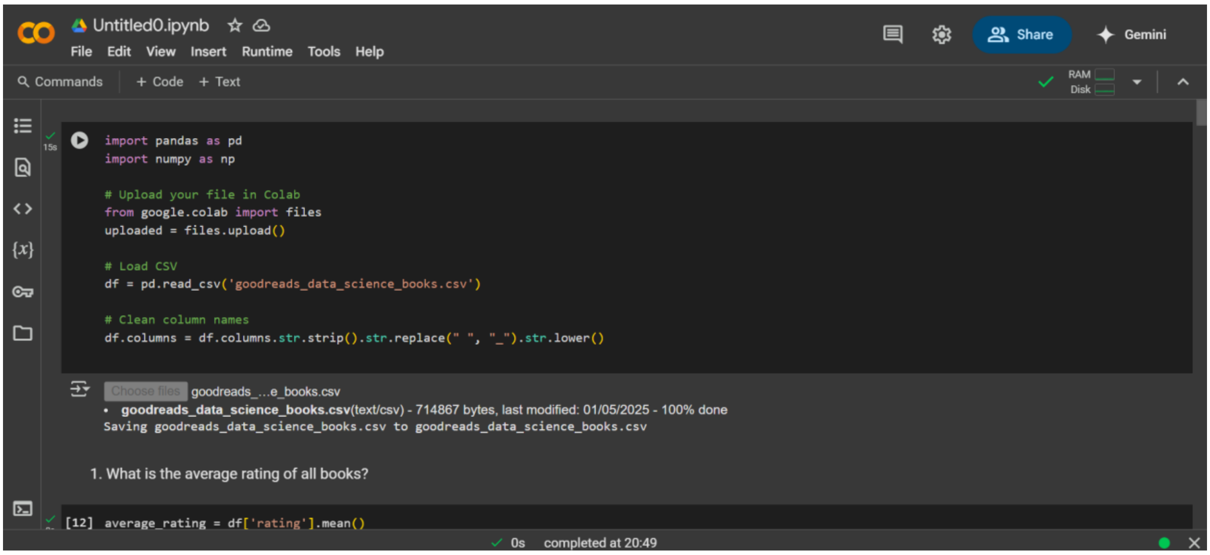
Name: Patel Saniya Mazherpasha

PRN: 20201040074

Batch: CS2

Roll No: 72

Topic: Goodreads Book Reviews



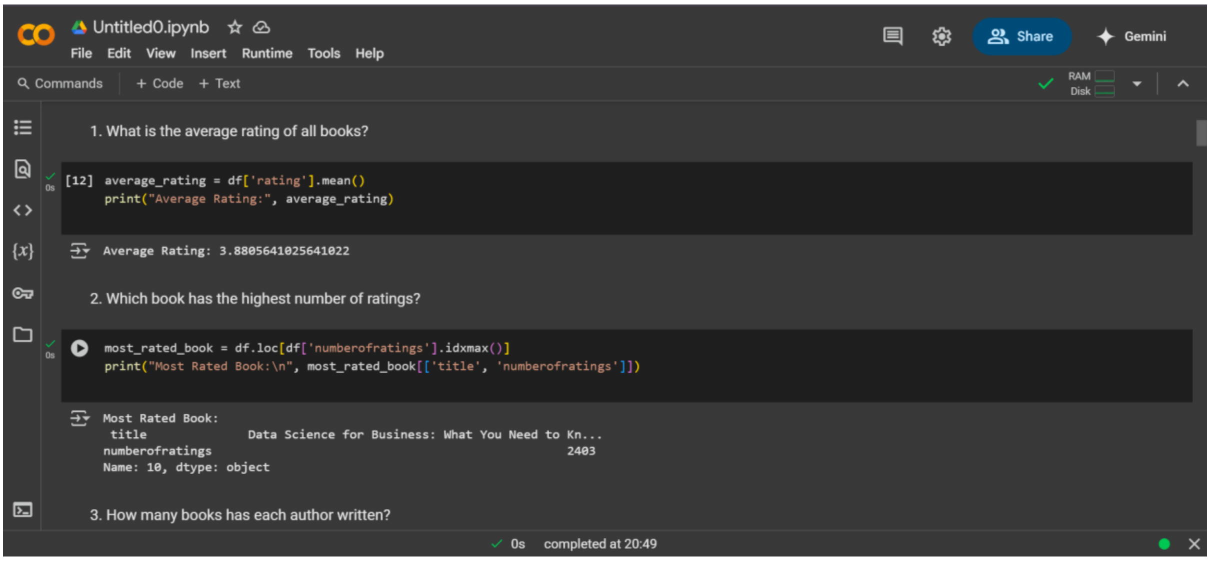
```
import pandas as pd
import numpy as np

# Upload your file in Colab
from google.colab import files
uploaded = files.upload()

# Load CSV
df = pd.read_csv('goodreads_data_science_books.csv')

# Clean column names
df.columns = df.columns.str.strip().str.replace(" ", "_").str.lower()
```

- **goodreads_data_science_books.csv**(text/csv) - 714867 bytes, last modified: 01/05/2025 - 100% done

```
Saving goodreads_data_science_books.csv to goodreads_data_science_books.csv
```

1. What is the average rating of all books?

```
[12] average_rating = df['rating'].mean()
```



1. What is the average rating of all books?

```
[12] average_rating = df['rating'].mean()
     print("Average Rating:", average_rating)
```

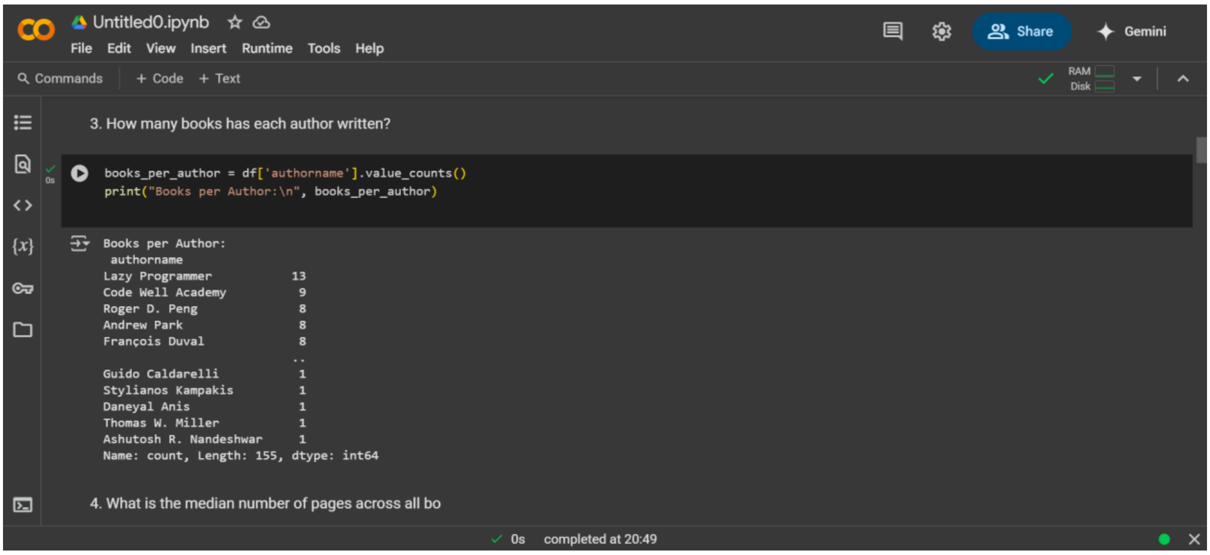
```
Average Rating: 3.8805641025641022
```

