**Project:** Book Rating Prediction Model

**Course:** Python Machine Learning Labs

DSTI

Omar AIT ALI

**Project Summary:**

The dataset provided is a curation of Goodreads books based on real user information. It can be used for many tasks like predicting a book’s rating or recommending new books.

Description the columns of this Dataset:

**1) bookID:** A unique identification number for each book.

**2) title:** The name under which the book was published.

**3) authors:** The names of the authors of the book. Multiple authors are delimited by “/”.

**4) average\_rating:** The average rating of the book received in total.

**5) isbn:** Another unique number to identify the book, known as the International Standard Book Number.

**6) isbn13:** A 13-digit ISBN to identify the book, instead of the standard 11-digit ISBN.

**7) language\_code:** Indicates the primary language of the book. For instance, “eng” is standard for English.

**8) num\_pages:** The number of pages the book contains.

**9) ratings\_count:** The total number of ratings the book received.

**10) text\_reviews\_count:** The total number of written text reviews the book received.

**11) publication\_date:** The date the book was published.

**12) publisher:** The name of the book publisher.

**Project Objectives:**

The objective of the project is to predict the average rating assigned to each book. The

project can be submitted as a Jupyter Notebook and should include exploratory analysis of

the data, feature engineering and selection, model training and evaluation.

**List of libraries:**

**Pandas:** is an open-source library that provides high-performance, easy-to-use data structures and data analysis tools for the Python programming language. It is used extensively in data science and machine learning applications.

**NumPy:** is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting and discrete Fourier transforms…

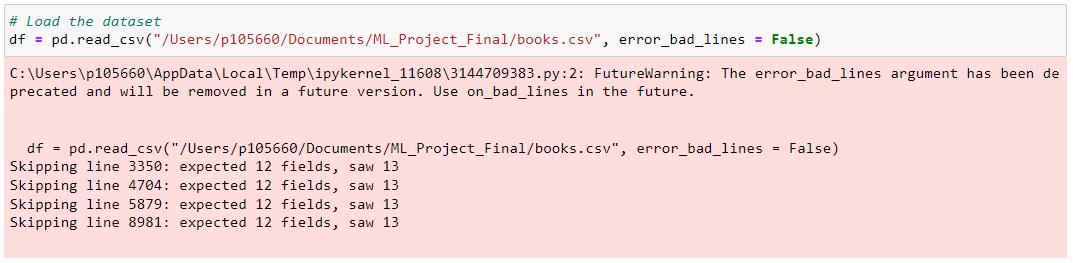
**Matplotlib:** is a comprehensive open-source library for creating static, animated, and interactive visualizations in Python.

**Seaborn:** is a Python data visualization library based on Matplotlib. It provides a high-level interface for creating informative and visually appealing statistical graphics.

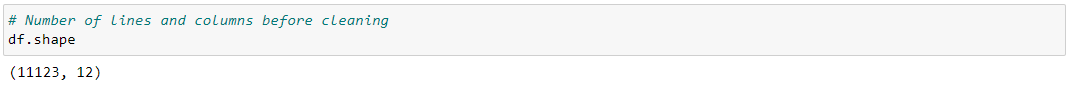
**Scikit-learn**: is a Python library for machine learning that provides a wide range of tools for predictive data analysis.

**I- Data Exploration**

**1. Read Data**



We observe that the dataset contains 4 bad lines. Therefore, it is necessary to proceed with the correction of the dataset.

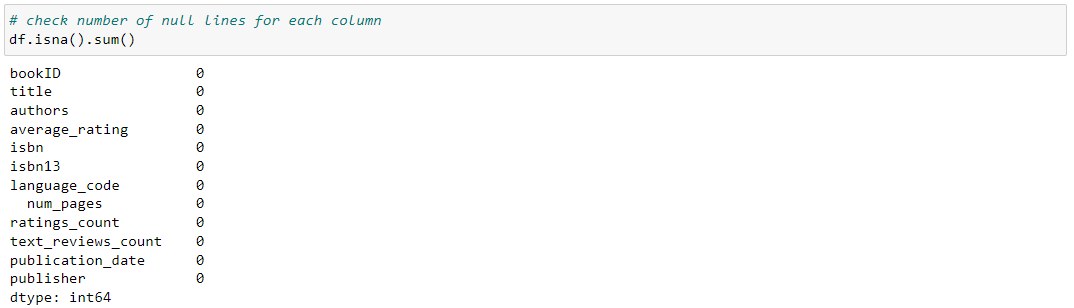


We have 11123 lines before correction.

