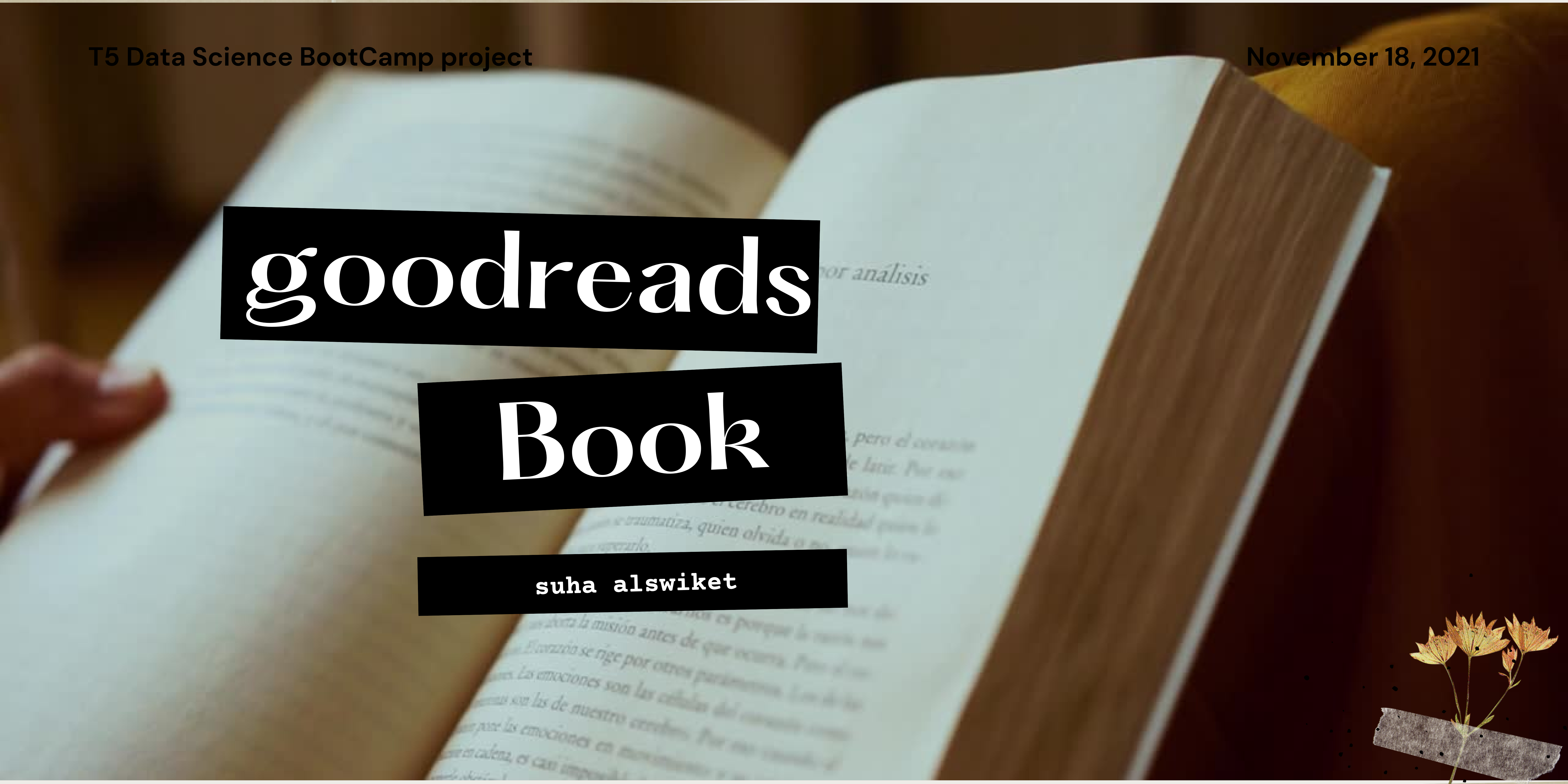
T5 Data Science BootCamp project

November 18, 2021

goodreads Book

suha alswiket



the goal

the goal of this project is to predict the reviews on the books based on the rating count "people who rated"

