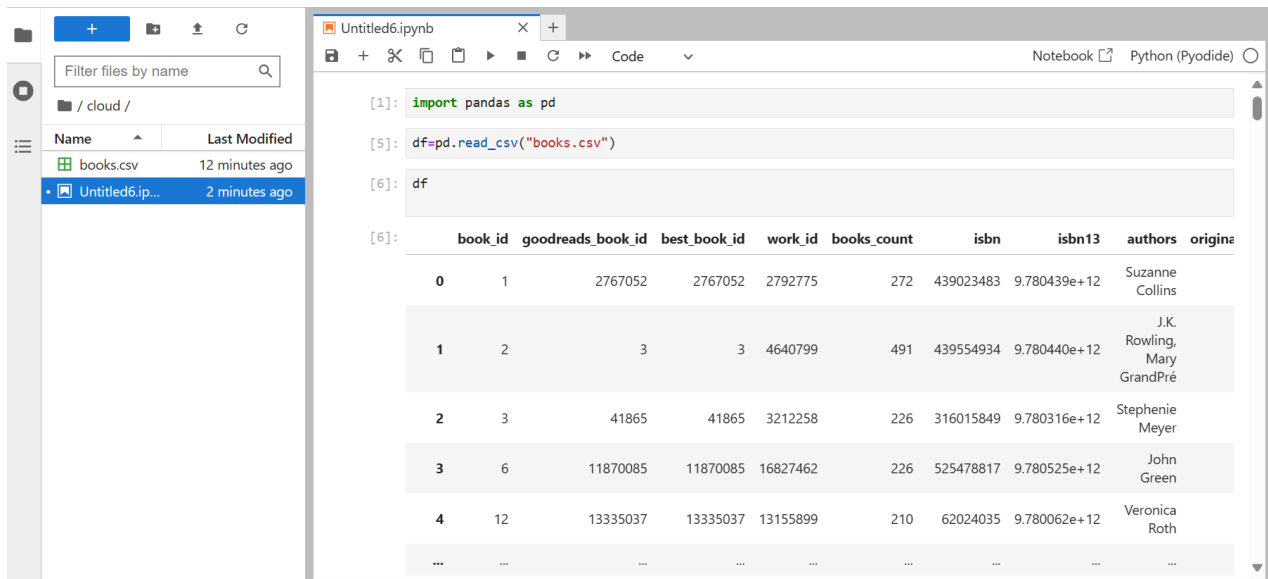


ASSIGNMENT 2 | DOCKERFILE AND DATA ANALYSIS WITH POPULAR BOOKS DATASET

****Jupyter notebook****

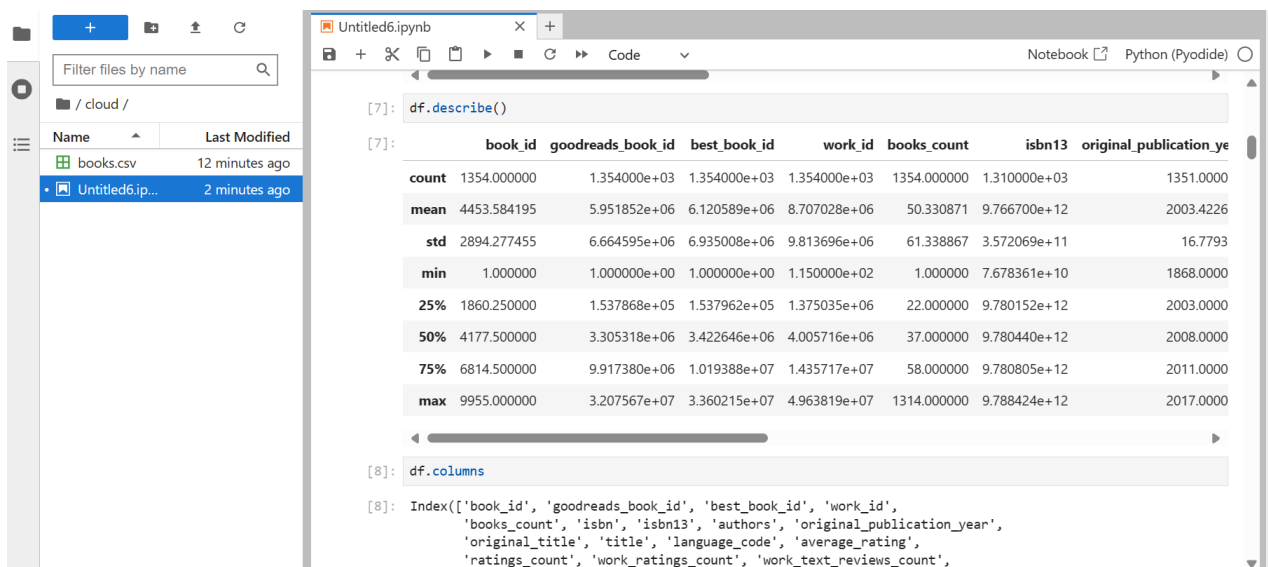


The screenshot shows a Jupyter notebook interface with a file browser on the left and a code editor on the right. The file browser shows a directory structure with 'books.csv' and 'Untitled6.ipynb'. The code editor shows the following code:

```
[1]: import pandas as pd
[5]: df=pd.read_csv("books.csv")
[6]: df
```

The output of the code is a table showing the first few rows of the 'books.csv' dataset:

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year
