**PRUDATE.AI**

**Task -1**

**Exploratory data analysis:**

***Importing the packages:***

#importing the necessary packages

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import os

import warnings

import plotly.express as px

***Importing the dataset:***

#dataset

data = pd.read\_csv(r"C:\Users\Selva Vignesh M\Desktop\TechnicalAssessment\Technical Assessment\task1-books.csv\books.csv")

data.head(10)

***EDA:***

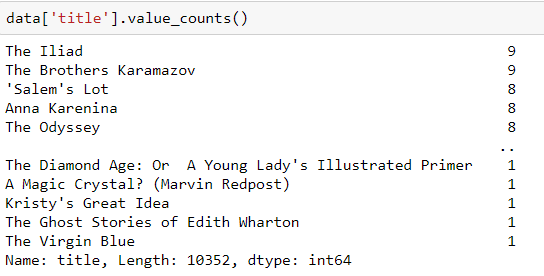
Exploratory Data analysis helps us to have better understanding of the data. Each and every variable contributes certain importance to the data. So eda for selected variables helps us to have a clear view of the data.



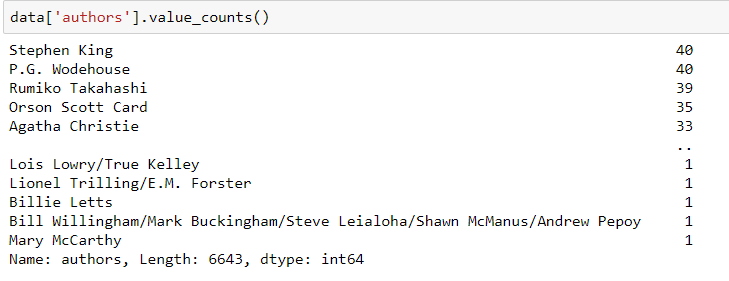
Here we check whether there are any null values in the dataset. Unfortunately we have about 11123 null values in the Unnamed column.



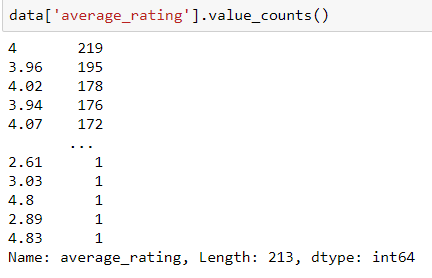
Now we have removed the unnamed column and the data is clean without the null values.



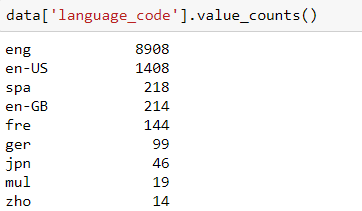
Here we explore the title variable and we can see that The Iliad book is the highest purchased book and we can see the ranks of other books.



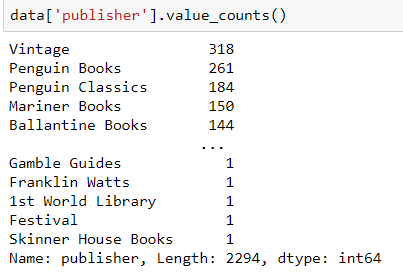
Here we explore the authors variable and we can see that Stephen King is the author whose books has been sold the most and we can see the ranks of other authors.



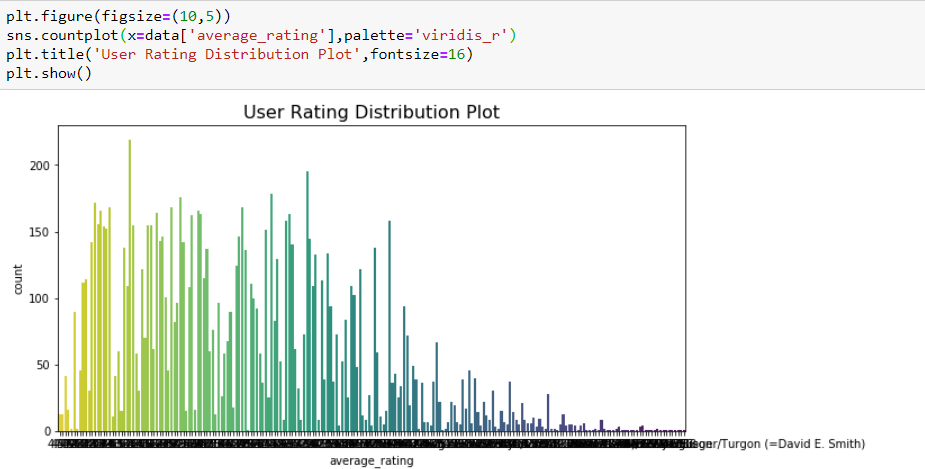
Here we explore the average rating variable and we can see that score 4 is the most given rating and we can see the ranks of other ratings.