goodreads books

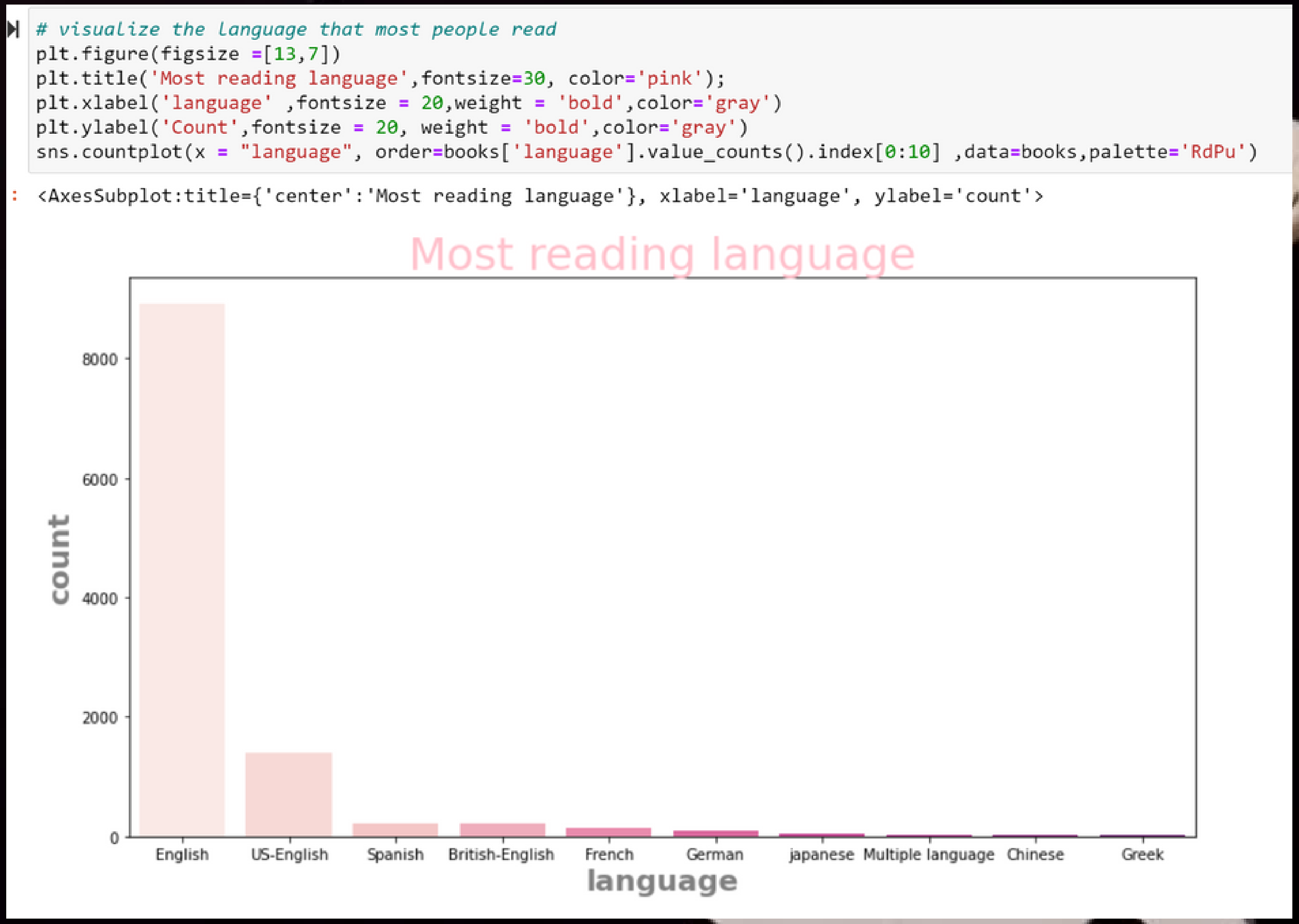
the dataset provided in .csv format from kaggle , it contains 11123 rows and 12 column (feature)

process data

- EDA
- changing name of the feature
- add extra feature
- changing names of the rows