2. Which book has the highest number of ratings?

```
most_rated_book = df.loc[df['numberofratings'].idxmax()]
print("Most Rated Book:\n", most_rated_book[['title', 'numberofratings']])
```

```
Most Rated Book:
 title                  Data Science for Business: What You Need to Kn...
numberofratings                                                     2403
Name: 10, dtype: object
```

3. How many books has each author written?

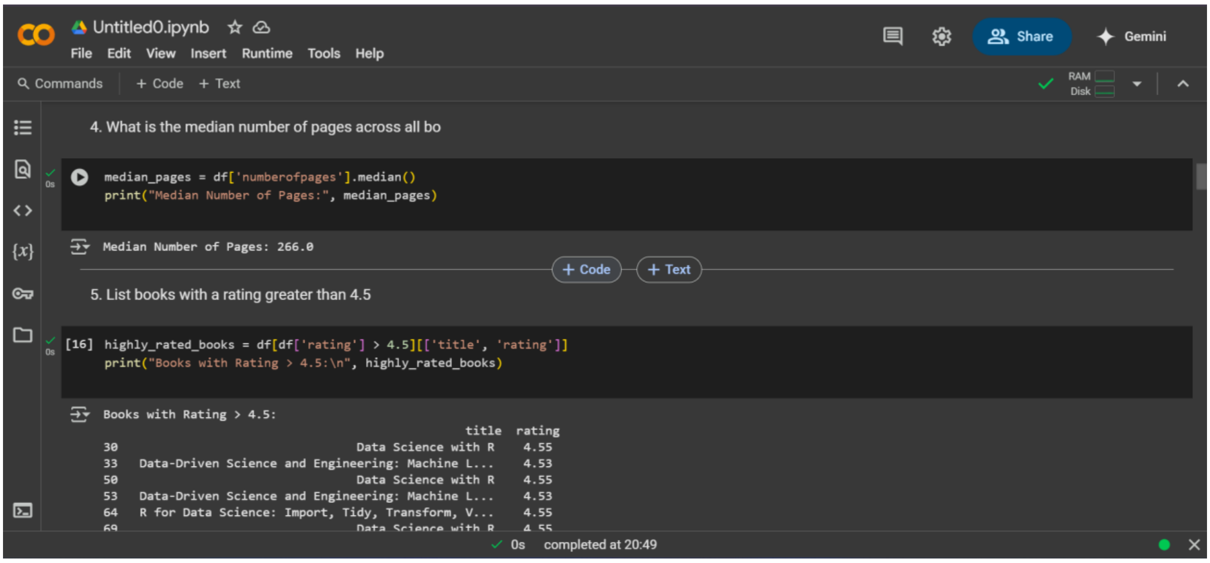


3. How many books has each author written?

```
books_per_author = df['authorname'].value_counts()
print("Books per Author:\n", books_per_author)
```

```
Books per Author:
 authorname
Lazy Programmer           13
Code Well Academy          9
Roger D. Peng              8
Andrew Park                8
François Duval             8
                          ..
Guido Caldarelli           1
Stylianos Kampakis         1
Daneyal Anis               1
Thomas W. Miller           1
Ashutosh R. Nandeshwar     1
Name: count, Length: 155, dtype: int64
```

4. What is the median number of pages across all bo



4. What is the median number of pages across all bo

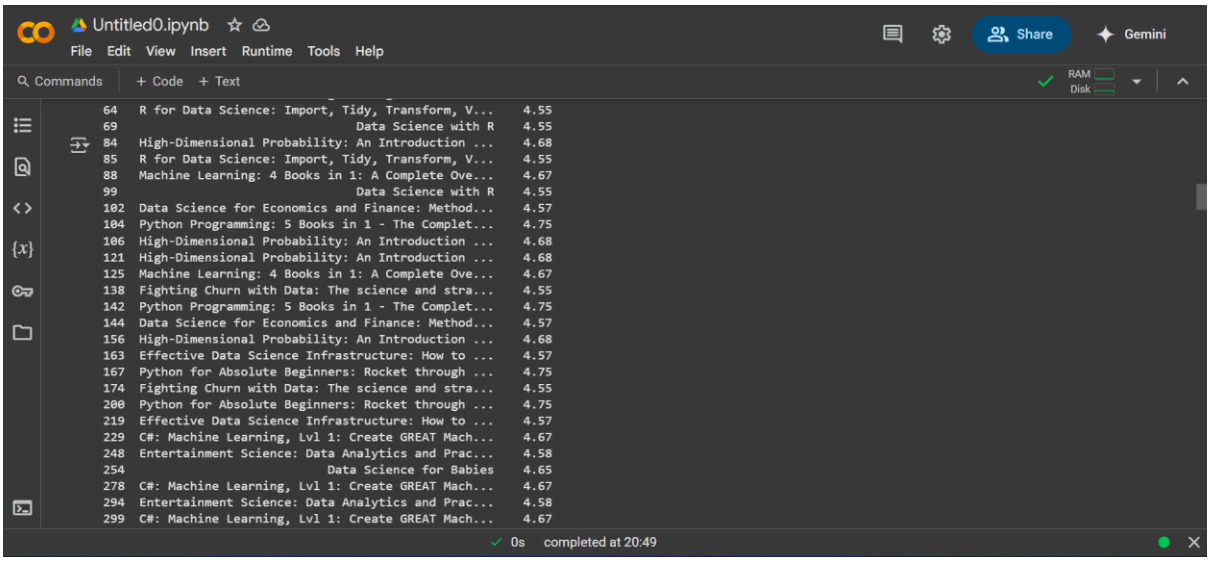
```
median_pages = df['numberofpages'].median()
print("Median Number of Pages:", median_pages)
```

```
Median Number of Pages: 266.0
```

