Analysis of Book Information**

Group – 10

**Team Members:**

1. **Ashoka Chakravarthy Sanjapu**
2. **Tejas Reddy Vepala**

**Table of Contents**

Abstract……………………………………………………………………………………………………….…………………………….2

Introduction………………………………………………………………………………………….……………………………………2

Data Description ………………..…………………………………………………………………….……………………………….4

Correlation between Variables…………………………………………………………………………………………………..6

Exploratory Data Analysis ………………………………………………………………………………………………………….7

Models……………………………………………………………………………………………………..……………………………..17

Model Evaluation………………………………………..…………………………………….…………………………………….20

Conclusion…………………………………………………………………………………………………………………………….…24

References……………………………………………………………………………….………………………………………………25

**Abstract:**

The Amazon Books dataset provides a wealth of data about an extensive collection of books available on one of the biggest e-commerce platforms in the world. The project offers a thorough analysis of this dataset that considers a number of facets of the publishing business, such as authorship, reader reviews, pricing, and publication dates. The most important aspect of every book dataset is authorship. We explore the contributions of prolific writers and their impact on the book market. Additionally, so as to shed knowledge on the variables influencing readers' preferences, we look into the connection between author reputation and book ratings. The discovery and selection of publications is greatly influenced by user ratings and reviews. We examined the dataset to see how user evaluations, reviews, and sales of books are related. The study will throw light on the dynamics of user-generated material and how it impacts book sales.

**Introduction:**

This project's main goal is to reveal the Amazon Books dataset's hidden patterns, trends, and correlations. We seek to provide a thorough insight of the complicated book market by evaluating the dataset's variables, including Name, Author, User Ratings, Reviews, Price, and Year. We believe that this analysis will assist in shedding light on an array of issues that authors, publishers, marketers, researchers, and avid readers alike are interested in. During the last few years, the publishing industry has seen considerable chaos, partly fueled by the rise of e-commerce behemoths like Amazon and the digitization of books. Understanding the changing patterns of book sales, reader preferences, and market trends has never been more important than it is right now, when consumers have unprecedented access to books from around the world. To meet this demand, we conducted a thorough examination of the Amazon Books dataset, a goldmine of data containing details about an extensive range of books, their authors, user ratings, reviews, pricing, and releasing years. The publishing industry has changed as a result of Amazon, a pioneer in e-commerce, which has also revolutionized the way books are bought and sold. The website has a vast library of books that covers several genres, languages, and cultures. This dataset offers a singular opportunity to obtain insights into the world of literature and reading habits by reflecting the worldwide literary ecosystem. The importance of authors to the literary world cannot be emphasized. In this study, we identify authors who produce a lot of work, examine their contributions to the Amazon Books marketplace, and evaluate how their writing impacts the book industry. In order to understand how well-known authors perform compared to developing authors, we will investigate the connection between author reputation and book ratings. The book business is dynamic and constantly changing. We will look into how Amazon's book distribution has evolved over time while maintaining an eye on the publishing industry changes. We seek to comprehend the dynamics of modern major films and the lasting popularity of literary classics by examining whether there is an association between book success and the year of publication. This project provides light on the intricate web of variables that affects book sales, reader preferences, and the always changing literary scene through a thorough exploration of the Amazon Books dataset. We cordially encourage writers, publishers, booksellers, academics, and book lovers to join us as we explore this approach to acquire more understanding of the complex world of books and reading. This analysis will offer useful data to help you make decisions and enhance your understanding of literature, whether your goals are to write more effectively, promote your books more successfully, or simply find your next favourite book to read.

**Data Description:**

The Amazon Books dataset is a comprehensive informational source which comprises an extensive range of books provided on the Amazon marketplace. Name, Author, User Ratings, Reviews, Price, and Year are the six main elements. We provide great depth on each variable in this section, in addition to the significance it is to the dataset.

**Name:** The name of the book is given by the "Name" variable. Each book in the dataset has it as its main identifier. The core and subject matter of a book frequently appears in its title, making it essential to book discovery and identification. The dataset might include variations in book names, such as subtitles and editions, which should be taken into consideration when doing an analysis.

**Author:** The name of the author or authors who wrote the book can be found in the "Author" variable. We can look at the contributions made by different writers to the Amazon Books marketplace thanks to this variable because authorship is a crucial component of a book's success. The dataset may contain books produced by a single author or by many co-authors.

