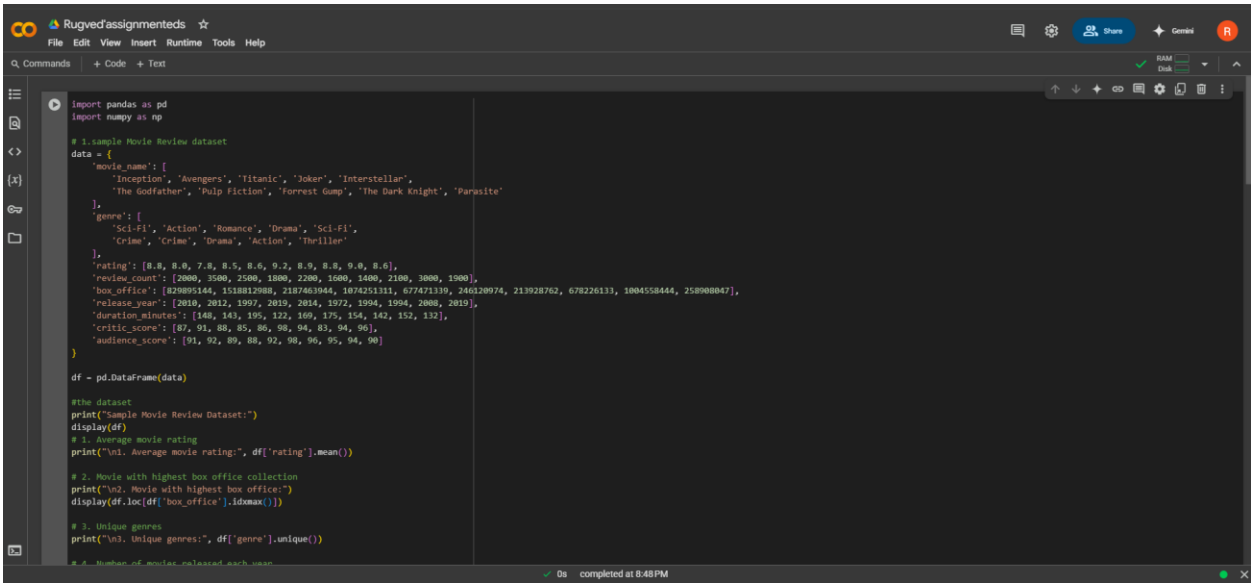
Rugved Suryawanshi

202401100041

CS5-14

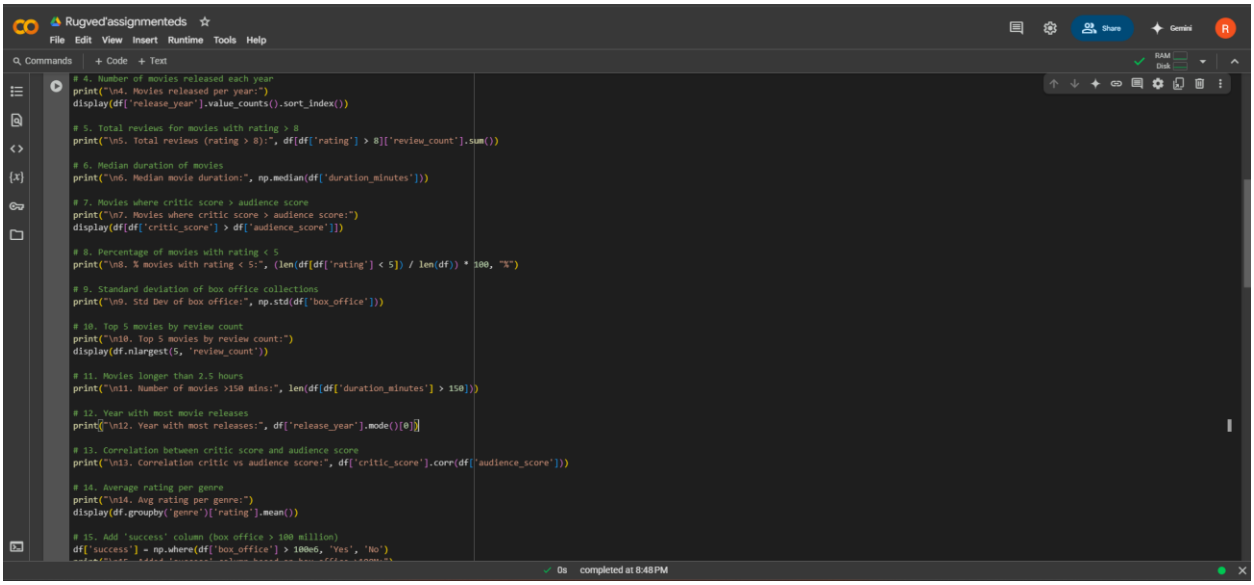
Dataset: Movie review

Below is the code and statements in form of comments in it (I have used Google Colab for running the code)