visualization

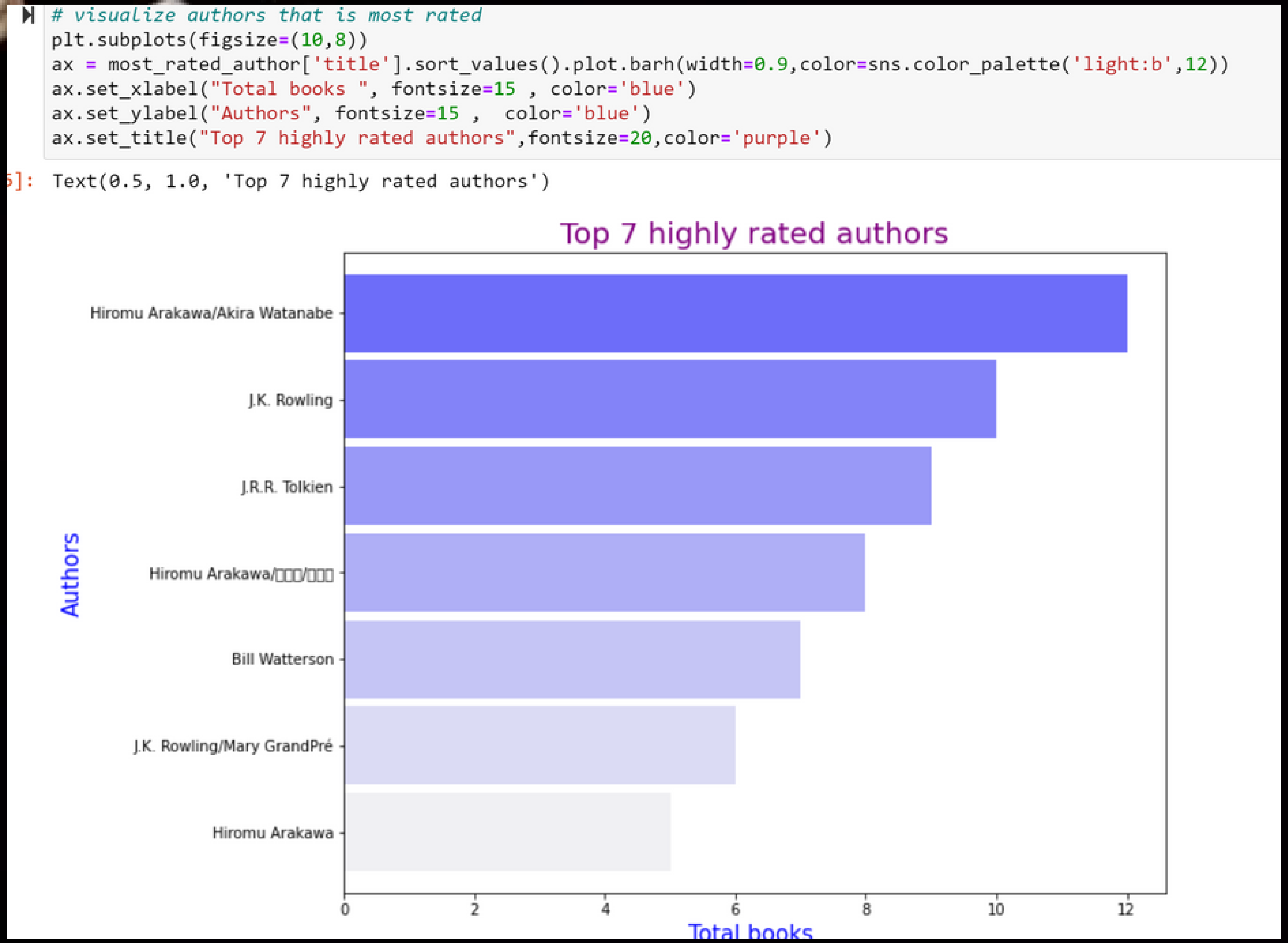
most language written by :



```
# visualize the language that most people read
plt.figure(figsize =[13,7])
plt.title('Most reading language',fontsize=30, color='pink');
plt.xlabel('language' ,fontsize = 20,weight = 'bold',color='gray')
plt.ylabel('Count',fontsize = 20, weight = 'bold',color='gray')
sns.countplot(x = "language", order=books['language'].value_counts().index[0:10] ,data=books,palette='RdPu')
```

```
<AxesSubplot:title={'center':'Most reading language'}, xlabel='language', ylabel='count'>
```

highly rated authors :