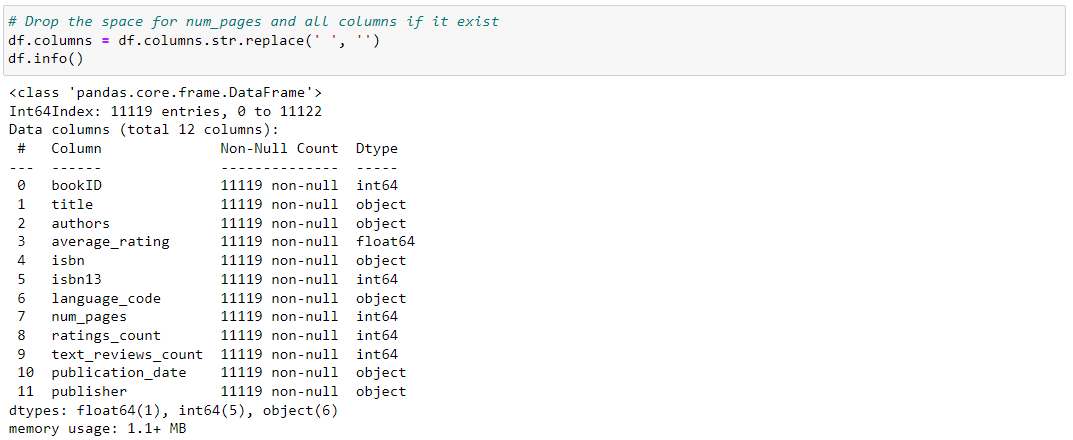
**2- Check invalid values**



I drop the bad lines to have just the data with the good format.

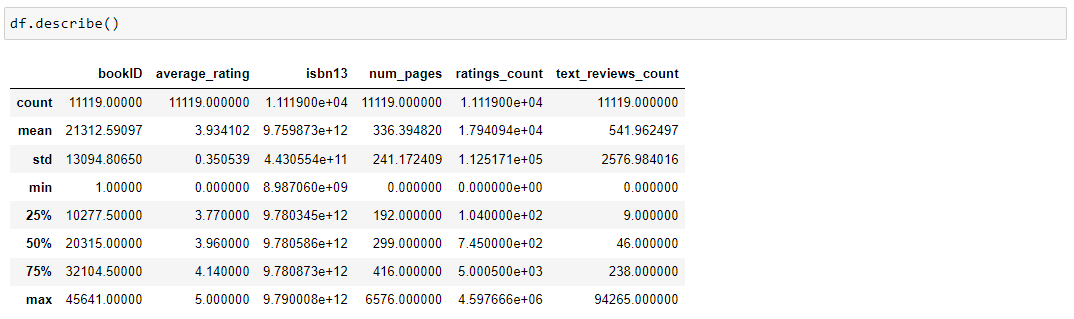


Check number of null lines for each column.

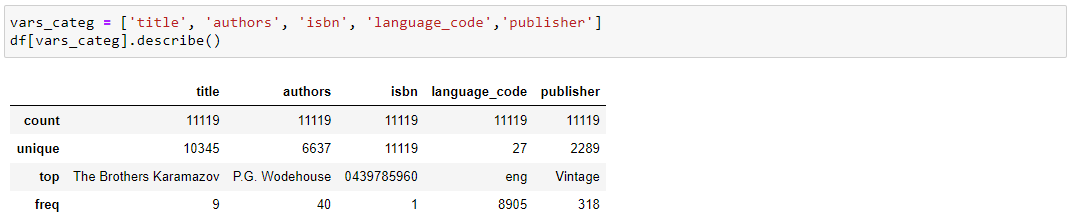


Drop the space if it exists for all columns before starting the analysis part.

**3- Describe the data**



Generate descriptive statistics for all numeric columns in the DataFrame.



Generate descriptive statistics for all categorical columns in the DataFrame.

**Remak :**

* We have 11119 books.
* Most of books are in English (8905/11119).
* P.G. Wodehouse is the author with the greatest number of books (40 books).
* The Brothers Karamazov is the most recurrent book (9 times).
* Most publisher is Vintage (318 books).

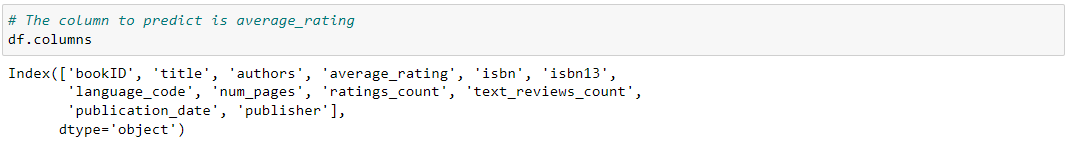
**II. Data Analysis and ML**

**1- Type of problem**

We are going to try to predict the average rating of each book, we are in the context of a supervised problem.