5. List books with a rating greater than 4.5

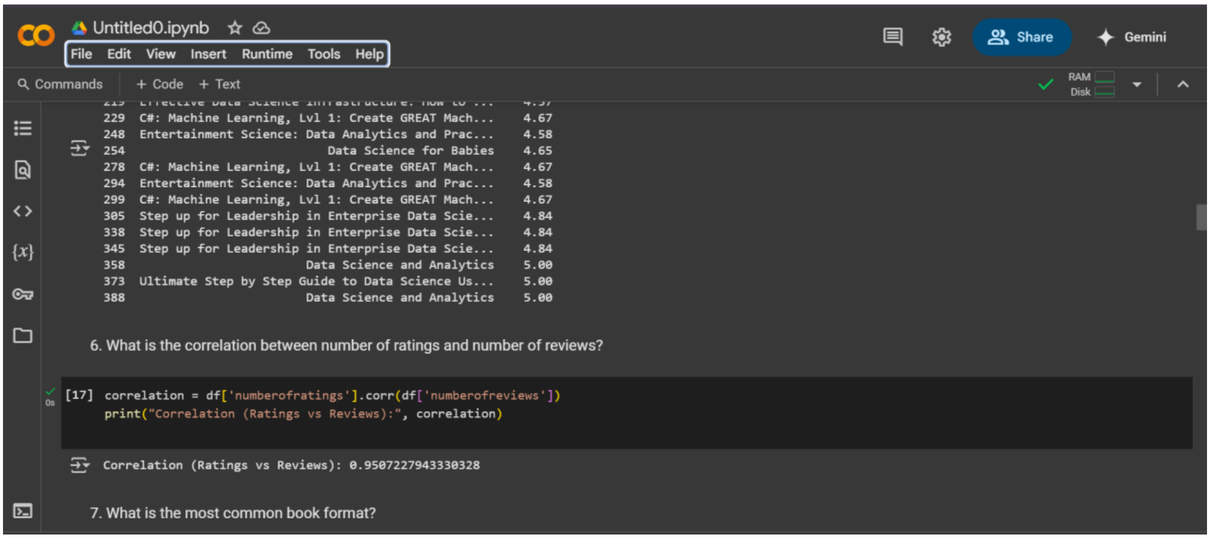
```
[16] highly_rated_books = df[df['rating'] > 4.5][['title', 'rating']]
     print("Books with Rating > 4.5:\n", highly_rated_books)
```

```
Books with Rating > 4.5:
                                                 title  rating
30                                  Data Science with R    4.55
33   Data-Driven Science and Engineering: Machine L...    4.53
50                                  Data Science with R    4.55
53   Data-Driven Science and Engineering: Machine L...    4.53
64   R for Data Science: Import, Tidy, Transform, V...    4.55
69                                  Data Science with R    4.55
```



```
64   R for Data Science: Import, Tidy, Transform, V...    4.55
69                                   Data Science with R    4.55
84   High-Dimensional Probability: An Introduction ...    4.68
85   R for Data Science: Import, Tidy, Transform, V...    4.55
88   Machine Learning: 4 Books in 1: A Complete Ove...    4.67
99                                   Data Science with R    4.55
102  Data Science for Economics and Finance: Method...    4.57
104  Python Programming: 5 Books in 1 - The Complet...    4.75
106  High-Dimensional Probability: An Introduction ...    4.68
121  High-Dimensional Probability: An Introduction ...    4.68
125  Machine Learning: 4 Books in 1: A Complete Ove...    4.67
138  Fighting Churn with Data: The science and stra...    4.55
142  Python Programming: 5 Books in 1 - The Complet...    4.75
144  Data Science for Economics and Finance: Method...    4.57
156  High-Dimensional Probability: An Introduction ...    4.68
163  Effective Data Science Infrastructure: How to ...    4.57
167  Python for Absolute Beginners: Rocket through ...    4.75
174  Fighting Churn with Data: The science and stra...    4.55
200  Python for Absolute Beginners: Rocket through ...    4.75
219  Effective Data Science Infrastructure: How to ...    4.57
229  C#: Machine Learning, Lvl 1: Create GREAT Mach...    4.67
248  Entertainment Science: Data Analytics and Prac...    4.58
254                              Data Science for Babies    4.65
278  C#: Machine Learning, Lvl 1: Create GREAT Mach...    4.67
294  Entertainment Science: Data Analytics and Prac...    4.58
299  C#: Machine Learning, Lvl 1: Create GREAT Mach...    4.67
```

```
64   R for Data Science: Import, Tidy, Transform, V...    4.55
69                                   Data Science with R    4.55
84   High-Dimensional Probability: An Introduction ...    4.68
85   R for Data Science: Import, Tidy, Transform, V...    4.55
88   Machine Learning: 4 Books in 1: A Complete Ove...    4.67
99                                   Data Science with R    4.55
102  Data Science for Economics and Finance: Method...    4.57
104  Python Programming: 5 Books in 1 - The Complet...    4.75
106  High-Dimensional Probability: An Introduction ...    4.68
121  High-Dimensional Probability: An Introduction ...    4.68
125  Machine Learning: 4 Books in 1: A Complete Ove...    4.67
138  Fighting Churn with Data: The science and stra...    4.55
142  Python Programming: 5 Books in 1 - The Complet...    4.75
144  Data Science for Economics and Finance: Method...    4.57
156  High-Dimensional Probability: An Introduction ...    4.68
163  Effective Data Science Infrastructure: How to ...    4.57
167  Python for Absolute Beginners: Rocket through ...    4.75
174  Fighting Churn with Data: The science and stra...    4.55
200  Python for Absolute Beginners: Rocket through ...    4.75
219  Effective Data Science Infrastructure: How to ...    4.57
229  C#: Machine Learning, Lvl 1: Create GREAT Mach...    4.67
248  Entertainment Science: Data Analytics and Prac...    4.58
254                              Data Science for Babies    4.65
278  C#: Machine Learning, Lvl 1: Create GREAT Mach...    4.67
294  Entertainment Science: Data Analytics and Prac...    4.58
299  C#: Machine Learning, Lvl 1: Create GREAT Mach...    4.67
```



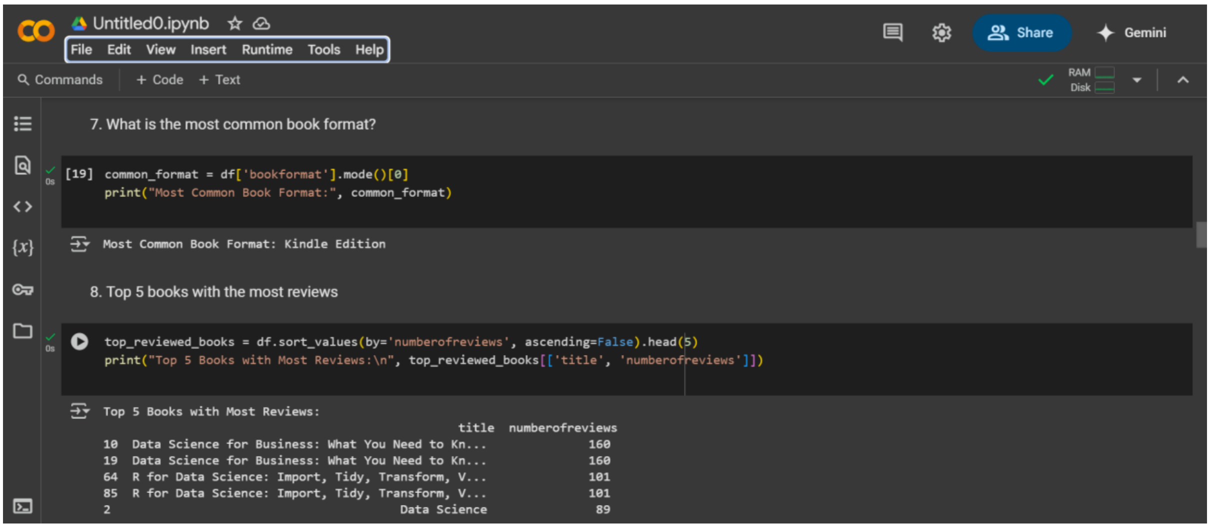
```
219  Effective Data Science Infrastructure: How to ...    4.57
229  C#: Machine Learning, Lvl 1: Create GREAT Mach...    4.67
248  Entertainment Science: Data Analytics and Prac...    4.58
254                              Data Science for Babies    4.65
278  C#: Machine Learning, Lvl 1: Create GREAT Mach...    4.67
294  Entertainment Science: Data Analytics and Prac...    4.58
299  C#: Machine Learning, Lvl 1: Create GREAT Mach...    4.67
305  Step up for Leadership in Enterprise Data Scie...    4.84
338  Step up for Leadership in Enterprise Data Scie...    4.84
345  Step up for Leadership in Enterprise Data Scie...    4.84
358                           Data Science and Analytics    5.00
373  Ultimate Step by Step Guide to Data Science Us...    5.00
388                           Data Science and Analytics    5.00
```

6. What is the correlation between number of ratings and number of reviews?

```
[17] correlation = df['numberofratings'].corr(df['numberofreviews'])
     print("Correlation (Ratings vs Reviews):", correlation)
```

```
Correlation (Ratings vs Reviews): 0.9507227943330328
```