**User Ratings:**

The "User Ratings" column displays the typical evaluations that consumers and critics have given the book. A book's overall quality and reader popularity can be determined by its user ratings. Typically, ratings range from 1 to 5, with higher ratings indicating more satisfied readers.

**Reviews:**

The number of user-generated reviews for each book is tracked by the "Reviews" variable.

Reviews offer comprehensive input from readers and insights into a book's benefits and drawbacks. Potential readers' trust in a book's ratings and their choice to buy or read it can be affected by the volume of reviews.

**Price:** Price indicates the price to buy the book on the Amazon platform, usually in the local currency. Pricing has an important effect on both sales and profitability when deciding whether or not to purchase a book. The format (hardcover, paperback, ebook), edition, and availability of a book are only a few instances of the factors that might affect price.

**Year:** The book's year of release is indicated by the "Year" variable. An important component of a book's information is its publication year, which affects its audience's appeal and relevancy. Through assessing publishing industry changes, this variable enables us to examine how the distribution of books on Amazon has evolved over time.

It's important to remember that the Amazon Books dataset may include a wide variety of book genres, languages, and editions, making it an accurate depiction of the the globe's literature. To provide accurate and valuable insights, future data cleaning, normalization, and handling of missing information should be taken into consideration during the data analysis procedure. In the context of a large online marketplace, this dataset offers a singular chance to investigate the intricate relationships between book titles, authors, reader reviews, cost tactics, and release years. We may learn important things regarding the dynamics of the book business by examining these factors. This understanding will aid authors, publishers, marketers, and readers in making wise choices and navigating the ever-shifting landscape of literature.

**Correlation between Variables:**

A correlation plot of numerical data, commonly called a heatmap, is a form of visualization. The primary goal is to show the direction and intensity of correlations between different numerical variables. The plot's name, "Correlation Plot of Numerical Variables," suggests that the goal is to find linear correlations between the data displayed. The following variables are shown on the heatmap: "User Rating," "Price," and "Year." The cell's color saturation and any associated numbers stand in for the correlation coefficients between these variables. The relationship is -0.14 in strength. This indicates that user reviews and pricing have a shaky negative linear relationship. In simpler terms, a product or item's rating among consumers slightly declines when its price rises, and vice versa. The correlation coefficient is 0.8. This indicates a weak positive link between year and user ratings. This may suggest that user ratings gradually rise over time. The correlation coefficient in the present instance is -0.044, indicating a tiny difference between the two variables.

A graph of a number of numerical variables

Description automatically generated

**Exploratory Data Analysis:**

shows a table with details on five books, including their titles, authors, ratings by users, number of reviews, costs, and years of publication. Books by authors like J.J. Smith, Stephen King, and the writer George Orwell are among the titles.

A screenshot of a computer

Description automatically generated

The program checks a dataset called "data" for any null values. The output lists how many values are missing for each column: "Author" has 124, "User Rating" and "Reviews" each have 75, and "Price" has 67. Rows holding these null values are eliminated by a later line of code.

A screenshot of a computer program

Description automatically generated

Using a dataset and the **describe()** method, perhaps from the Python Pandas package. The User Rating, Price, Year, and Price/Rating columns in the table each have descriptive statistics given to them.

The following statistics are shown for each of these columns:

* count: Total number of values that are present.
* mean: A typical value.
* std: The measure of variability is the standard deviation.
* min: The bare minimum.
* 25%: The value below which 25% of the data falls, or the 25th percentile.
* 50%: The median or the midpoint.
* The value below which 75% of the data falls is known as the 75th percentile.
* max: The highest value.

A table with numbers and letters

Description automatically generated

A horizontal bar chart labelled "Books that appear 5 or more times" can be seen in the plot. The chart seeks to display how frequently particular books come up, maybe indicating their popularity or frequent occurrences in a dataset. The x-axis shows the number of times each book has appeared, and the bars, which have different colors of blue, represent various books. With its bar stretching closest to the 10 point on the x-axis, the "**Publication Manual of the American Psychological Association, 6th Edition**" appears to have the highest frequency. StrengthsFinder 2.0, Oh, the Places You'll Go!, and "The 7 Habits of Highly Effective People: Powerful Lessons in Personal Change" are three other well-known titles that are referenced between 8 and 9 times each. This list also includes well-known novels like "To Kill a Mockingbird" and beloved children's books like "The Very Hungry Caterpillar". The presence of inspirational books is evident in titles like "The Four Agreements: A Practical Guide to Personal Freedom (A Toltec Wisdom Book)" and "Jesus Calling: Enjoying Peace in His Presence (with Scripture References)".

