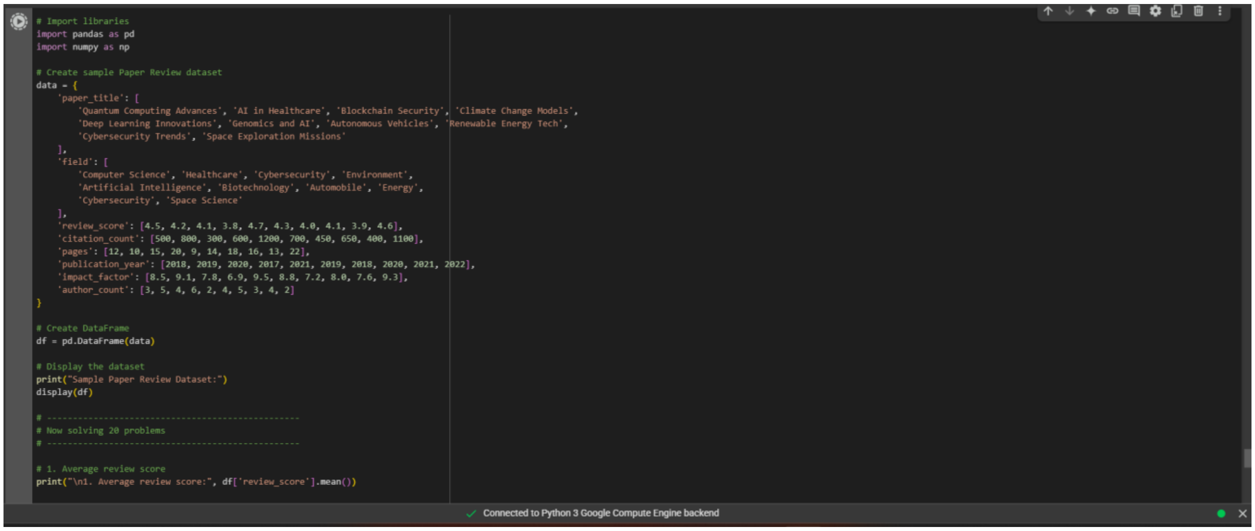
Atharv Korde

202401100050

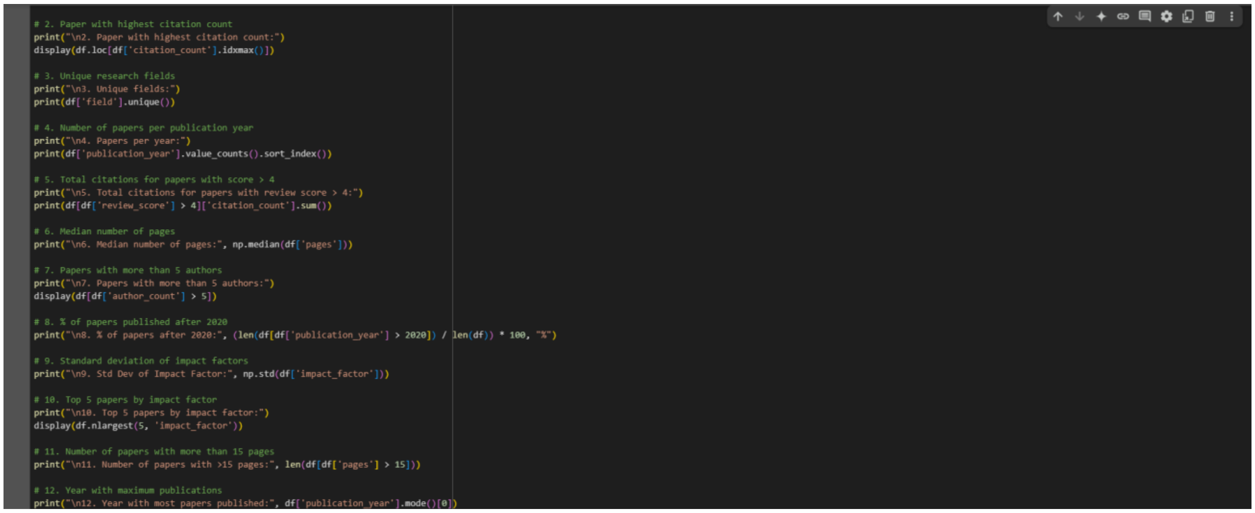
CS5-19

Paper review Dataset

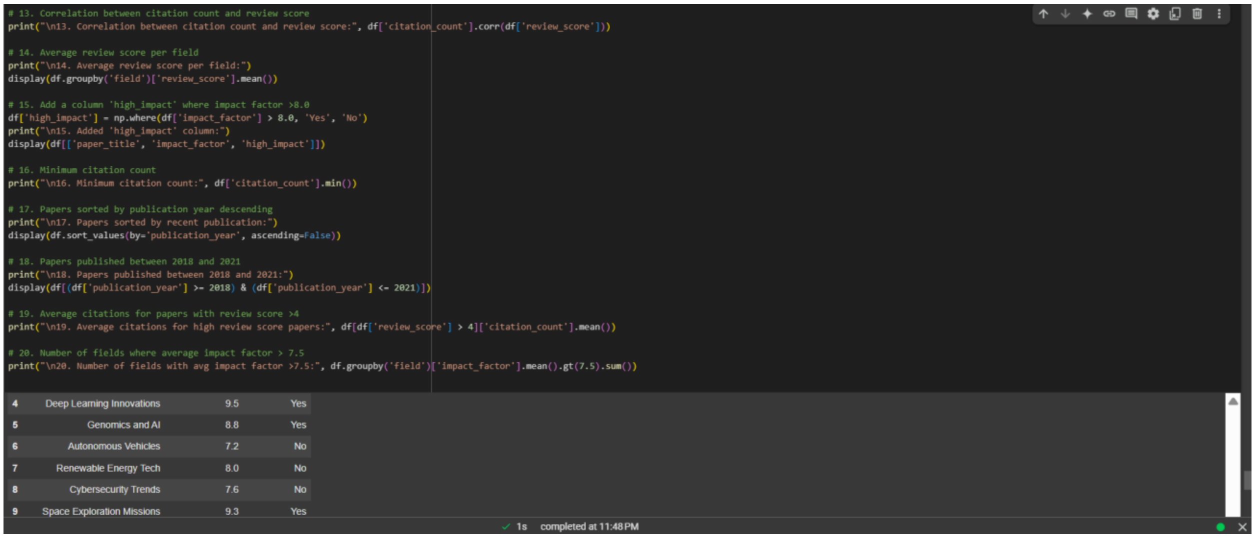


```
# Import libraries
import pandas as pd
import numpy as np

# Create sample Paper Review dataset
data = {
    'paper_title': [
        'Quantum Computing Advances', 'AI in Healthcare', 'Blockchain Security', 'Climate Change Models',
        'Deep Learning Innovations', 'Genomics and AI', 'Autonomous Vehicles', 'Renewable Energy Tech',
        'Cybersecurity Trends', 'Space Exploration Missions'
    ],
    'field': [
        'Computer Science', 'Healthcare', 'Cybersecurity', 'Environment',
        'Artificial Intelligence', 'Biotechnology', 'Automobile', 'Energy',
        'Cybersecurity', 'Space Science'
    ],
    'review_score': [4.5, 4.2, 4.1, 3.8, 4.7, 4.3, 4.0, 4.1, 3.9, 4.6],
    'citation_count': [500, 800, 300, 600, 1200, 700, 450, 650, 400, 1100],
    'pages': [12, 10, 15, 20, 9, 14, 18, 16, 13, 22],
    'publication_year': [2018, 2019, 2020, 2017, 2021, 2019, 2018, 2020, 2021, 2022],
    'impact_factor': [8.5, 9.1, 7.8, 6.9, 9.5, 8.8, 7.2, 8.0, 7.6, 9.3],
    'author_count': [3, 5, 4, 6, 2, 4, 5, 3, 4, 2]
}

# Create DataFrame
df = pd.DataFrame(data)

# Display the dataset
print("Sample Paper Review Dataset:")
display(df)

# ------------------------------------------------
# Now solving 20 problems
# ------------------------------------------------

# 1. Average review score
print("\n1. Average review score:", df['review_score'].mean())
```