7. What is the most common book format?



7. What is the most common book format?

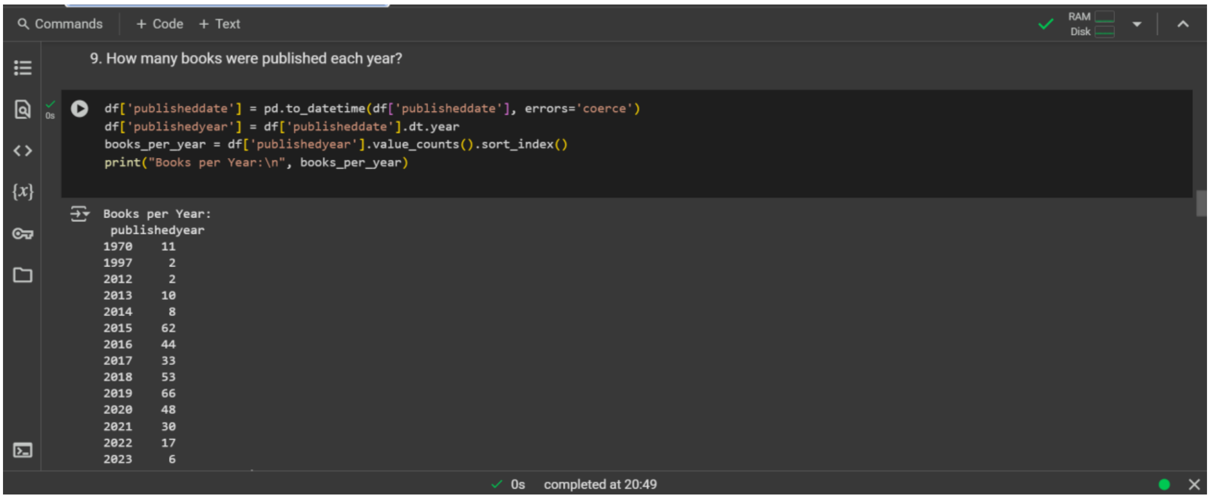
```
[19] common_format = df['bookformat'].mode()[0]
     print("Most Common Book Format:", common_format)
```

```
Most Common Book Format: Kindle Edition
```

8. Top 5 books with the most reviews

```
top_reviewed_books = df.sort_values(by='numberofreviews', ascending=False).head(5)
print("Top 5 Books with Most Reviews:\n", top_reviewed_books[['title', 'numberofreviews']])
```

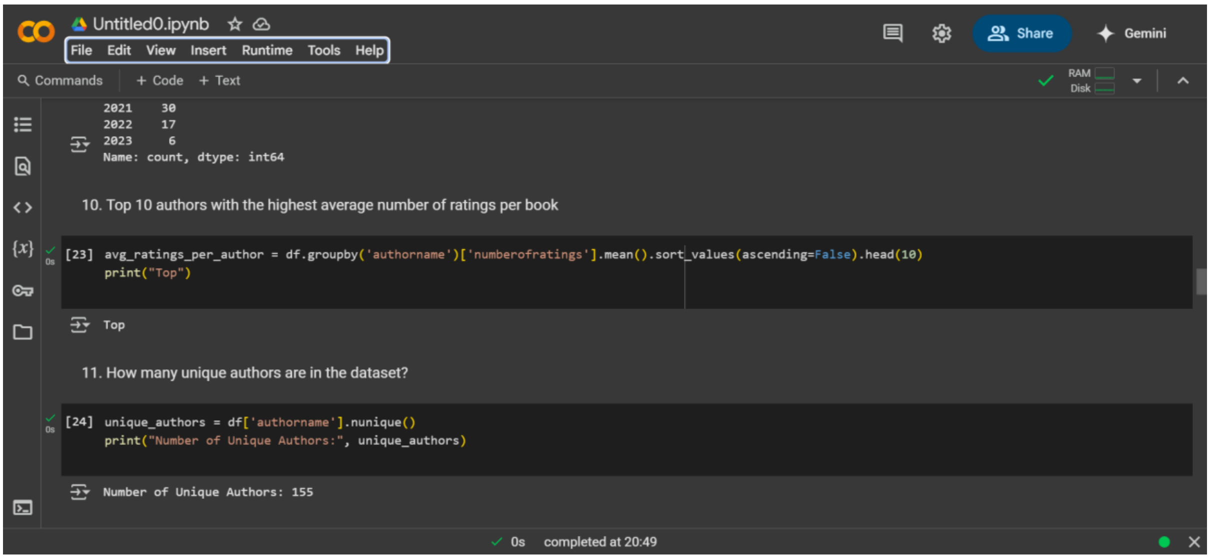
```
Top 5 Books with Most Reviews:
                                                 title  numberofreviews
10  Data Science for Business: What You Need to Kn...              160
19  Data Science for Business: What You Need to Kn...              160
64  R for Data Science: Import, Tidy, Transform, V...              101
85  R for Data Science: Import, Tidy, Transform, V...              101
2                                        Data Science               89
```



9. How many books were published each year?

```
df['publisheddate'] = pd.to_datetime(df['publisheddate'], errors='coerce')
df['publishedyear'] = df['publisheddate'].dt.year
books_per_year = df['publishedyear'].value_counts().sort_index()
print("Books per Year:\n", books_per_year)
```

```
Books per Year:
 publishedyear
1970    11
1997     2
2012     2
2013    10
2014     8
2015    62
2016    44
2017    33
2018    53
2019    66
2020    48
2021    30
2022    17
2023     6
```



```
2021    30
2022    17
2023     6
Name: count, dtype: int64
```

10. Top 10 authors with the highest average number of ratings per book

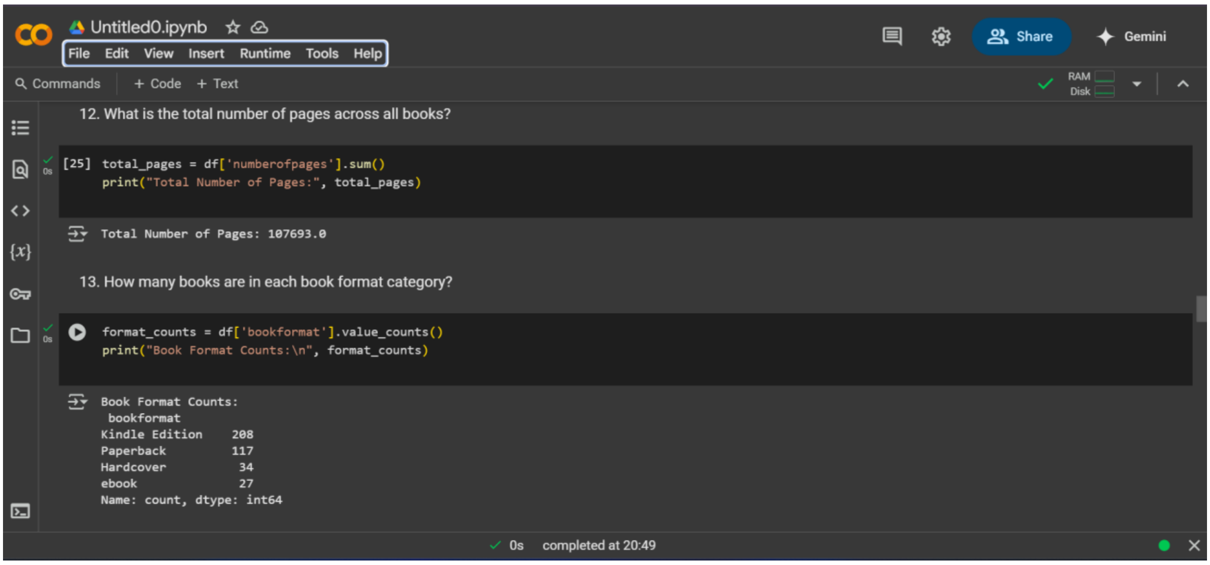
```
[23] avg_ratings_per_author = df.groupby('authorname')['numberofratings'].mean().sort_values(ascending=False).head(10)
     print("Top")
```

```
Top
```