0	1	2767052	2767052	2792775	272	439023483	9.780439e+12	Suzanne Collins	2003.0000
1	2	3	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPré	2003.4226
2	3	41865	41865	3212258	226	316015849	9.780316e+12	Stephenie Meyer	2003.0000
3	6	11870085	11870085	16827462	226	525478817	9.780525e+12	John Green	2008.0000
4	12	13335037	13335037	13155899	210	62024035	9.780062e+12	Veronica Roth	2011.0000
...



The screenshot shows the same Jupyter notebook interface. The code editor shows the following code:

```
[7]: df.describe()
[8]: df.columns
```

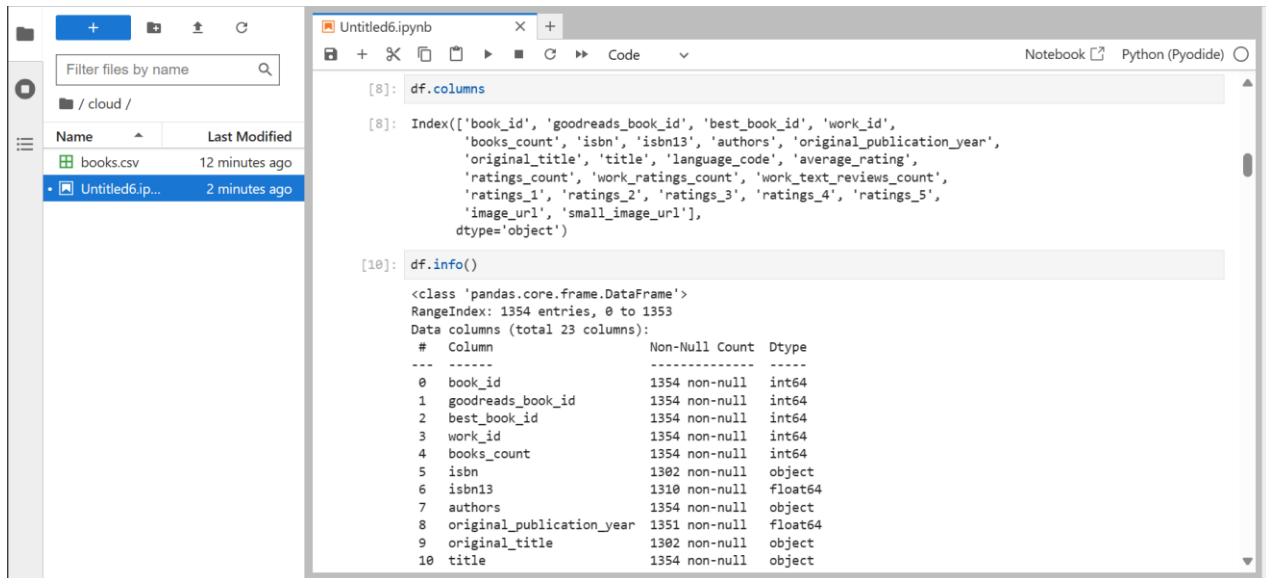
The output of the code is a summary statistics table and a list of column names.