Connected to Python 3 Google Compute Engine backend

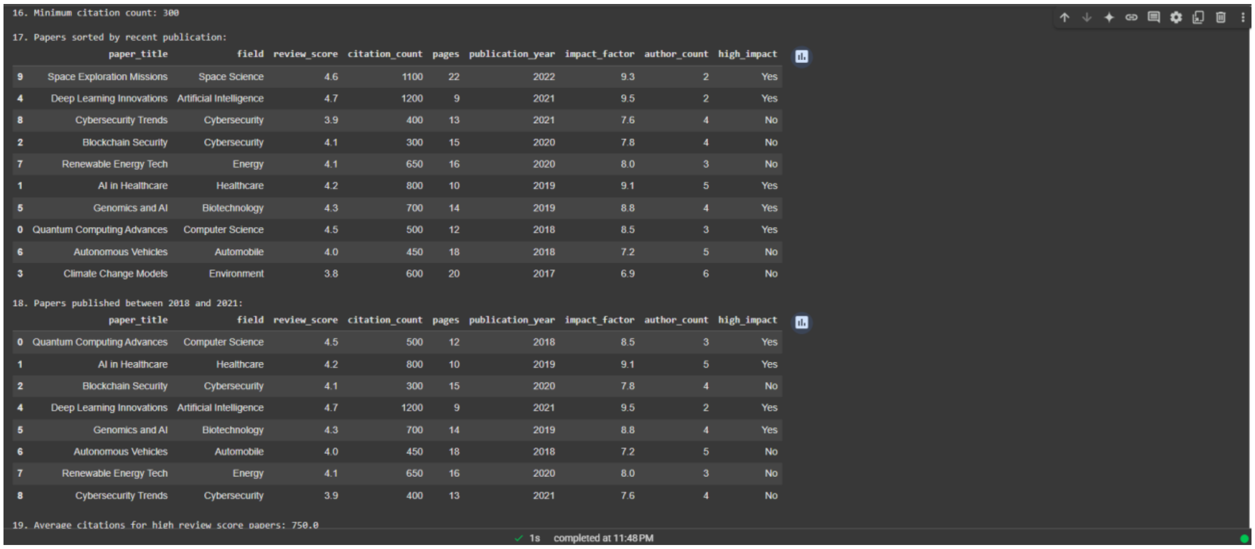


```
# 2. Paper with highest citation count
print("\n2. Paper with highest citation count:")
display(df.loc[df['citation_count'].idxmax()])

# 3. Unique research fields
print("\n3. Unique fields:")
print(df['field'].unique())

# 4. Number of papers per publication year
print("\n4. Papers per year:")
print(df['publication_year'].value_counts().sort_index())

# 5. Total citations for papers with score > 4
print("\n5. Total citations for papers with review score > 4:")
print(df[df['review_score'] > 4]['citation_count'].sum())

# 6. Median number of pages
print("\n6. Median number of pages:", np.median(df['pages']))

# 7. Papers with more than 5 authors
print("\n7. Papers with more than 5 authors:")
display(df[df['author_count'] > 5])

# 8. % of papers published after 2020
print("\n8. % of papers after 2020:", (len(df[df['publication_year'] > 2020]) / len(df)) * 100, "%")

# 9. Standard deviation of impact factors
print("\n9. Std Dev of Impact Factor:", np.std(df['impact_factor']))

# 10. Top 5 papers by impact factor
print("\n10. Top 5 papers by impact factor:")
display(df.nlargest(5, 'impact_factor'))

# 11. Number of papers with more than 15 pages
print("\n11. Number of papers with >15 pages:", len(df[df['pages'] > 15]))

# 12. Year with maximum publications
print("\n12. Year with most papers published:", df['publication_year'].mode()[0])
```



```python
# 13. Correlation between citation count and review score
print("\n13. Correlation between citation count and review score:", df['citation_count'].corr(df['review_score']))

# 14. Average review score per field
print("\n14. Average review score per field:")
display(df.groupby('field')['review_score'].mean())

# 15. Add a column 'high_impact' where impact factor >8.0
df['high_impact'] = np.where(df['impact_factor'] > 8.0, 'Yes', 'No')
print("\n15. Added 'high_impact' column:")
display(df[['paper_title', 'impact_factor', 'high_impact']])

# 16. Minimum citation count
print("\n16. Minimum citation count:", df['citation_count'].min())

# 17. Papers sorted by publication year descending
print("\n17. Papers sorted by recent publication:")
display(df.sort_values(by='publication_year', ascending=False))

# 18. Papers published between 2018 and 2021
print("\n18. Papers published between 2018 and 2021:")
display(df[(df['publication_year'] >= 2018) & (df['publication_year'] <= 2021)])

# 19. Average citations for papers with review score >4
print("\n19. Average citations for high review score papers:", df[df['review_score'] > 4]['citation_count'].mean())

# 20. Number of fields where average impact factor > 7.5
print("\n20. Number of fields with avg impact factor >7.5:", df.groupby('field')['impact_factor'].mean().gt(7.5).sum())
```

