* 1. The pandas library is used to read the CSV file. This library provides powerful data manipulation tools in Python. In the Book class, the readCSV method utilizes the pd.read\_csv(self.path) function to read the CSV file located at the specified path.   
     *“  
     def readCSV(self):   
      bookInformation = pd.read\_csv(self.path)   
      return bookInformation  
     ”*
  2. The books.head() method from the pandas library is used to display the first 5 rows of the DataFrame. This is useful for getting a quick look at the data structure and ensuring the data has been loaded correctly. **Shows initial entries of books with details bookID, title, authors, average\_rating, isbn, isbn13, language\_code, num\_pages,ratings\_count and text\_reviews\_count.**“*print("First 5 rows of the DataFrame")*

*print(books.head())  
”*

* 1. The books.shape attribute returns a tuple representing the dimensions of the DataFrame (number of rows, number of columns). This provides a quick overview of the dataset size.   
     **The DataFrame has 13,719 rows and 9 columns, indicating a substantial dataset for analysis.**  
     *“  
     print("Number of rows and columns in the DataFrame")   
     print(books.shape)  
     “*
  2. The books.describe() method generates descriptive statistics that summarize the central tendency, dispersion, and shape of the dataset’s distribution, excluding NaN values. There are 7,605 unique authors. The most frequent author is Agatha Christie, appearing 69 times.  
     **Displays statistical details:**
* **count: Number of entries for each column.**
* **mean: Average values.**
* **std: Standard deviation, indicating variability.**
* **min and max: Minimum and maximum values.**
* **Quartiles (25%, 50%, 75%): Distribution spread.**

“  
*print("Summarize Data")   
print(books.describe())*  
“

* 1. The books['authors'].describe() method provides descriptive statistics specifically for the 'authors' column, including the count of unique authors and the most frequent author.  
     **There are 7,605 unique authors. The most frequent author is Agatha Christie, appearing 69 times.**  
     *“  
     print("Number of Unique Authors in the dataset and most Frequent Author") print(books['authors'].describe())  
     “*
  2. The dependent variable ‘dependentVariable’ is set to the 'average\_rating' column, which we aim to predict. The independent variables independentVariables are the columns 'num\_pages', 'ratings\_count', and 'text\_reviews\_count'. The stsm.add\_constant(independentVariables) function adds an intercept to the model, necessary for the OLS regression. The stsm.OLS(dependentVariable, independentVariables).fit() method fits the OLS regression model. The olsModel.summary() method displays the model's summary, including statistics like R-squared, F-statistic, coefficients, standard errors, and p-values, helping understand the relationship between dependent and independent variables.
* **R-squared**: 0.029, indicating the model explains 2.9% of the variance in average ratings.
* **Coefficients**:
* num\_pages: Positive impact on average rating (significant with p-value < 0.000).
* ratings\_count: Positive but very small impact (significant with p-value = 0.023).
* text\_reviews\_count: Not significant (p-value = 0.916).
* **Conclusion:** While num\_pages and ratings\_count have a statistically significant impact on average\_rating, the overall model explains a small portion of the variability, suggesting other factors also influence book ratings. The large condition number suggests potential multicollinearity issues, indicating that some independent variables might be highly correlated.

*“  
dependentVariable = books['average\_rating']  
independentVariables = books[['num\_pages', 'ratings\_count', 'text\_reviews\_count']]  
independentVariables = stsm.add\_constant(independentVariables)  
olsModel = stsm.OLS(dependentVariable, independentVariables).fit()  
print(olsModel.summary())  
“*