README

https://github.com/HamidouTH/DSTI-Python-Project

In the ***projectpython.HamidouTHIAM.zip*** file, we have several folders:

A ***data in*** folder which the initial database contains, a ***data out*** folder which contains all the results obtained, ***image*** folder where we have some output graphs in python and and the ***code*** folder where we have all the codes.

This folder code contains 4 .py files which are:

- ***preprocessing.py*** : is a code that processes the data by providing a year column keeping only the years, deleting the lines of the average_rating column containing not a rating but text, renaming the column num_pages and encoding certain columns as authors…. Except the title column which was removed beforehand.
- ***importance_variables.py*** : In this file on two of the code snippets, the first of which calculates correlations between variables using a heatmap The second part concerns the study of the anova. NB: these data are in data out and are presented in the report.
- ***training_modelling.py*** : this file develops the two models, namely the linear regression model and the random forest model.
- ***Prediction_result.py*** : This file gives the output of model predictions as well as specified indices of model performance (results are presented in detail in the report.

To compile the code and see the results, all you have to do at the beginning of the file mentioned above is to change the import and export path of the data according to their location in your computer.

NB : In addition to these folders, we have the project_report.pdf file as well as the power Bi report where we can have fun with the filter on the results for better visualization.