Predicting a value: it’s a regression. We have a regression problem.

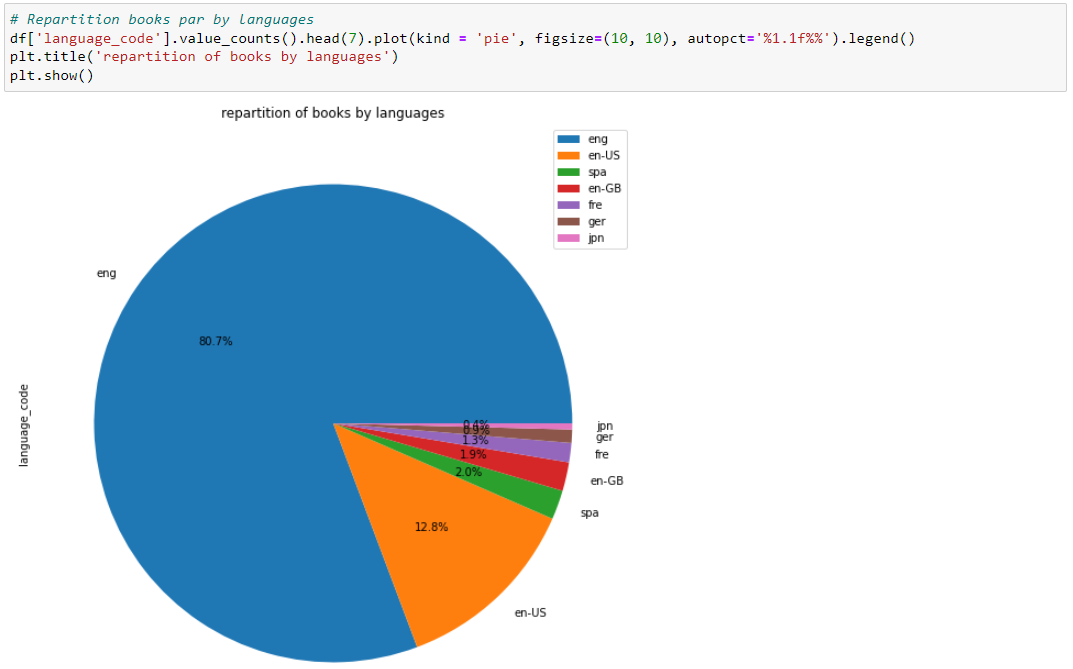
**2- Data analysis**



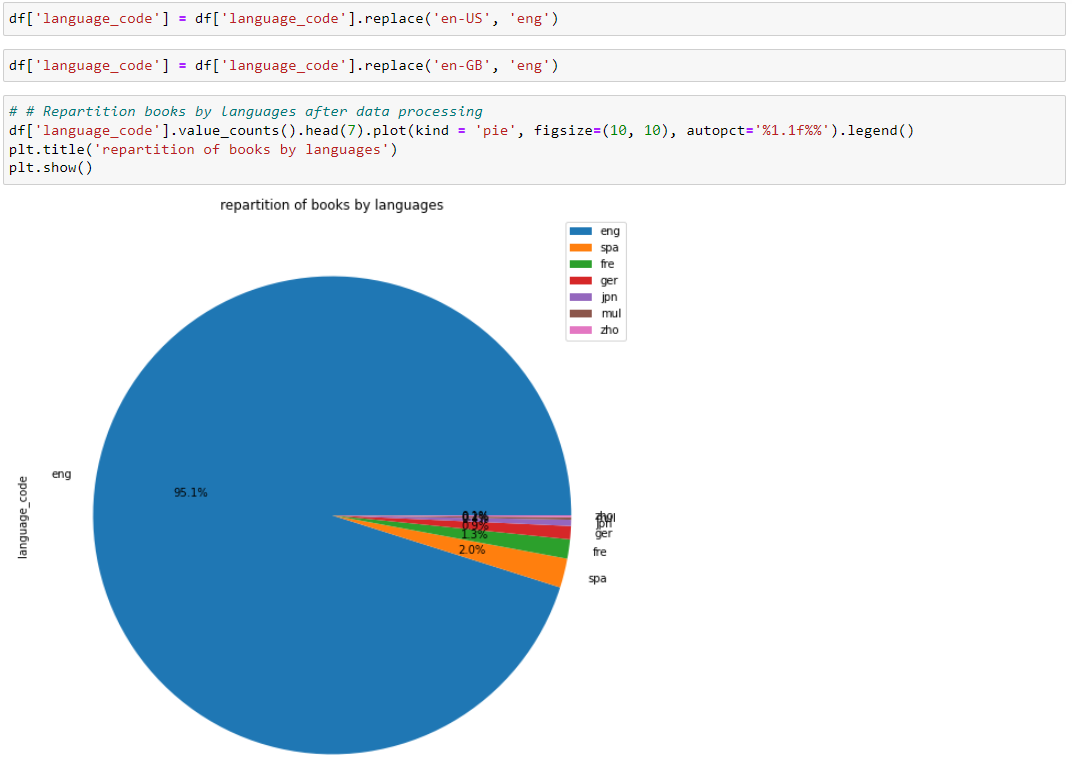
The column to predict is average\_rating, we need to predict the average and compare with the real average rating.