Summary statistics table:

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn13	original_publication_year
count	1354.000000	1.354000e+03	1.354000e+03	1.354000e+03	1354.000000	1.310000e+03	1351.0000
mean	4453.584195	5.951852e+06	6.120589e+06	8.707028e+06	50.330871	9.766700e+12	2003.4226
std	2894.277455	6.664595e+06	6.935008e+06	9.813696e+06	61.338867	3.572069e+11	16.7793
min	1.000000	1.000000e+00	1.000000e+00	1.150000e+02	1.000000	7.678361e+10	1868.0000
25%	1860.250000	1.537868e+05	1.537962e+05	1.375035e+06	22.000000	9.780152e+12	2003.0000
50%	4177.500000	3.305318e+06	3.422646e+06	4.005716e+06	37.000000	9.780440e+12	2008.0000
75%	6814.500000	9.917380e+06	1.019388e+07	1.435717e+07	58.000000	9.780805e+12	2011.0000
max	9955.000000	3.207567e+07	3.360215e+07	4.963819e+07	1314.000000	9.788424e+12	2017.0000

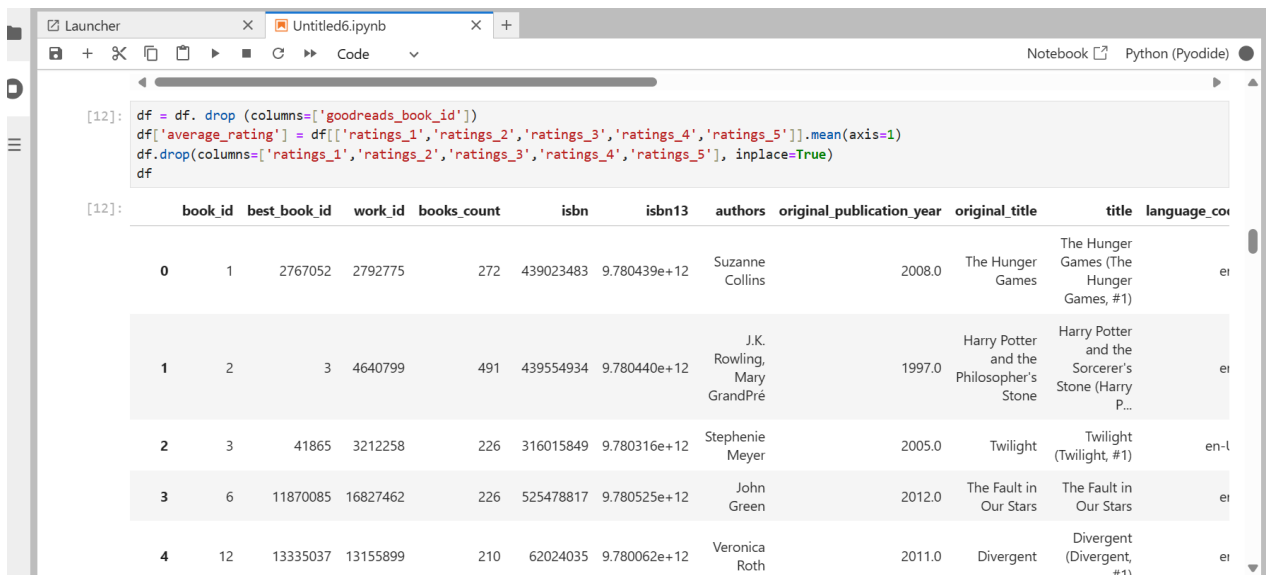
Column names:

```
Index(['book_id', 'goodreads_book_id', 'best_book_id', 'work_id', 'books_count', 'isbn', 'isbn13', 'authors', 'original_publication_year', 'original_title', 'title', 'language_code', 'average_rating', 'ratings_count', 'work_ratings_count', 'work_text_reviews_count', 'text_reviews_count'], dtype=object)
```