11. How many unique authors are in the dataset?

```
[24] unique_authors = df['authorname'].nunique()
     print("Number of Unique Authors:", unique_authors)
```

```
Number of Unique Authors: 155
```



12. What is the total number of pages across all books?

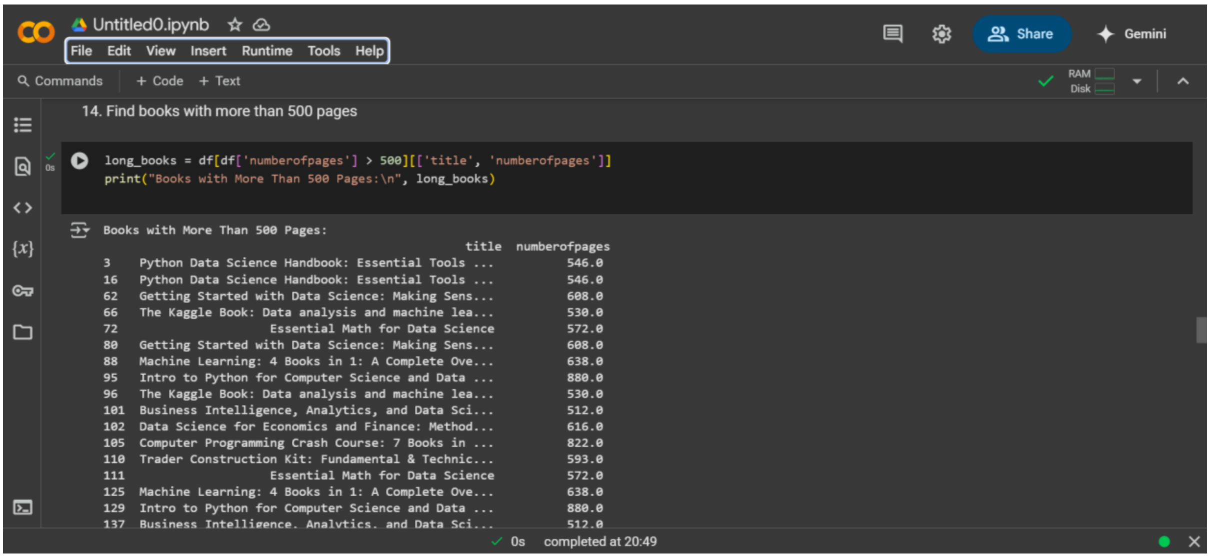
```
[25] total_pages = df['numberofpages'].sum()
     print("Total Number of Pages:", total_pages)
```

```
Total Number of Pages: 107693.0
```

13. How many books are in each book format category?

```
format_counts = df['bookformat'].value_counts()
print("Book Format Counts:\n", format_counts)
```

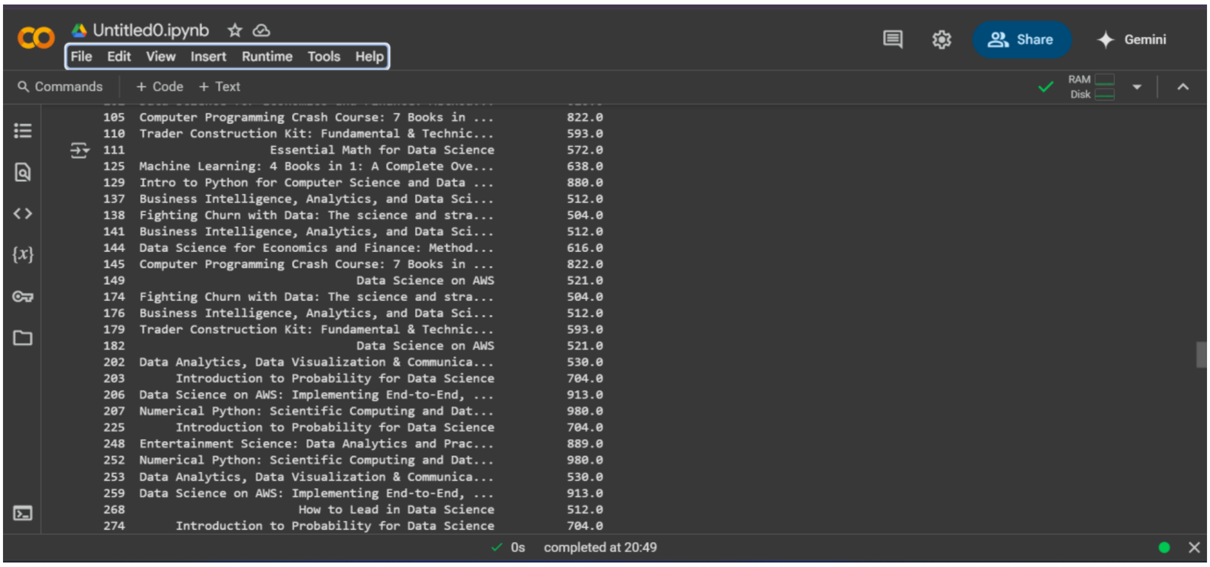
```
Book Format Counts:
 bookformat
Kindle Edition    208
Paperback         117
Hardcover          34
ebook              27
Name: count, dtype: int64
```