A screenshot of a computer

Description automatically generated

Authors that appear more than 5 times are shown in a horizontal bar chart in the image. This visualization appears to draw attention to the frequency with which specific authors appear in a dataset, providing a clue as to their fame or popularity in the appropriate setting. The scale on the x-axis ranges from 0 to 12, and each horizontal bar represents an author. The length of the bar indicates the number of times the author has been cited. With his bar achieving the maximum score of 12, "Jeff Kinney" stands out as the author that is mentioned the most frequently. The bars behind other well-known authors, such as "Rick Riordan" and "Suzanne Collins," are almost as long as Jeff Kinney's. The presence of prominent individuals like "Gary Chapman," "American Psychological Association," and "Gallup" in the list also serves to highlight their significance to the dataset. Being included on the list of well-known children's writers like "Dr. Seuss" and "Eric Carle" demonstrates their ongoing appeal. Additionally, modern writers like "Stephenie Meyer," "Dav Pilkey," and "E L James" are featured, showing their current relevance. Famous people like "J.K. Rowling" and "Harper Lee" are also included on the list, highlighting the broad range of authors that are there. That some of the authors or organizations, such "The College Board" and "[Al Sweigart]," appear to be more specialized indicates that the dataset may include an extensive variety of book genres or subjects.

A red and white graph

Description automatically generated

The plot displays a vertical bar chart that shows the number of items—likely books or events—over a number of years, from 2009 to 2019. The depiction of each year is a distinctly colored bar, which makes it easy to differentiate between the many years. The light green bar on the left represents the year 2009, and as we travel to the right, the subsequent bars indicate each following year up to the lightest purple bar, that represents the year 2019. One can infer the count for each year based on the height of the bars using the y-axis, called "count," that displays numerical values from 0 to just above 100. Each bar appears to be hanging between 80 and 100 counts, which suggests a regular and relatively high frequency of items for each year. A consistent and stable collection or publication rate over the course of this decade is shown by the absence of significant rise or drop trend from year to year. Additionally, by using a different hue for each year, the chart is not only pleasing to the eye but also makes it easy to tell which year is which. The gradual transition of shades from cool to warm to cool helps direct the viewer's eye from the start to the end of the data series.

A graph of different colored vertical lines

Description automatically generated

The plot shows a scatterplot with the headline "Top Authors and their Books/years." It depicts the total number of books published by different writers over the years 2009 to 2019. We find a list of prominent writers on the y-axis with the label "Author," including individuals like Suzanne Collins, J.K. Rowling, Harper Lee, Gary Chapman, and the American Psychological Association. The study years appear sequentially on the "Year"-labeled x-axis. Each color on the graph indicates a different number of books, and the data points are shown as colored circles. The color scheme is explained using the key on the right:

* 1 book is denoted in dark blue.
* Green means 2 books.
* Light blue denotes 3 books, whereas purple denotes 4 books.
* Orange stands for an outstanding "6 books."

A screenshot of a computer screen

Description automatically generated

The audience gets an immediate visual clue from the color-coded system about how prolific each author was in a certain year. For instance, it's obvious that Gary Chapman published a book in 2010 called "The 5 Love Languages: The Secret to Love That Lasts" (shown in dark blue). The majority of these top authors have constantly published throughout the decade, with a few peak years where their publication count was especially high, as is clear from the graph with a quick scan. The dynamic component of this visualization is an unique trait. When you hover over a data point, a pop-up box with more information on the number of books, the year they were released, the author's name, and the title of the book or books displays.