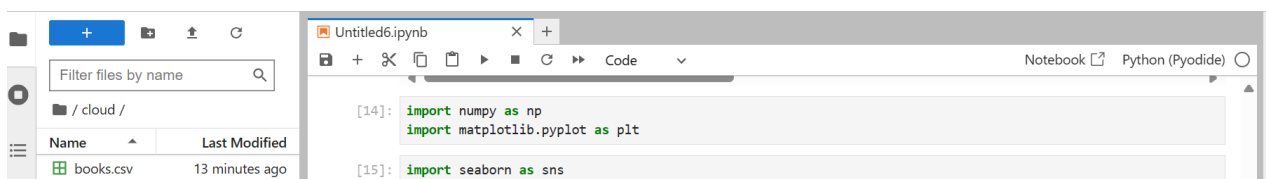
The screenshot shows a Jupyter Notebook interface. On the left, a file explorer sidebar displays a directory structure with a 'cloud' folder containing 'books.csv' (modified 12 minutes ago) and 'Untitled6.ipynb' (modified 2 minutes ago). The main notebook area, titled 'Untitled6.ipynb', shows two code cells. The first cell, [8], contains `df.columns`, which outputs a list of 23 column names: 'book_id', 'goodreads_book_id', 'best_book_id', 'work_id', 'books_count', 'isbn', 'isbn13', 'authors', 'original_publication_year', 'original_title', 'title', 'language_code', 'average_rating', 'ratings_count', 'work_ratings_count', 'work_text_reviews_count', 'ratings_1', 'ratings_2', 'ratings_3', 'ratings_4', 'ratings_5', 'image_url', and 'small_image_url'. The second cell, [10], contains `df.info()`, which outputs a summary of the DataFrame: 1354 entries, 0 to 1353, with 23 columns. The output table shows the data types and non-null counts for each column.

#	Column	Non-Null Count	Dtype
0	book_id	1354 non-null	int64
1	goodreads_book_id	1354 non-null	int64
2	best_book_id	1354 non-null	int64
3	work_id	1354 non-null	int64
4	books_count	1354 non-null	int64
5	isbn	1302 non-null	object
6	isbn13	1310 non-null	float64
7	authors	1354 non-null	object
8	original_publication_year	1351 non-null	float64
9	original_title	1302 non-null	object
10	title	1354 non-null	object