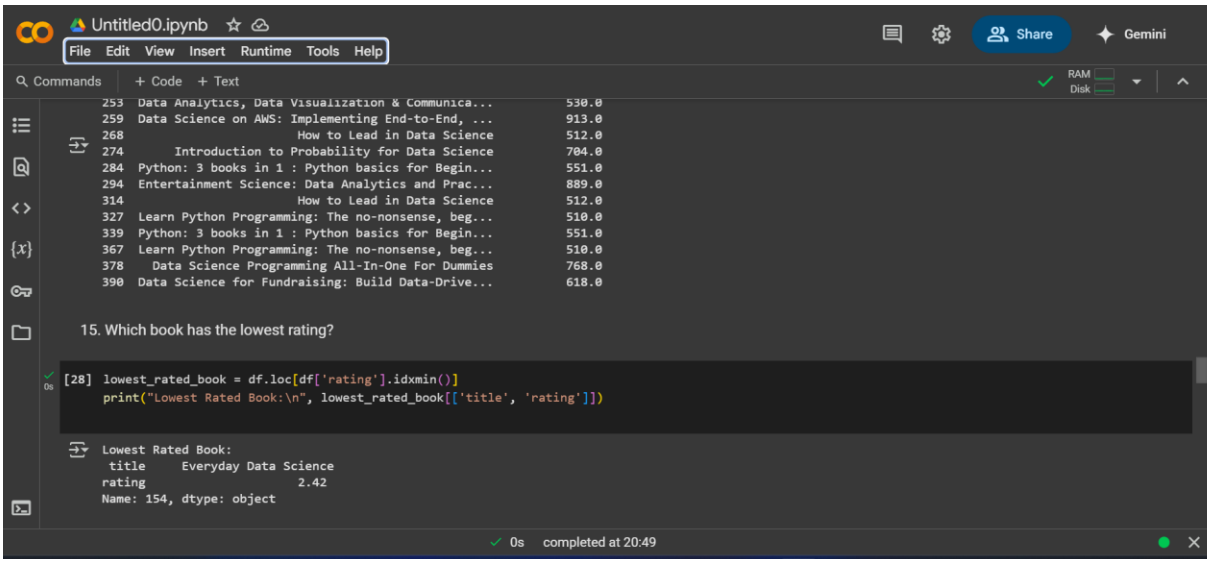
14. Find books with more than 500 pages

```
long_books = df[df['numberofpages'] > 500][['title', 'numberofpages']]
print("Books with More Than 500 Pages:\n", long_books)
```

```
Books with More Than 500 Pages:
                                                  title  numberofpages
3    Python Data Science Handbook: Essential Tools ...          546.0
16   Python Data Science Handbook: Essential Tools ...          546.0
62   Getting Started with Data Science: Making Sens...          608.0
66   The Kaggle Book: Data analysis and machine lea...          530.0
72                     Essential Math for Data Science          572.0
80   Getting Started with Data Science: Making Sens...          608.0
88   Machine Learning: 4 Books in 1: A Complete Ove...          638.0
95   Intro to Python for Computer Science and Data ...          880.0
96   The Kaggle Book: Data analysis and machine lea...          530.0
101  Business Intelligence, Analytics, and Data Sci...          512.0
102  Data Science for Economics and Finance: Method...          616.0
105  Computer Programming Crash Course: 7 Books in ...          822.0
110  Trader Construction Kit: Fundamental & Technic...          593.0
111                    Essential Math for Data Science          572.0
125  Machine Learning: 4 Books in 1: A Complete Ove...          638.0
129  Intro to Python for Computer Science and Data ...          880.0
137  Business Intelligence, Analytics, and Data Sci...          512.0
```



```
105  Computer Programming Crash Course: 7 Books in ...          822.0
110  Trader Construction Kit: Fundamental & Technic...          593.0
111                    Essential Math for Data Science          572.0
125  Machine Learning: 4 Books in 1: A Complete Ove...          638.0
129  Intro to Python for Computer Science and Data ...          880.0
137  Business Intelligence, Analytics, and Data Sci...          512.0
138  Fighting Churn with Data: The science and stra...          504.0
141  Business Intelligence, Analytics, and Data Sci...          512.0
144  Data Science for Economics and Finance: Method...          616.0
145  Computer Programming Crash Course: 7 Books in ...          822.0
149                              Data Science on AWS          521.0
174  Fighting Churn with Data: The science and stra...          504.0
176  Business Intelligence, Analytics, and Data Sci...          512.0
179  Trader Construction Kit: Fundamental & Technic...          593.0
182                              Data Science on AWS          521.0
202  Data Analytics, Data Visualization & Communica...          530.0
203       Introduction to Probability for Data Science          704.0
206  Data Science on AWS: Implementing End-to-End, ...          913.0
207  Numerical Python: Scientific Computing and Dat...          980.0
225       Introduction to Probability for Data Science          704.0
248  Entertainment Science: Data Analytics and Prac...          889.0
252  Numerical Python: Scientific Computing and Dat...          980.0
253  Data Analytics, Data Visualization & Communica...          530.0
259  Data Science on AWS: Implementing End-to-End, ...          913.0
268                        How to Lead in Data Science          512.0
274       Introduction to Probability for Data Science          704.0
```

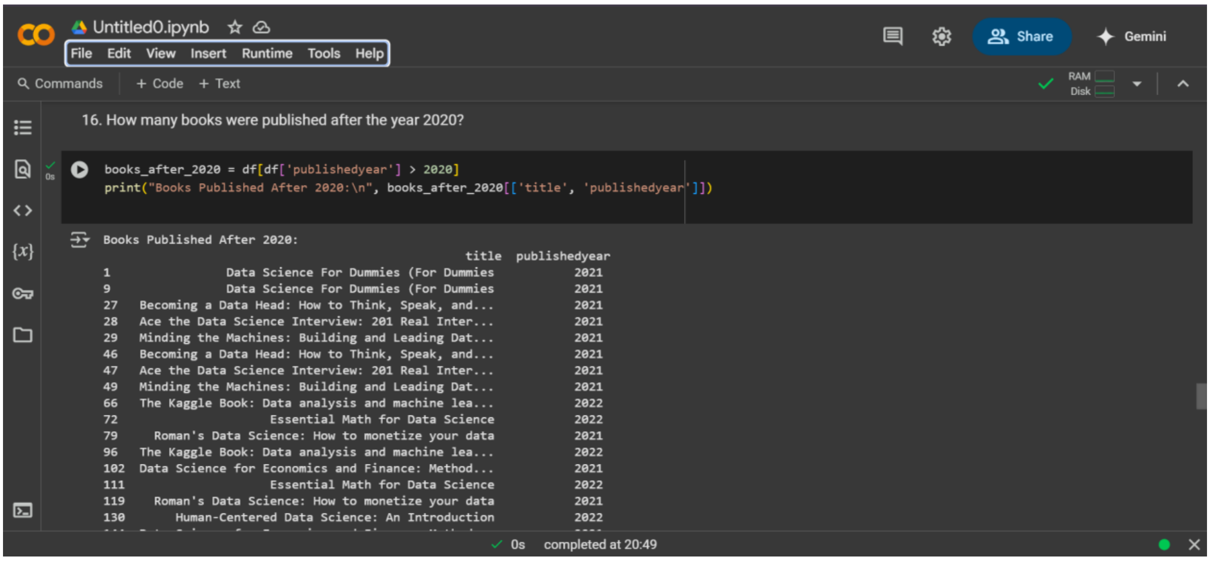


```
253  Data Analytics, Data Visualization & Communica...        530.0
259  Data Science on AWS: Implementing End-to-End, ...        913.0
268                        How to Lead in Data Science        512.0
274       Introduction to Probability for Data Science        704.0
284  Python: 3 books in 1 : Python basics for Begin...        551.0
294  Entertainment Science: Data Analytics and Prac...        889.0
314                        How to Lead in Data Science        512.0
327  Learn Python Programming: The no-nonsense, beg...        510.0
339  Python: 3 books in 1 : Python basics for Begin...        551.0
367  Learn Python Programming: The no-nonsense, beg...        510.0
378    Data Science Programming All-In-One For Dummies        768.0
390  Data Science for Fundraising: Build Data-Drive...        618.0
```

15. Which book has the lowest rating?

```
[28] lowest_rated_book = df.loc[df['rating'].idxmin()]
     print("Lowest Rated Book:\n", lowest_rated_book[['title', 'rating']])
```

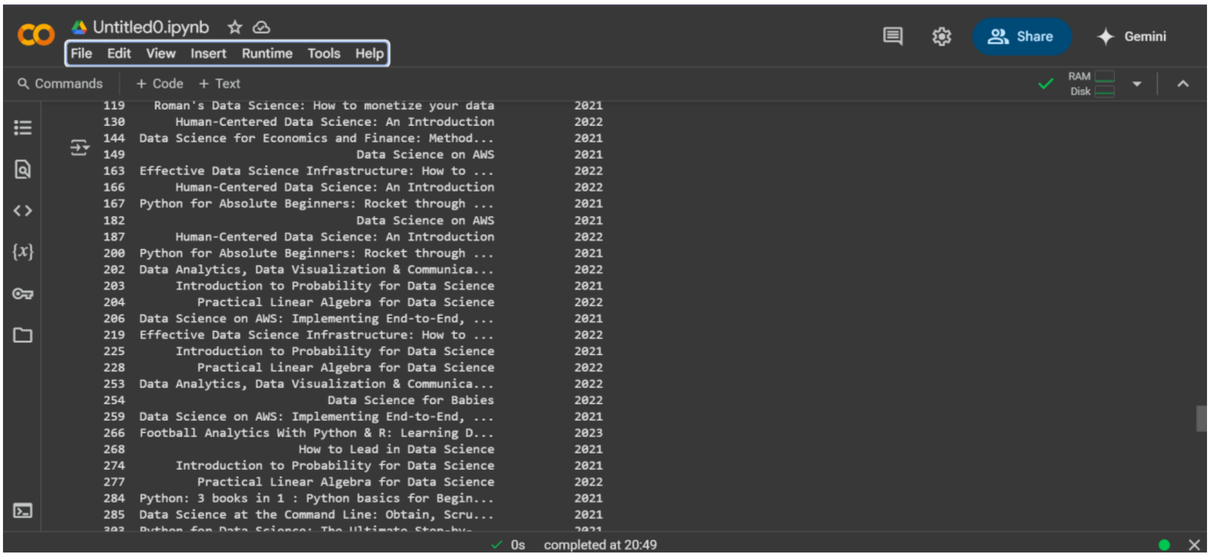
```
Lowest Rated Book:
 title     Everyday Data Science
rating                     2.42
Name: 154, dtype: object
```