The histogram in the plot is labeled "Histogram of Users Rating." In this instance, a histogram showing the distribution of user ratings is used to illustrate the distribution of a dataset. The "User Rating," that uses a scale from 0 to 5, appears on the x-axis. The "count," displaying the number of occurrences for each rating value, is shown on the y-axis. The bars on the chart are primarily an inviting shade of blue, and as you move from left to right, you can see a clear pattern. The graph's leftmost side, where ratings are lowest (between one and 2.5), has the smallest amount of representation. Around the 3 and 3.5 points, representation slightly improves. From rating 4 onward, the distribution sharply rises, reaching a clear peak between ratings of 4.5 and 5. To achieve this peak, a significant percentage of feedback from consumers must fall inside this upper range. A small pop-up window that appears when the cursor is over a particular bar gives a more thorough breakdown by displaying the precise rating and its count. For instance, one of the top bars' pop-up boxes displays a "User Rating" of 4.6 and a corresponding "count" of 191, suggesting that 191 users gave the bar a rating of 4.6. This histogram suggests the majority of customers gave favorable ratings, especially between 4.5 and 5. This shows a broad pattern of customer satisfaction with or favorable attitudes about the good, service, or literature being reviewed. Even while there are a few worse reviews, they are considerably outnumbered by their higher substitutes, emphasizing the users' overall good reaction.

A graph with blue and white bars

Description automatically generated

The graph displays a scatter plot labeled "User Rating vs. Price," which compares the correlation between user reviews and book pricing. We have "Price," which goes along the x-axis from 0 to approximately 250, and "User Rating," which runs down the y-axis from 1 to 5. A book is symbolized by each of the many data points that are scattered across the storyline. A legend to the right names these authors, including popular ones like "JJ Smith," "Stephen King," "George Orwell," "George R. R. Martin," and "Veronica Roth," among others. These points are color-coded, representing various authors. Regardless of price, an interesting pattern is that the majority of the books fall in the 4 to 5 star grouping. This cluster ranges in price from the lowest end near zero to the mid-ranges around fifty and one hundredth. Particularly between the price range of 0 and 50, there is a dense concentration of points, indicating that there are more books in this more affordable range. In the higher price ranges, especially between 100 and 250, there are fewer amounts sent apart. The ones that do fall within this pricing range, however, maintain comparatively good customer ratings, usually between 4 and 4.5. The portrayal is not as extensive in the user rating range of 2 to 3.5. Most of those that relate within this category are in the price range of 0 to 50, with a few outliers going up to 100. The apparent absence of books with the lowest user ratings (1 to 2) and the higher price range (over 100) is one of the significant findings. This implies that more costly materials typically guarantee a certain degree of appeal or quality, preventing them from earning lowest reviews.

A graph with many colored dots

Description automatically generated

The displayed graph is a scatter plot called the "Price Reviews Scatter Plot." The horizontal axis represents "Price," ranging from 0 to nothing over 250, and the vertical axis indicates an unnamed value or metric (maybe reviews or ratings). This graphic shows the connection between the two variables. A more detailed examination reveals that the data points, which have a distinctly yellow colour, clustered near the top of the y-axis, suggesting a higher value or metric for goods or things within the lower price range. The density of the data points noticeably declines as we move from left to right, indicating a rise in price. The substantial amount of products or items with lower prices earning higher evaluations or ratings is indicated by the dense cluster of points in the top-left corner. There is a vertical concentration of data points between the price range of 0 to 50, indicating that many things are concentrated inside this range with variable assessments or ratings. The concentration of points starts to spread out more horizontally as the price rises from 50 to about 150, suggesting a more uniform distribution of reviews or ratings in this price range.

A chart with a green dot

Description automatically generated with medium confidence

**Models:**

For regressive instances, linear regression is a fundamental and widely used supervised machine learning approach. By fitting a linear equation to the observed data, one can model the relationship among a dependent variable (the target) and a number of independent variables (features). One independent variable and a straight line are employed to Using simple linear regression, simulate the relationship between the aim and the independent variable. There are many benefits to linear regression. It is interpretable in the sense that the connections between each feature and the aim are simple to comprehend. In addition, it has good computational speed and handles both small and large datasets with ease. Both univariate and multivariate regression issues are solvable with linear regression. But there are also drawbacks to linear regression. It relies on the assumption that the relationship between the features and the target is linear, which may not always be the case in real-world situations. It could fail to identify intricate, non-linear correlations in the data, and it might be sensitive to outliers.

A screen shot of a computer code

Description automatically generated

The dataset is divided into training and testing sets using the function **train\_test\_split**. The features (independent variables) are represented by X, while the target (dependent variable) is represented by Y.

Due to the argument test\_size=0.2, 20% of the data will be used for testing, and the remaining **80% for training**. Using random\_state=42 makes sure the split may be repeated. The random number generator is given a seed, ensuring that the same subset is created each time the code is executed. **LinearRegression()** puts up a linear regression model. By doing this, a LinearRegression class instance has been generated, ready for data training. To train the model, use the fit strategy. For it to teach the algorithm the parameters, training data (X\_train and y\_train) has to be provided.