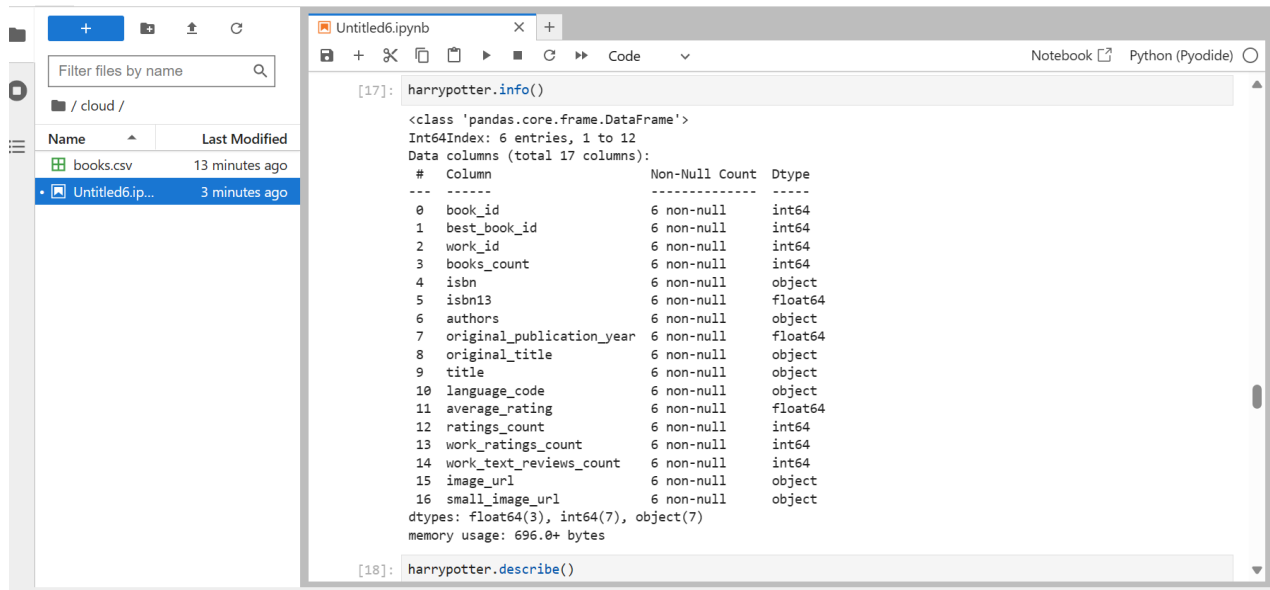


The screenshot shows the same Jupyter Notebook interface. The code cell [12] contains the following operations: `df = df.drop(columns=['goodreads_book_id'])`, `df['average_rating'] = df[['ratings_1', 'ratings_2', 'ratings_3', 'ratings_4', 'ratings_5']].mean(axis=1)`, and `df.drop(columns=['ratings_1', 'ratings_2', 'ratings_3', 'ratings_4', 'ratings_5'], inplace=True)`. Below the code, a preview of the DataFrame is shown, displaying columns: book_id, best_book_id, work_id, books_count, isbn, isbn13, authors, original_publication_year, original_title, title, and language_code. The preview shows the first five rows of data.

	book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	title	language_co
0	1	2767052	2792775	272	439023483	9.780439e+12	Suzanne Collins	2008.0	The Hunger Games	The Hunger Games (The Hunger Games, #1)	en
1	2	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPré	1997.0	Harry Potter and the Philosopher's Stone	Harry Potter and the Sorcerer's Stone (Harry P...	en
2	3	41865	3212258	226	316015849	9.780316e+12	Stephenie Meyer	2005.0	Twilight	Twilight (Twilight, #1)	en-l
3	6	11870085	16827462	226	525478817	9.780525e+12	John Green	2012.0	The Fault in Our Stars	The Fault in Our Stars	en
4	12	13335037	13155899	210	62024035	9.780062e+12	Veronica Roth	2011.0	Divergent	Divergent (Divergent, #1)	en



The screenshot shows the same Jupyter Notebook interface. The code cell [14] contains the following import statements: `import numpy as np` and `import matplotlib.pyplot as plt`. The code cell [15] contains the following import statement: `import seaborn as sns`.