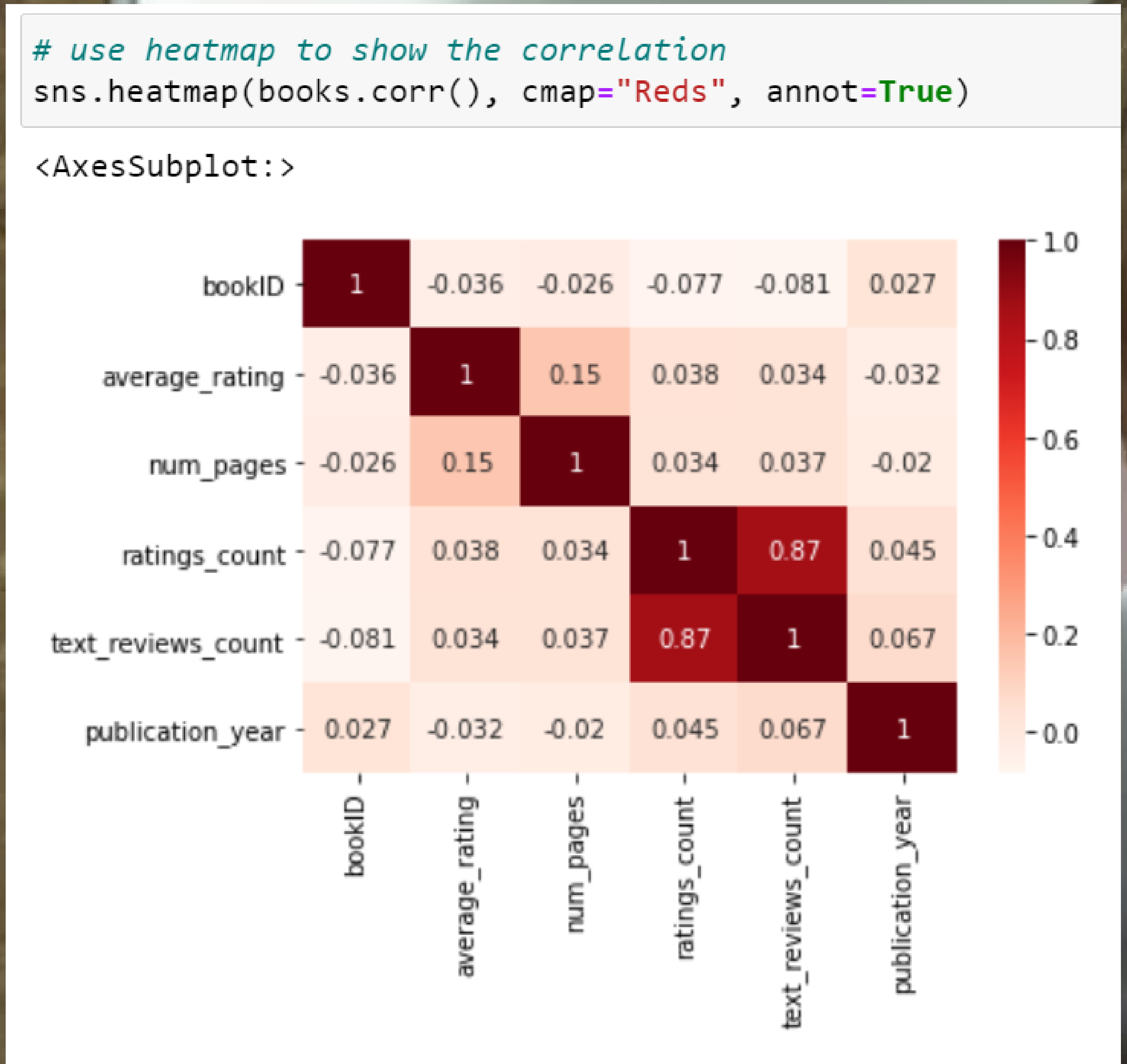


```
# visualize authors that is most rated
plt.subplots(figsize=(10,8))
ax = most_rated_author['title'].sort_values().plot.barh(width=0.9,color=sns.color_palette('light:b',12))
ax.set_xlabel("Total books ", fontsize=15 , color='blue')
ax.set_ylabel("Authors", fontsize=15 ,  color='blue')
ax.set_title("Top 7 highly rated authors",fontsize=20,color='purple')
```

```
Text(0.5, 1.0, 'Top 7 highly rated authors')
```