**K-Nearest Neighbors:**

The K-Nearest Neighbors (KNN) algorithm is a supervised machine learning method that is mostly employed for classification and regression applications. It is an algorithm that is simple but efficient and can be used to both kinds of issues. KNN is regarded as a non-parametric and instance-based learning method because it relies only on the actual data points in the training set and doesn't make any major assumptions about the underlying data distribution. The primary objective of KNN is to forecast a data point's class or value using the majority class or average of its k-nearest neighbors in the feature space. The "k" in KNN stands for the number of nearest neighbors had a bearing on the prediction. The algorithm calculates the distance (usually the Euclidean distance) between a new data point and every other point in the training set prior to making a prediction for that point. The prediction follows by using either majority voting (for classification) or averaging (for regression) among the k-nearest neighbors with the shortest distances. The ease and ease of use of KNN is one of its benefits. It is a lazy learner because it doesn't need instruction or model fitting. KNN is able to handle non-linear decision boundaries and is noise-resistant.

A screenshot of a computer program

Description automatically generated

The scikit-learn library's k-nearest neighbors (KNN) regression model is used in the example code provided. With the KNN algorithm, an instance-based learning technique, a value is predicted based on how much it resembles a group of nearby points in the dataset. Based on the average of its 'k' nearest neighbors' outputs, it makes a continuous output prediction in this regression context. The required module KNeighborsRegressor is initially loaded from **sklearn.neighbors**. The **n\_neighbors** option is then set to 5 when creating an instance of the KNN regressor. This means that when making predictions, the algorithm will take the five closest points in the training dataset into account. Once the model has been initialized, it is trained using the fit technique on the supplied data, where X\_train denotes the goal values and y\_train the characteristics of the training data. Once the model has been properly initialized, it is trained using the fit technique on the information that has been provided, where X\_train denotes the goal values and y\_train indicates the features of the training data. Once trained, the model can be utilized to make recommendations based on fresh, unexplored data. The test dataset (X\_test) uses the predict method to estimate target values. The variable y\_pred holds the forecasts.

**Model Evaluation:**

In the discipline of machine learning, evaluating a model is a basic step that serves as a benchmark for evaluating how well it can predict future outcomes. Its main goal is to examine a model's ability to generalize by evaluating how well it performs on test or unobserved information. By doing this, it aids in determining whether a model's predictions are accurate and trustworthy, ultimately directing decision-making and potential improvements. It is essential to compare various models or algorithms so as to select the one that performs the best. Additionally, it is essential for identifying overfitting (when a model performs well on training data but poorly on fresh data) and underfitting (when a model performs poorly on both training and test data). Depending on the type of task, such as classification or regression, several evaluation metrics are used to measure model performance. Metrics including accuracy, precision, recall, F1 score, mean squared error, and R-squared are often used. A model's advantages and drawbacks can also be fully understood by using assessment methods like cross-validation, confusion matrices, and learning curves. In order to make sure that machine learning efforts are knowledgeable, impartial, and effective, the choice of evaluation metrics and approaches is dependent on the particular problem, dataset characteristics, and project objectives.

A screenshot of a computer error

Description automatically generated

The process of calculating the Mean Squared Error (MSE), a popular statistic for regression issues. It computes the squared difference between the true values (y\_test) and the predicted values (y\_pred) of the model. The squaring guarantees that each term is positive, giving bigger errors more weight. In essence, a smaller MSE indicates greater agreement between the model and the data. The Root Mean Squared Error (RMSE) is obtained after the MSE calculation. The advantage of this being in the same units as the target variable is that it is simply the square root of the MSE. A simple, comprehensible indicator on how much, on average, the model's predictions leave from the actual values is provided by RMSE. Last but not least, the method determines the R-squared value (abbreviated as R2), a further significant indicator for regression models. The amount of the dependent variable's variance that can be predicted from the independent variables is determined by R-squared. A better model fit can be seen by higher values, which range from 0 to 1. An R-squared of 1 implies that the model adequately adjusts for the results' variability.

A K-Nearest Neighbors (KNN) regression model is used to illustrate how a predictive model gets assessed. KNN, a non-parametric method, forecasts unlabeled data points using a set of labeled data points. It is important to understand the model's accuracy and reliability during this evaluation phase since the model's capacity to generalize is measured by how well it works on fresh, untested information.