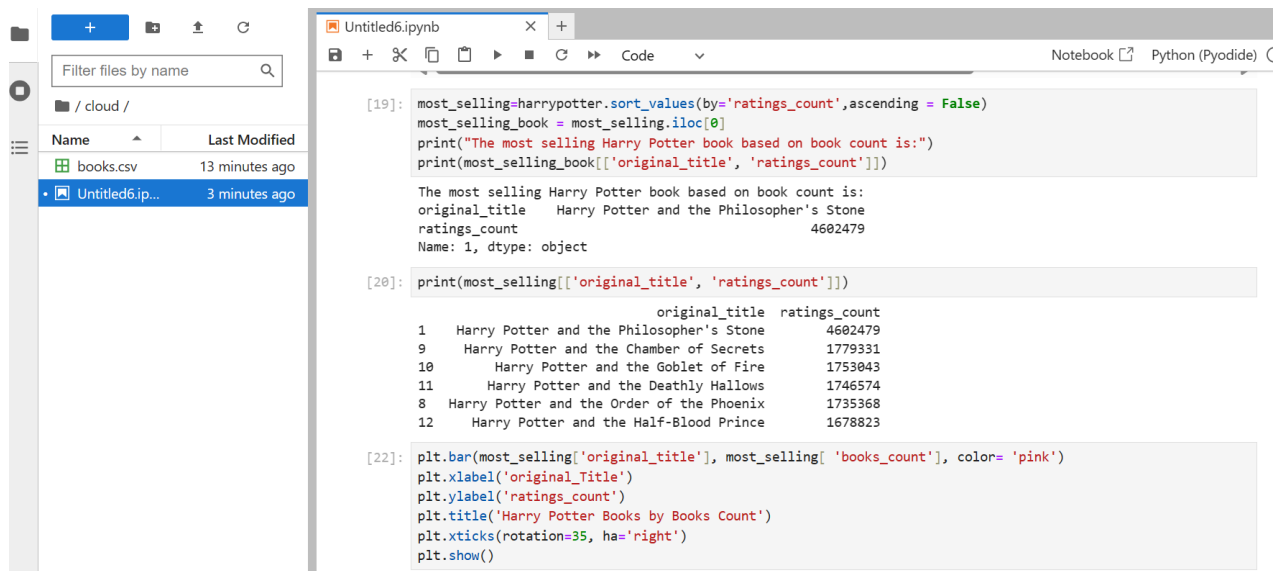


The screenshot shows a Jupyter Notebook with a file explorer on the left and a code editor on the right. The file explorer shows a folder named 'cloud' containing 'books.csv' and 'Untitled6.ip...'. The code editor shows the following code and output:

```
[17]: harrypotter.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 6 entries, 1 to 12
Data columns (total 17 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   book_id                               6 non-null      int64
1   best_book_id                         6 non-null      int64
2   work_id                              6 non-null      int64
3   books_count                          6 non-null      int64
4   isbn                                 6 non-null      object
5   isbn13                              6 non-null      float64
6   authors                             6 non-null      object
7   original_publication_year            6 non-null      float64
8   original_title                       6 non-null      object
9   title                               6 non-null      object
10  language_code                        6 non-null      object
11  average_rating                       6 non-null      float64
12  ratings_count                        6 non-null      int64
13  work_ratings_count                   6 non-null      int64
14  work_text_reviews_count              6 non-null      int64
15  image_url                            6 non-null      object
16  small_image_url                      6 non-null      object
dtypes: float64(3), int64(7), object(7)
memory usage: 696.0+ bytes
```

```
[18]: harrypotter.describe()
```



The screenshot shows a Jupyter Notebook with a file explorer on the left and a code editor on the right. The file explorer shows a folder named 'cloud' containing 'books.csv' and 'Untitled6.ip...'. The code editor shows the following code and output:

```
[19]: most_selling=harrypotter.sort_values(by='ratings_count',ascending = False)
most_selling_book = most_selling.iloc[0]
print("The most selling Harry Potter book based on book count is:")
print(most_selling_book[['original_title', 'ratings_count']])

The most selling Harry Potter book based on book count is:
original_title    Harry Potter and the Philosopher's Stone
ratings_count      4602479
Name: 1, dtype: object
```