correlation

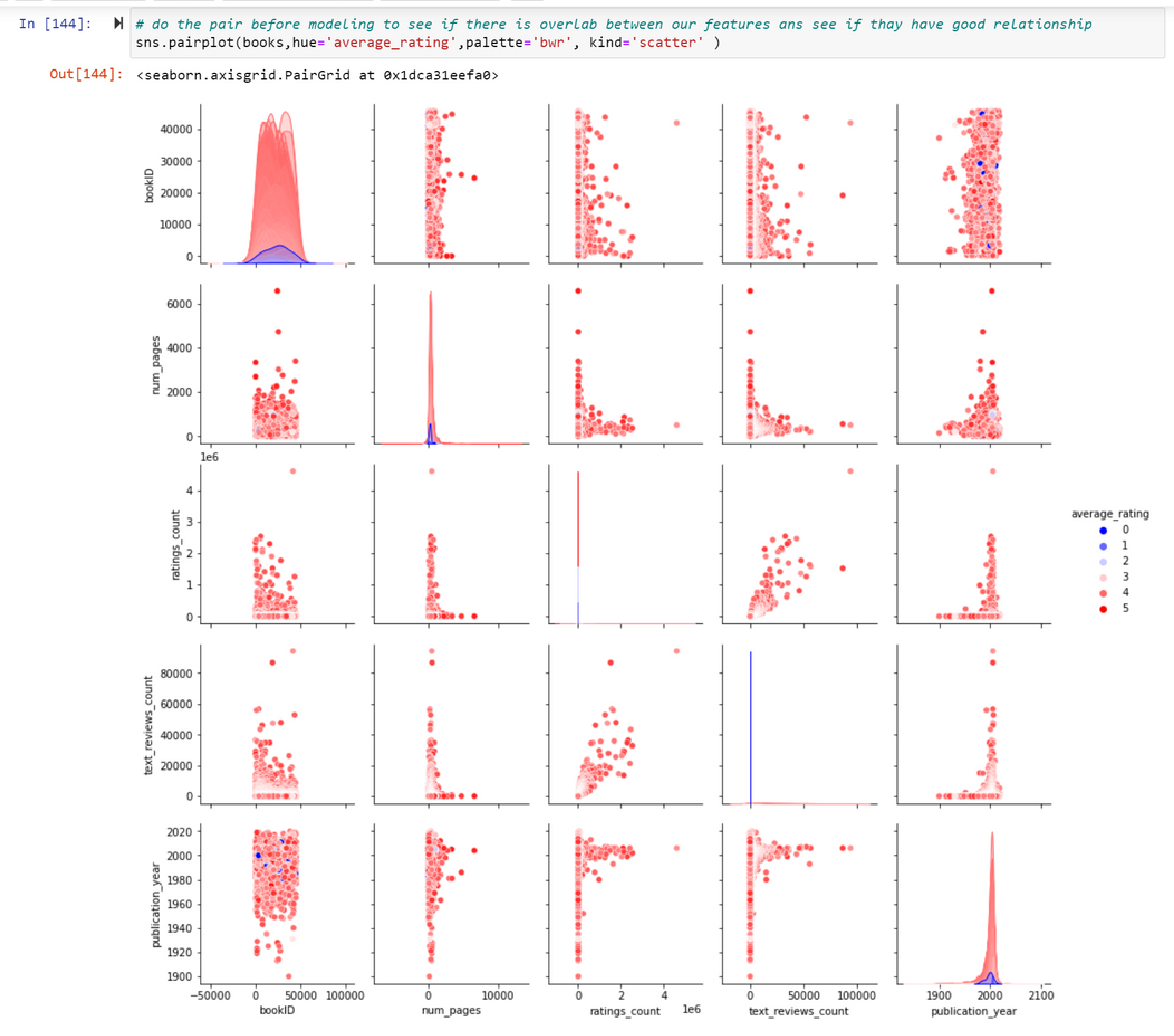
visualization



```
# use heatmap to show the correlation
sns.heatmap(books.corr(), cmap="Reds", annot=True)
```

```
<AxesSubplot:>
```