A screenshot of a computer code

Description automatically generated

The Mean Squared Error (MSE) is computed before evaluation. This measure takes the squared difference between the true values, denoted by y\_test, and the anticipated value from the model, indicated by y\_pred, and averaged it. The measure ensures that every value are positive and penalizes greater errors more severely than smaller ones by squaring these disparities. A single value that describes the typical prediction error is what the MSE is about. The approach then calculates the Root Mean Squared Error (RMSE) by taking the MSE's square root. In terms of interpretability, RMSE has a significant advantage over MSE. It expresses the model's average error magnitude in the dependent variable's unit, making it simpler to relate to. The coefficient of determination, also known as the R-squared value, is calculated at the final stage. The R-squared metric calculates the amount of the dependent variable's variance that the independent variables can be held accountable for. It essentially describes the extent to which the model fits the data. A score of 0 shows that the model does not explain any of the variance, whereas an R-squared value of 1 would imply a perfect match. According to the output, the Mean Squared Error for the K-Nearest Neighbors Regression Model is approximately 1043.13. Approximately 87.635% of the variance in the dependent variable is explained by the model, based to the R-squared value of 0.87635. This is quite high and suggests the KNN model fits the data well.

A screenshot of a computer code

Description automatically generated

using a K-Nearest Neighbors (KNN) regression model as a practical application of predictive analytics. Here, forecasting book prices using a set of predefined attributes is the major goal. Using the pd.DataFrame method from the pandas package, a new dataset with the name new\_data is generated. Each of the three data points (books) in this dataset has three features: "User Rating," "Reviews," and "Year." For instance, the 2015 release of the first book had a user rating of 4.5, 500 reviews, and 500 ratings. Similar to the first book, the second and third novels have unique features. The code first defines this dataset then predicts the prices of these books using the knn\_model.predict method. An array of anticipated prices given back by the predict method, which takes the new\_data as an input. The output provided displays predictions for each book. It is expected that the first book will run you $22.08, the second book $15.19, and the third book $34.61.

**Conclusion:**

In conclusion, this study focused on analyzing and predicting book prices using a dataset that included information about books such user ratings, reviews, year, authors, and names. We began a comprehensive tour through the worlds of exploratory data analysis, machine learning modeling, and data preprocessing. For us to make sure that the data was in a format that was appropriate for modeling, our data prior to treatment processes included dealing with missing values, cleaning, and transforming the data. Exploratory data analysis revealed significant patterns and trends within the dataset, providing useful knowledge about the connections between variables and allowing data-driven decisions. In conclusion, this experiment showed the effectiveness of statistical analysis and machine learning for forecasting book prices. KNN showed its adaptability by capturing complicated correlations in the data, whereas Linear Regression served as a reliable baseline. Our predictions' reliability and precision depended heavily on the model evaluation and model selection processes. This project's knowledge and expertise can be used in real-world book company environments to inform pricing decisions and decision-making processes.

**References:**

Analysis For The System Recommended Books That Are Fetched From The Available Dataset, 4th

International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli,

India, 2022, pp. 1801–1804, doi: 10.1109/ICSSIT53264.2022.9716522. S. Nagendram, M. S. R. K. Nag, S. H. Ahammad, K. Satish, and K. Saikumar.

BRAD 1.0: Book Reviews in Arabic dataset, 2016 IEEE/ACS 13th International Conference of

Computer Systems and Applications (AICCSA), Agadir, Morocco, pp. 1-8, doi:

10.1109/AICCSA.2016.7945800.

Combining Linear Regression Models, Zheng Yuan and Yuhong Yang, Journal of the American

Statistical Association, 100:472, 1202-1214, DOI: 10.1198/016214505000000088, 2005.

Evaluating KNN Performance on Dataset," 12th International Conference on Computational

Intelligence and Communication Networks (CICN), 2020, Bhimtal, India, pp. 60–62, doi:

10.1109/CICN49253.2020.9242568. D. Bajpai and L. He.

Amazon books reviews recommendations <https://paperswithcode.com/dataset/amazon-review>

Xie X, Yukun C, and Haoyun L. VR-Explorer: A Virtual Reality Interactive Data Exploration System

in Demonstration. Companion Proceedings of the ACM International Conference on

Supporting Group Work in 2023. (15-17)