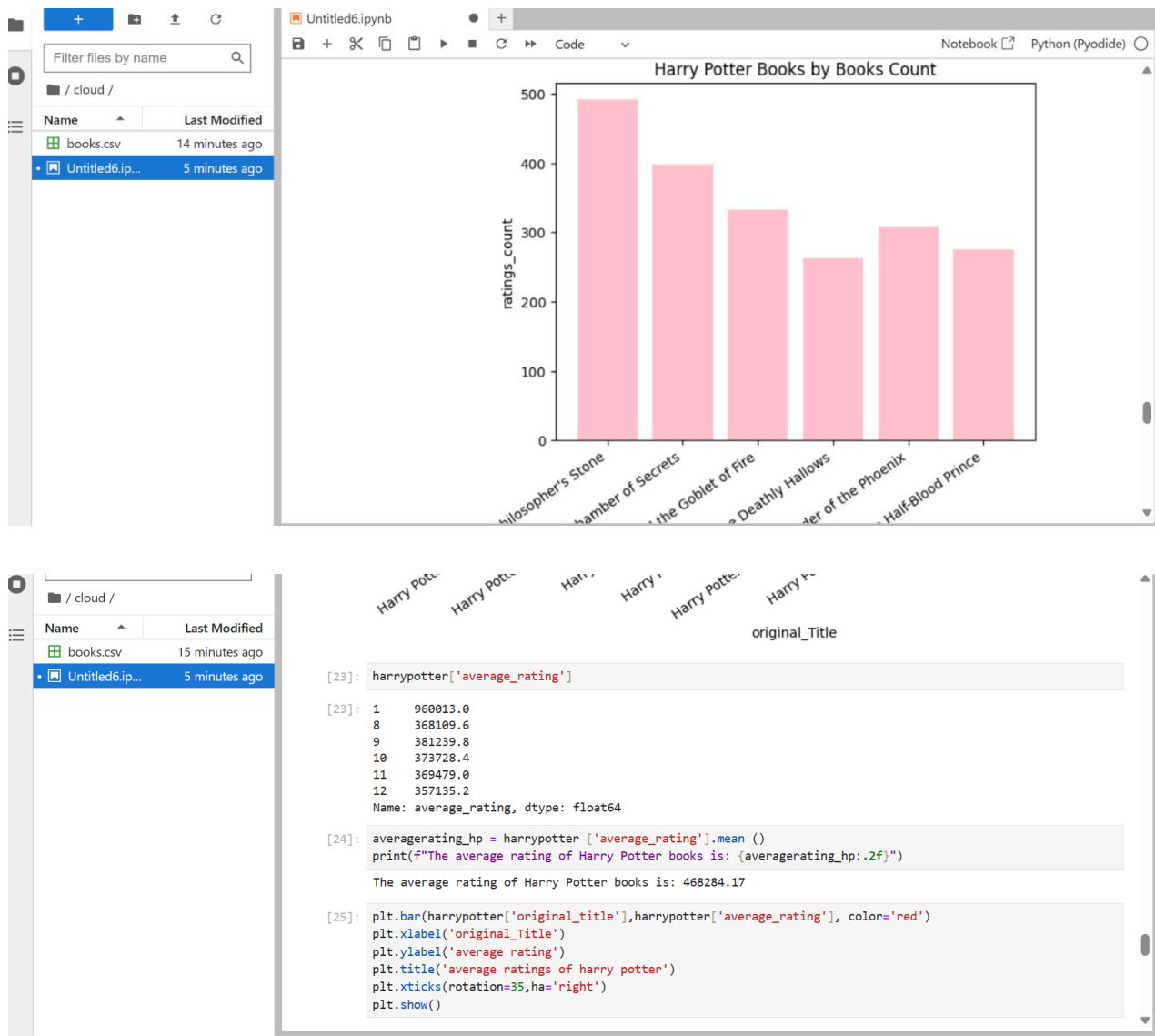
```
[20]: print(most_selling[['original_title', 'ratings_count']])

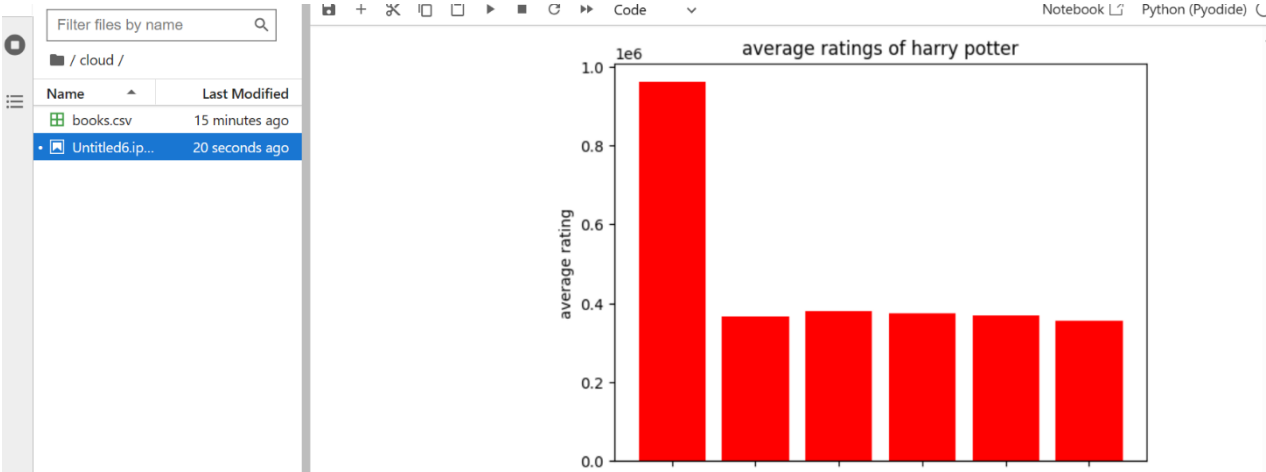
original_title    ratings_count
1   Harry Potter and the Philosopher's Stone    4602479
9   Harry Potter and the Chamber of Secrets    1779331
10  Harry Potter and the Goblet of Fire        1753043
11  Harry Potter and the Deathly Hallows       1746574
8   Harry Potter and the Order of the Phoenix  1735368
12  Harry Potter and the Half-Blood Prince     1678823
```

```
[22]: plt.bar(most_selling['original_title'], most_selling['books_count'], color= 'pink')
plt.xlabel('original_title')
plt.ylabel('ratings_count')
plt.title('Harry Potter Books by Books Count')
plt.xticks(rotation=35, ha='right')
plt.show()
```

WAGD HOSSAM
2206189

****RESULTS****





original_Title

Harry Potter and the Philosopher's Stone

Harry Potter and the Order of the Phoenix

Harry Potter and the Chamber of Secrets

Harry Potter and the Goblet of Fire

Harry Potter and the Deathly Hallows

Harry Potter and the Half-Blood Prince