We list the top 10 rated books. Twighlit is the top one with 4597666 rates.



Repartition books par by languages. 80% are in Eng and 12.8% are in Eng-US and the third category is Spanish with 2%.



Here we replace en-US and en-GB by eng only. we have now 95% of books are in eng.



Create a new dataset with only the top 4 languages. So here we have English, Spanish, French and German.

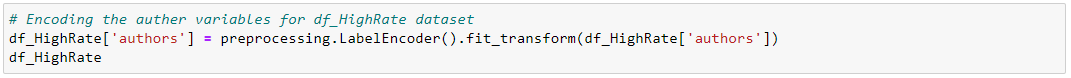
**3- Split the data**

The objective is to split the data of the new dataset df\_top\_4\_languages into two categories based on the ratings\_count

|  |  |
| --- | --- |
| df\_HighRate | df\_LowRate |
| It contains books that have received the total number of ratings the book received >= 100000. | It contains books that have received the total number of ratings the book received < 100000. |
| Length: (352, 12) | Length: (10635, 12) |

**4- Feature engineering**

In this step, we will encode the categorical variables for our two datasets (df\_HighRat, df\_LowRat). The objective of this step is to assign numerical values to all textual variables.



We encore authors, title, and publisher variables exactly with the same logic. Each author for example should have a specific id.



We Encode the language\_code variables using get\_dummies() function, and after we create a new column for each language, then we identify using binary code witch language is right for each book.

5- Selected model

For this case I'm ready to build the Linear Regression Model because Linear regression models are easier to adjust than models that are nonlinearly related to their parameters, and the statistical properties of the resulting estimators are easier to determine.



We split the data on two set, 80% as a training data and 20% as a test data.



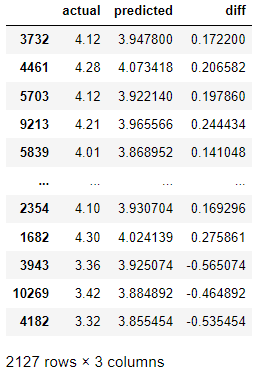
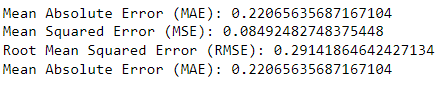
After splitting data into two set (80% training set and 20% test set), we create a linear regression model using the function LinearRegression(). The next step is to train the model using the training data.

We compute and display the regression coefficients. Finally, we display the actual and predicted values.

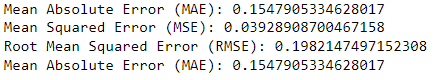
Performance of the model:

The performance of a linear regression model is typically evaluated using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

* Performance for df\_LowRate (model1):



* Performance for df\_HighRate (model2):



**Performances Analysis :**

Based on the provided error values, model2 appears to be the better one.

* Mean Absolute Error (MAE): model2 has a MAE of 0.1547905334628017, which is lower than the MAE of model1, which is 0.22065635687167104. This means that, on average, the predictions of model2 are closer to the actual values than those of model1.
* Mean Squared Error (MSE): model2 has a MSE of 0.03928908700467158, which is lower than the MSE of model1, which is 0.08492482748375448. The MSE gives more weight to larger errors because it squares the residuals. A lower MSE therefore indicates that model2 has fewer large errors than model1.
* Root Mean Squared Error (RMSE): model2 has a RMSE of 0.1982147497152308, which is lower than the RMSE of model1, which is 0.29141864642427134. This means that, on average, the square root of the squared differences between the predicted and actual values in model2 is lower than that of model1.