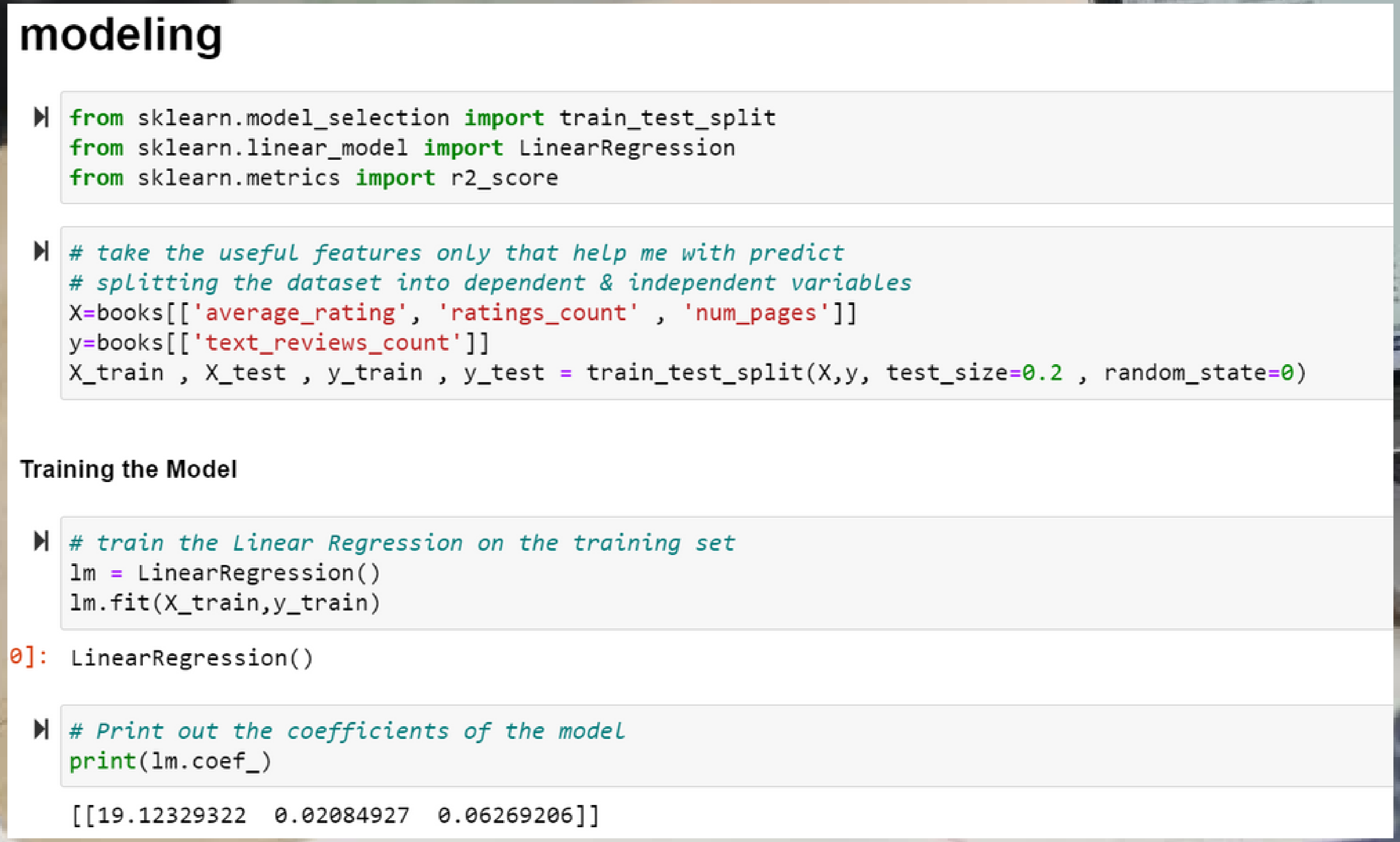
	bookID	average_rating	num_pages	ratings_count	text_reviews_count	publication_year
bookID	1	-0.036	-0.026	-0.077	-0.081	0.027
average_rating	-0.036	1	0.15	0.038	0.034	-0.032
num_pages	-0.026	0.15	1	0.034	0.037	-0.02
ratings_count	-0.077	0.038	0.034	1	0.87	0.045
text_reviews_count	-0.081	0.034	0.037	0.87	1	0.067
publication_year	0.027	-0.032	-0.02	0.045	0.067	1



```
In [144]:  # do the pair before modeling to see if there is overlab between our features ans see if thay have good relationship
           sns.pairplot(books,hue='average_rating',palette='bwr', kind='scatter' )
```

```
Out[144]: <seaborn.axisgrid.PairGrid at 0x1dca31eefa0>
```