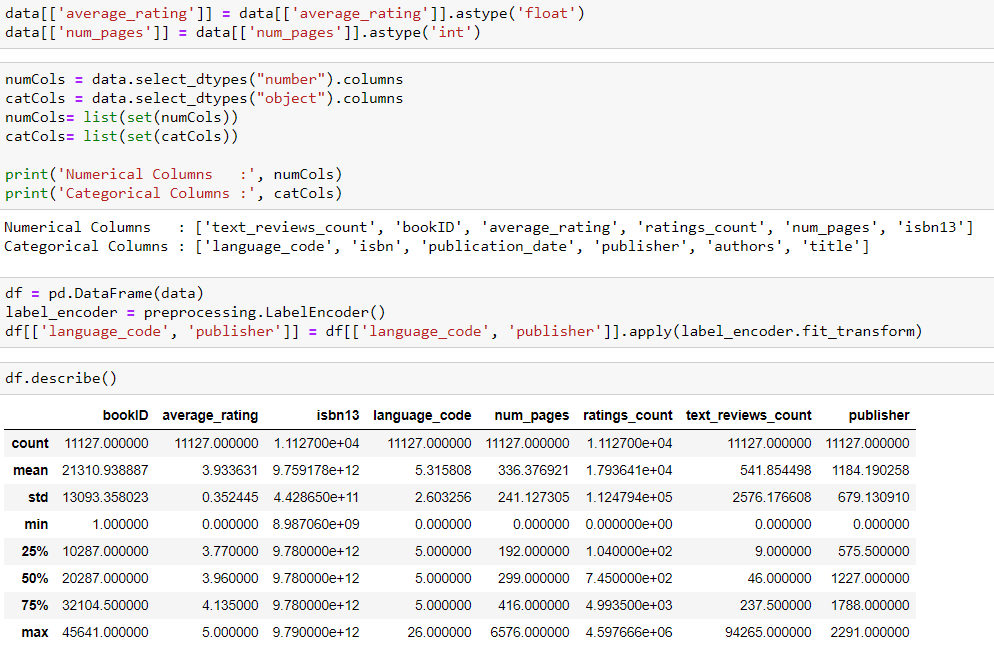
Here we explore the language code variable and we can see that score English is the most used language and we can see the ranks of other languages.



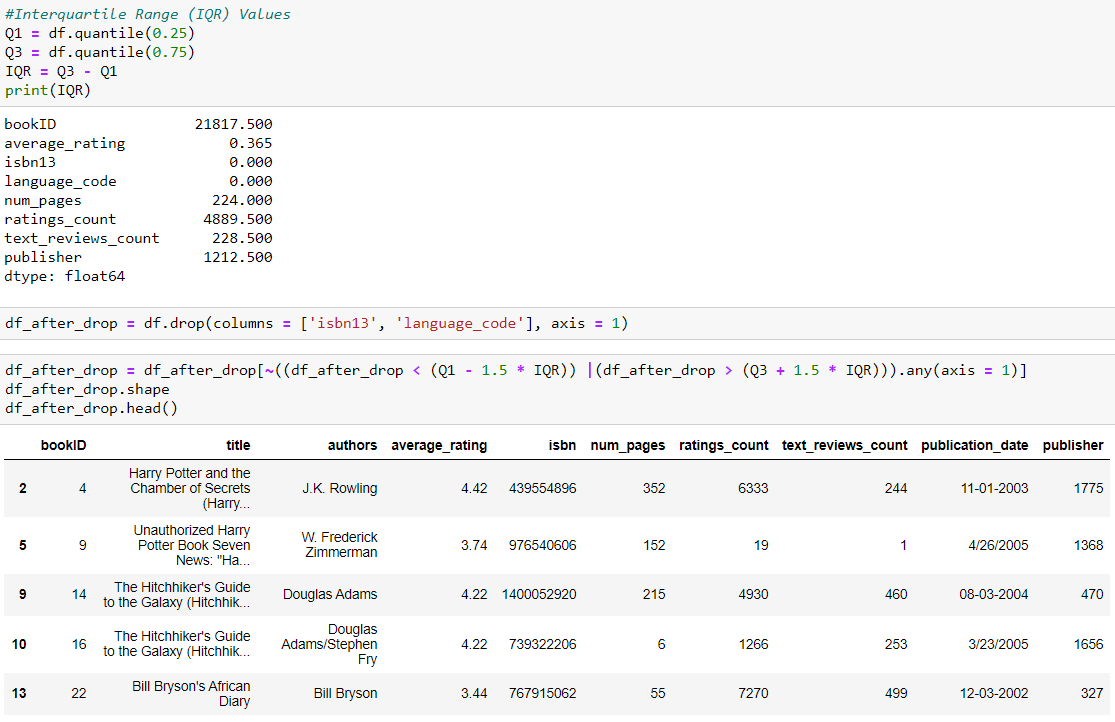
Here we explore the publisher variable and we can see that score Vintage is the most sold publisher and we can see the ranks of other publisher.



Here is a bar graph that represents the user’s rating distribution plot to show their distribution among the books.



Here we have converted the variables average rating and number of pages to float and int respectively. Then we are separating the numerical and categorical columns. Then we use Label encoder to convert the categorical variables to numerical variables.



Next we have defined quartiles to divide the data into 4 equal halves and their respective variable scores are shown.