4	Deep Learning Innovations	9.5	Yes
5	Genomics and AI	8.8	Yes
6	Autonomous Vehicles	7.2	No
7	Renewable Energy Tech	8.0	No
8	Cybersecurity Trends	7.6	No
9	Space Exploration Missions	9.3	Yes



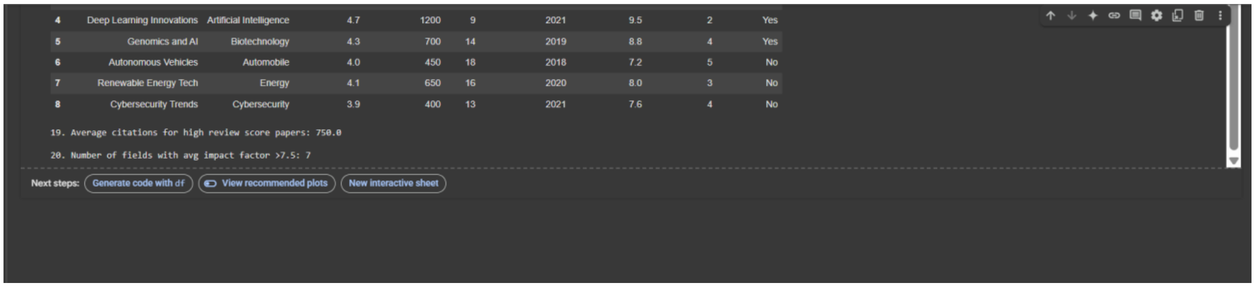
```
16. Minimum citation count: 300

17. Papers sorted by recent publication:
```

	paper_title	field	review_score	citation_count	pages	publication_year	impact_factor	author_count	high_impact
9	Space Exploration Missions	Space Science	4.6	1100	22	2022	9.3	2	Yes
4	Deep Learning Innovations	Artificial Intelligence	4.7	1200	9	2021	9.5	2	Yes
8	Cybersecurity Trends	Cybersecurity	3.9	400	13	2021	7.6	4	No
2	Blockchain Security	Cybersecurity	4.1	300	15	2020	7.8	4	No
7	Renewable Energy Tech	Energy	4.1	650	16	2020	8.0	3	No
1	AI in Healthcare	Healthcare	4.2	800	10	2019	9.1	5	Yes
5	Genomics and AI	Biotechnology	4.3	700	14	2019	8.8	4	Yes
0	Quantum Computing Advances	Computer Science	4.5	500	12	2018	8.5	3	Yes
6	Autonomous Vehicles	Automobile	4.0	450	18	2018	7.2	5	No
3	Climate Change Models	Environment	3.8	600	20	2017	6.9	6	No

```
18. Papers published between 2018 and 2021:
```

	paper_title	field	review_score	citation_count	pages	publication_year	impact_factor	author_count	high_impact
0	Quantum Computing Advances	Computer Science	4.5	500	12	2018	8.5	3	Yes
1	AI in Healthcare	Healthcare	4.2	800	10	2019	9.1	5	Yes
2	Blockchain Security	Cybersecurity	4.1	300	15	2020	7.8	4	No
4	Deep Learning Innovations	Artificial Intelligence	4.7	1200	9	2021	9.5	2	Yes
5	Genomics and AI	Biotechnology	4.3	700	14	2019	8.8	4	Yes
6	Autonomous Vehicles	Automobile	4.0	450	18	2018	7.2	5	No
7	Renewable Energy Tech	Energy	4.1	650	16	2020	8.0	3	No
8	Cybersecurity Trends	Cybersecurity	3.9	400	13	2021	7.6	4	No

```
19. Average citations for high review score papers: 750.0
```



4	Deep Learning Innovations	Artificial Intelligence	4.7	1200	9	2021	9.5	2	Yes
5	Genomics and AI	Biotechnology	4.3	700	14	2019	8.8	4	Yes
6	Autonomous Vehicles	Automobile	4.0	450	18	2018	7.2	5	No
7	Renewable Energy Tech	Energy	4.1	650	16	2020	8.0	3	No
8	Cybersecurity Trends	Cybersecurity	3.9	400	13	2021	7.6	4	No

```
19. Average citations for high review score papers: 750.0

20. Number of fields with avg impact factor >7.5: 7
```