Modeling



modeling

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
```

```
# take the useful features only that help me with predict
# splitting the dataset into dependent & independent variables
X=books[['average_rating', 'ratings_count' , 'num_pages']]
y=books[['text_reviews_count']]
X_train , X_test , y_train , y_test = train_test_split(X,y, test_size=0.2 , random_state=0)
```

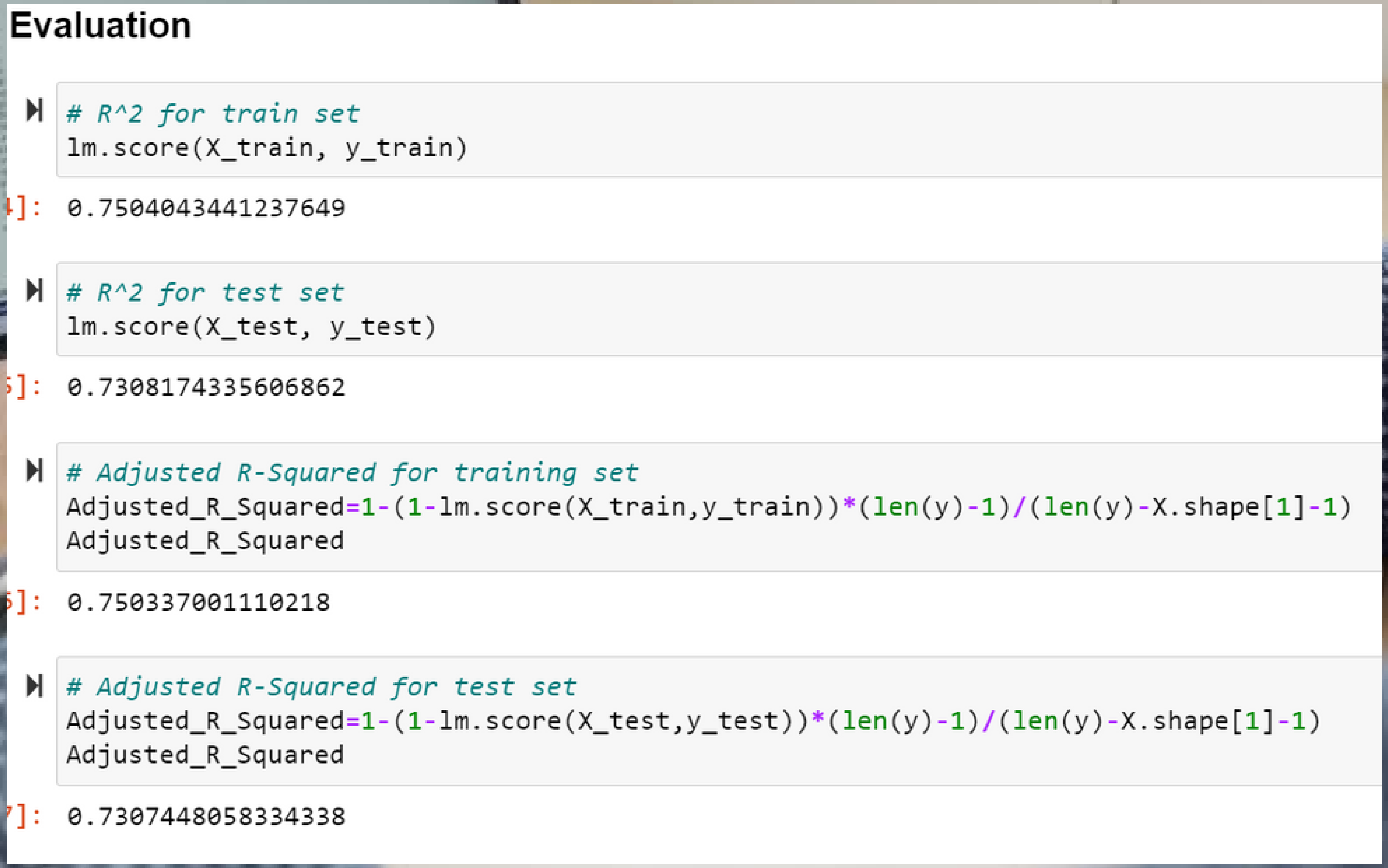
Training the Model

```
# train the Linear Regression on the training set
lm = LinearRegression()
lm.fit(X_train,y_train)
```