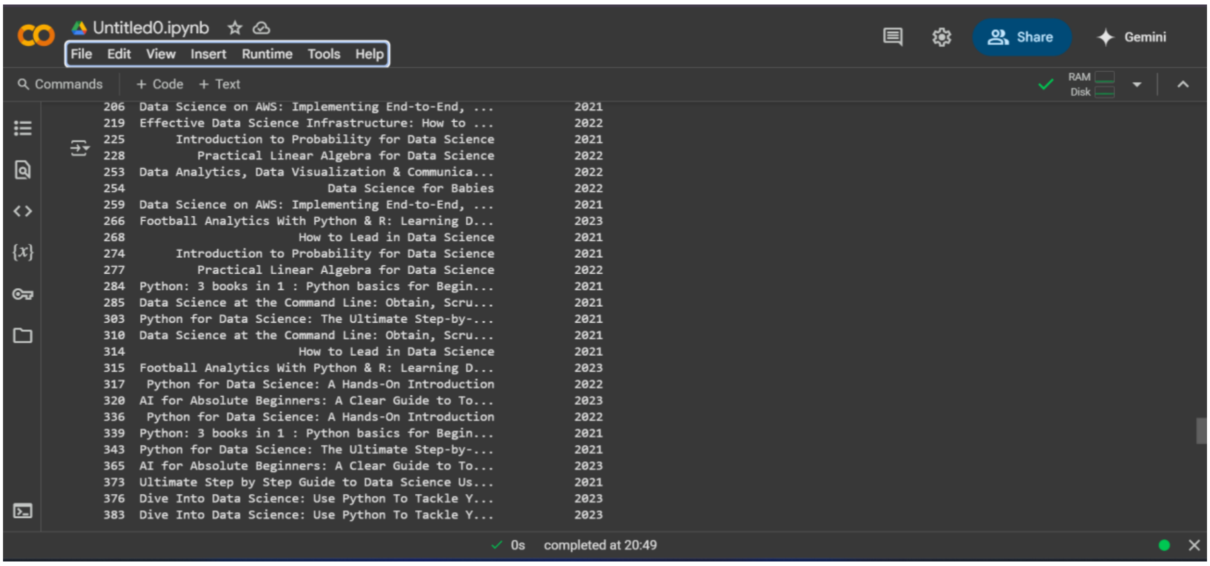
16. How many books were published after the year 2020?

```
books_after_2020 = df[df['publishedyear'] > 2020]
print("Books Published After 2020:\n", books_after_2020[['title', 'publishedyear']])
```

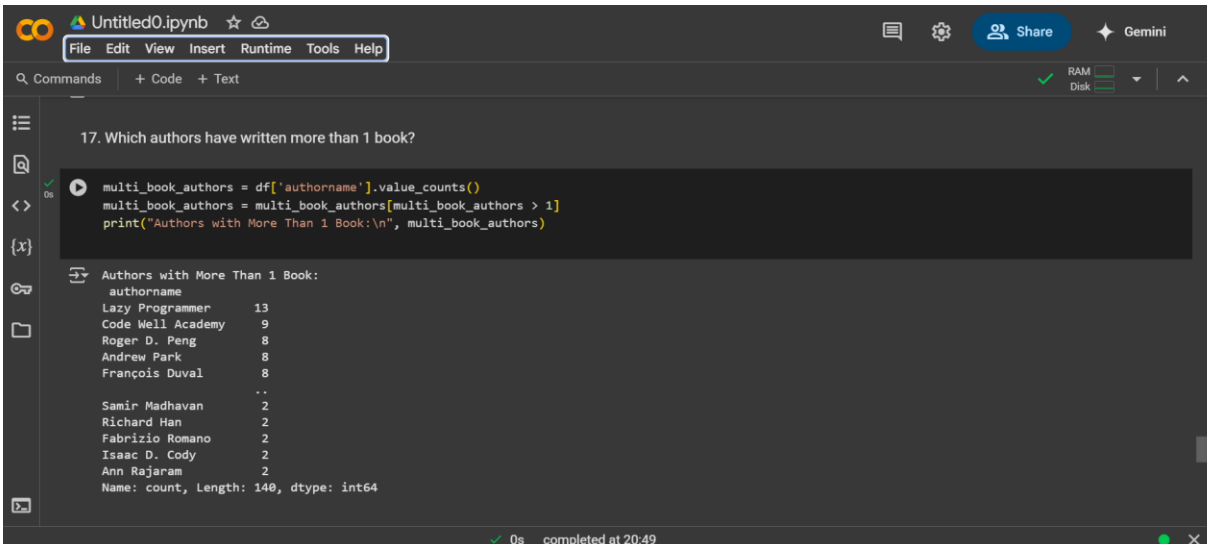
```
Books Published After 2020:
                                                 title  publishedyear
1                Data Science For Dummies (For Dummies           2021
9                Data Science For Dummies (For Dummies           2021
27   Becoming a Data Head: How to Think, Speak, and...           2021
28   Ace the Data Science Interview: 201 Real Inter...           2021
29   Minding the Machines: Building and Leading Dat...           2021
46   Becoming a Data Head: How to Think, Speak, and...           2021
47   Ace the Data Science Interview: 201 Real Inter...           2021
49   Minding the Machines: Building and Leading Dat...           2021
66   The Kaggle Book: Data analysis and machine lea...           2022
72                      Essential Math for Data Science           2022
79     Roman's Data Science: How to monetize your data           2021
96   The Kaggle Book: Data analysis and machine lea...           2022
102  Data Science for Economics and Finance: Method...           2021
111                     Essential Math for Data Science           2022
119    Roman's Data Science: How to monetize your data           2021
130       Human-Centered Data Science: An Introduction           2022
```



```
119    Roman's Data Science: How to monetize your data           2021
130       Human-Centered Data Science: An Introduction           2022
144  Data Science for Economics and Finance: Method...           2021
149                                Data Science on AWS           2021
163  Effective Data Science Infrastructure: How to ...           2022
166       Human-Centered Data Science: An Introduction           2022
167  Python for Absolute Beginners: Rocket through ...           2021
182                                Data Science on AWS           2021
187       Human-Centered Data Science: An Introduction           2022
200  Python for Absolute Beginners: Rocket through ...           2021
202  Data Analytics, Data Visualization & Communica...           2022
203       Introduction to Probability for Data Science           2021
204          Practical Linear Algebra for Data Science           2022
206  Data Science on AWS: Implementing End-to-End, ...           2021
219  Effective Data Science Infrastructure: How to ...           2022
225       Introduction to Probability for Data Science           2021
228          Practical Linear Algebra for Data Science           2022
253  Data Analytics, Data Visualization & Communica...           2022
254                            Data Science for Babies           2022
259  Data Science on AWS: Implementing End-to-End, ...           2021
266  Football Analytics With Python & R: Learning D...           2023
268                        How to Lead in Data Science           2021
274       Introduction to Probability for Data Science           2021
277          Practical Linear Algebra for Data Science           2022
284  Python: 3 books in 1 : Python basics for Begin...           2021
285  Data Science at the Command Line: Obtain, Scru...           2021
303  Python for Data Science: The Ultimate Step-by-...           2021
```



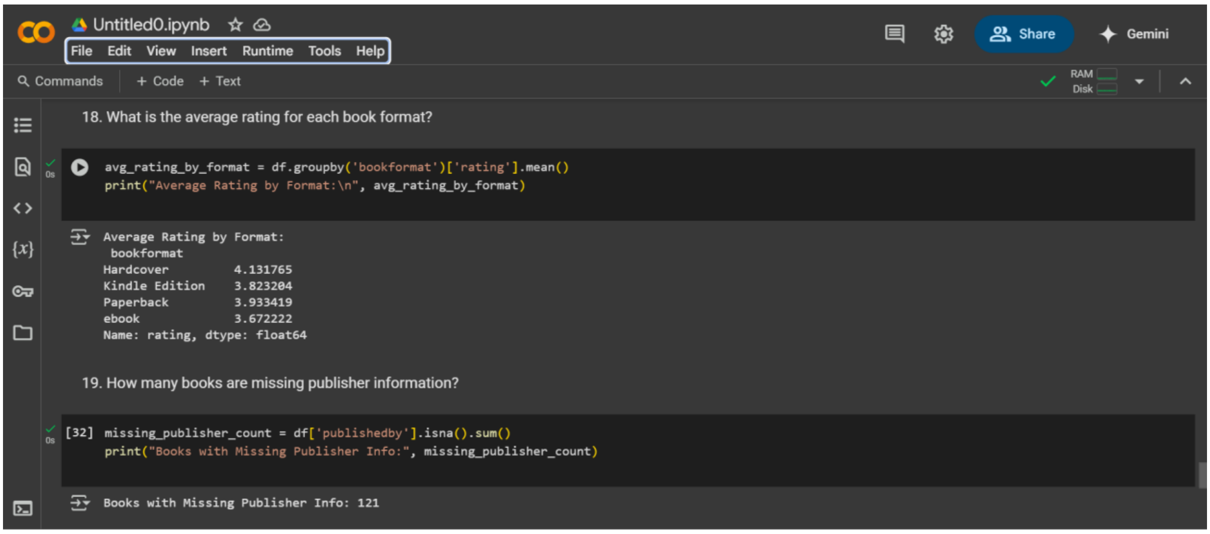
```
206  Data Science on AWS: Implementing End-to-End, ...        2021
219  Effective Data Science Infrastructure: How to ...        2022
225       Introduction to Probability for Data Science        2021
228          Practical Linear Algebra for Data Science        2022
253  Data Analytics, Data Visualization & Communica...        2022
254                           Data Science for Babies        2022
259  Data Science on AWS: Implementing End-to-End, ...        2021
266  Football Analytics With Python & R: Learning D...        2023
268                     How to Lead in Data Science        2021
274       Introduction to Probability for Data Science        2021
277          Practical Linear Algebra for Data Science        2022
284  Python: 3 books in 1 : Python basics for Begin...        2021
285  Data Science at the Command Line: Obtain, Scru...        2021
303  Python for Data Science: The Ultimate Step-by-...        2021
310  Data Science at the Command Line: Obtain, Scru...        2021
314                     How to Lead in Data Science        2021
315  Football Analytics With Python & R: Learning D...        2023
317   Python for Data Science: A Hands-On Introduction        2022
320  AI for Absolute Beginners: A Clear Guide to To...        2023
336   Python for Data Science: A Hands-On Introduction        2022
339  Python: 3 books in 1 : Python basics for Begin...        2021
343  Python for Data Science: The Ultimate Step-by-...        2021
365  AI for Absolute Beginners: A Clear Guide to To...        2023
373  Ultimate Step by Step Guide to Data Science Us...        2021
376  Dive Into Data Science: Use Python To Tackle Y...        2023
383  Dive Into Data Science: Use Python To Tackle Y...        2023
```



17. Which authors have written more than 1 book?

```
multi_book_authors = df['authorname'].value_counts()
multi_book_authors = multi_book_authors[multi_book_authors > 1]
print("Authors with More Than 1 Book:\n", multi_book_authors)
```

```
Authors with More Than 1 Book:
 authorname
Lazy Programmer      13
Code Well Academy     9
Roger D. Peng         8
Andrew Park           8
François Duval        8
                     ..
Samir Madhavan        2
Richard Han           2
Fabrizio Romano       2
Isaac D. Cody         2
Ann Rajaram           2
Name: count, Length: 140, dtype: int64
```



18. What is the average rating for each book format?

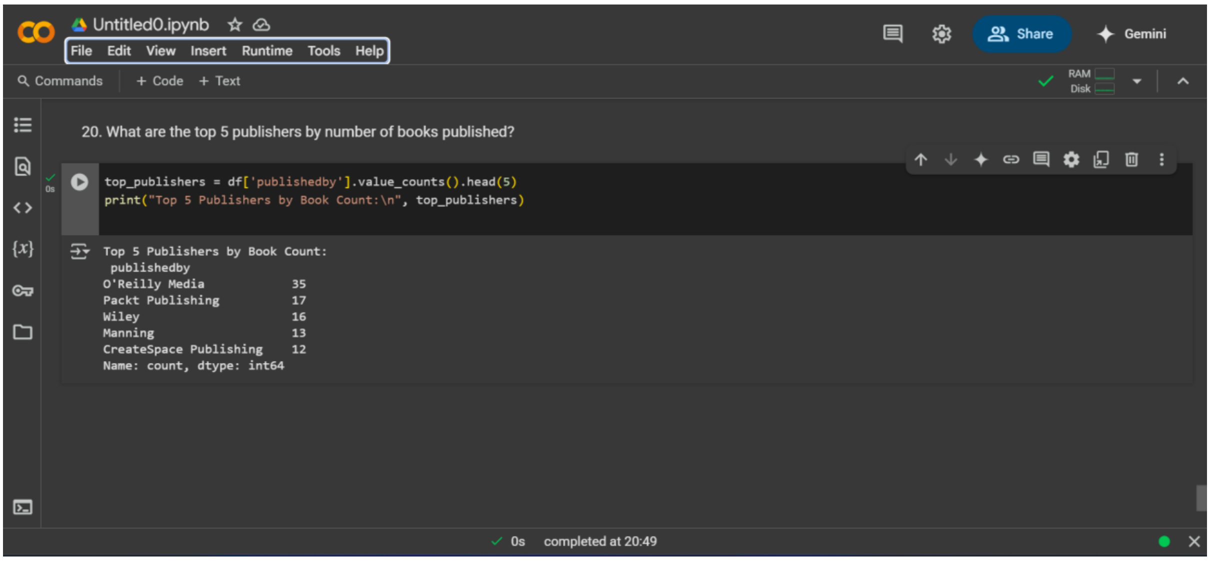
```
avg_rating_by_format = df.groupby('bookformat')['rating'].mean()
print("Average Rating by Format:\n", avg_rating_by_format)
```

```
Average Rating by Format:
 bookformat
Hardcover         4.131765
Kindle Edition    3.823204
Paperback         3.933419
ebook             3.672222
Name: rating, dtype: float64
```

19. How many books are missing publisher information?

```
[32] missing_publisher_count = df['publishedby'].isna().sum()
     print("Books with Missing Publisher Info:", missing_publisher_count)
```

```
Books with Missing Publisher Info: 121
```



20. What are the top 5 publishers by number of books published?

```
top_publishers = df['publishedby'].value_counts().head(5)
print("Top 5 Publishers by Book Count:\n", top_publishers)
```

```
Top 5 Publishers by Book Count:
 publishedby
O'Reilly Media          35
Packt Publishing        17
Wiley                   16
Manning                 13
CreateSpace Publishing  12
Name: count, dtype: int64
```