```
0]:  LinearRegression()
```

```
# Print out the coefficients of the model
print(lm.coef_)
```

```
[[19.12329322  0.02084927  0.06269206]]
```



Evaluation

```
# R^2 for train set
lm.score(X_train, y_train)
```

```
4]:  0.7504043441237649
```

```
# R^2 for test set
lm.score(X_test, y_test)
```

```
5]:  0.7308174335606862
```

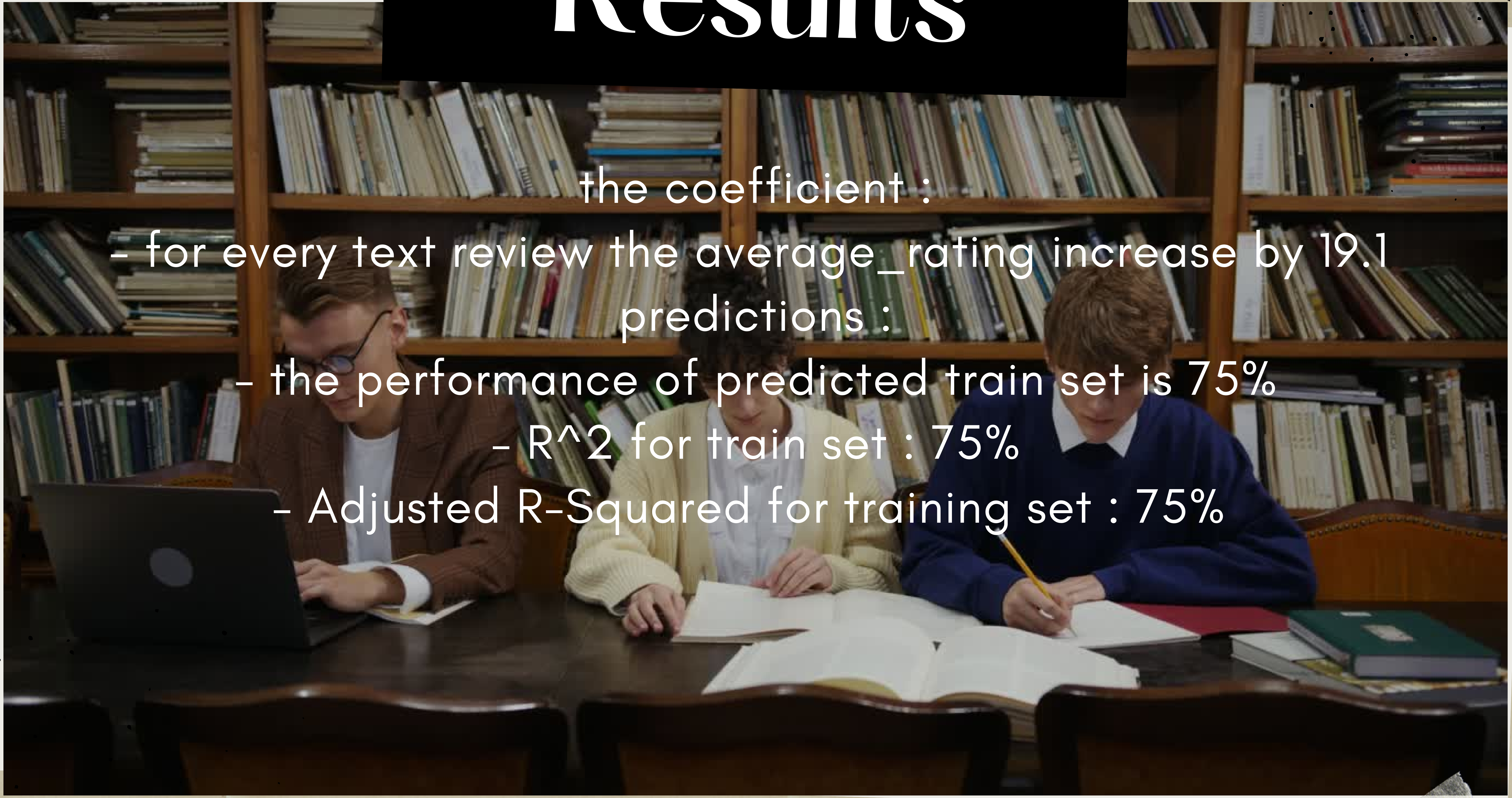
```
# Adjusted R-Squared for training set
Adjusted_R_Squared=1-(1-lm.score(X_train,y_train))*(len(y)-1)/(len(y)-X.shape[1]-1)
Adjusted_R_Squared
```

```
6]:  0.750337001110218
```

```
# Adjusted R-Squared for test set
Adjusted_R_Squared=1-(1-lm.score(X_test,y_test))*(len(y)-1)/(len(y)-X.shape[1]-1)
Adjusted_R_Squared
```

```
7]:  0.7307448058334338
```

Results



the coefficient :

- for every text review the average_rating increase by 19.1

predictions :

- the performance of predicted train set is 75%
- R^2 for train set : 75%
- Adjusted R-Squared for training set : 75%

T5 Data Science BootCamp project

November 18, 2021

Thank you