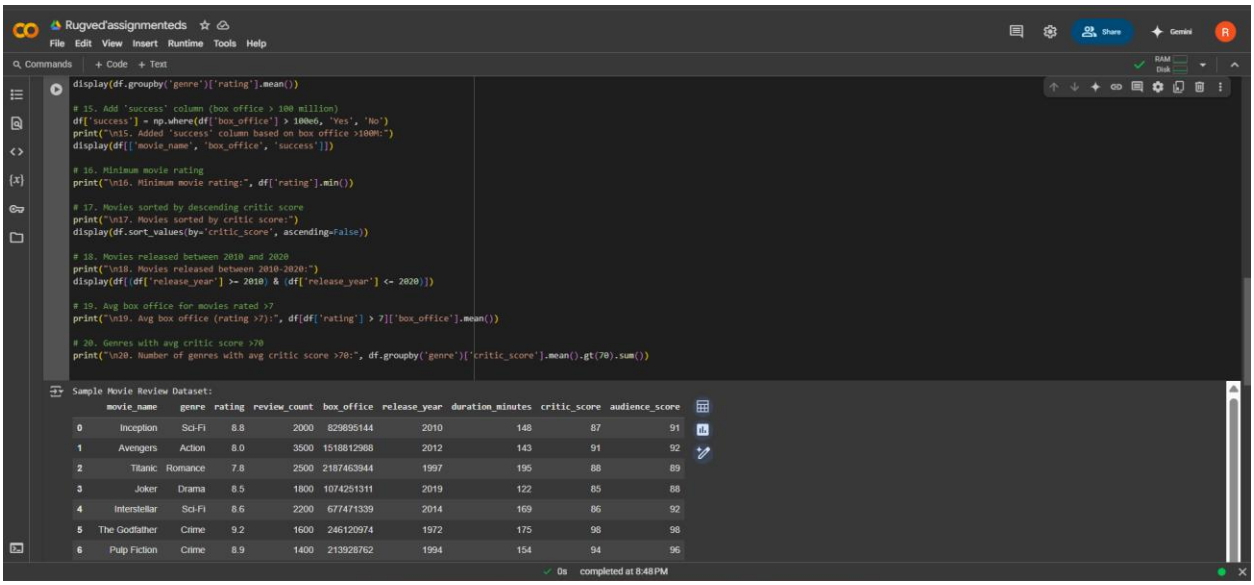
```
import pandas as pd
import numpy as np

# 1.sample Movie Review dataset
data = {
    'movie_name': [
        'Inception', 'Avengers', 'Titanic', 'Joker', 'Interstellar',
        'The Godfather', 'Pulp Fiction', 'Forrest Gump', 'The Dark Knight', 'Parasite'
    ],
    'genre': [
        'Sci-Fi', 'Action', 'Romance', 'Drama', 'Sci-Fi',
        'Crime', 'Crime', 'Drama', 'Action', 'Thriller'
    ],
    'rating': [8.8, 8.0, 7.8, 8.5, 8.6, 9.2, 8.9, 8.8, 9.0, 8.6],
    'review_count': [2000, 3500, 2500, 1800, 2200, 1600, 1400, 2100, 3000, 1900],
    'box_office': [829895144, 1518812988, 2187463944, 1074251311, 677471339, 246120974, 213928762, 678226133, 1004558444, 258908047],
    'release_year': [2010, 2012, 1997, 2019, 2014, 1972, 1994, 1994, 2008, 2019],
    'duration_minutes': [148, 143, 195, 122, 169, 175, 154, 142, 152, 132],
    'critic_score': [87, 91, 88, 85, 86, 98, 94, 83, 94, 96],
    'audience_score': [91, 92, 89, 88, 92, 98, 96, 95, 94, 90]
}

df = pd.DataFrame(data)

#the dataset
print("Sample Movie Review Dataset:")
display(df)
# 1. Average movie rating
print("\n1. Average movie rating:", df['rating'].mean())

# 2. Movie with highest box office collection
print("\n2. Movie with highest box office:")
display(df.loc[df['box_office'].idxmax()])

# 3. Unique genres
print("\n3. Unique genres:", df['genre'].unique())

# 4. Number of movies released each year
```

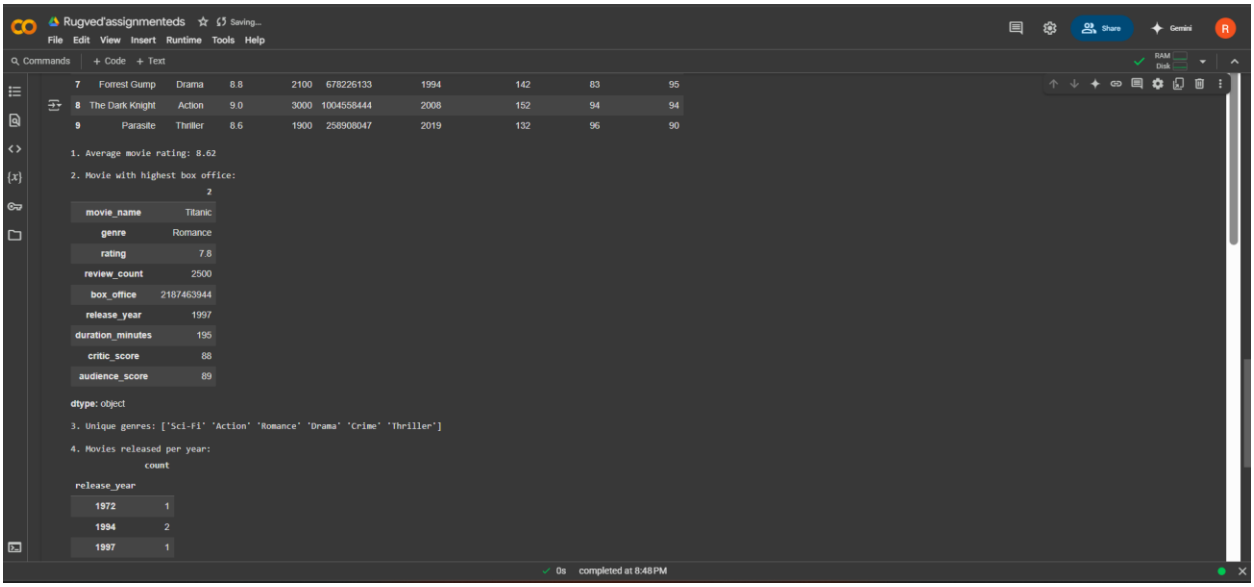
0s completed at 8:48PM

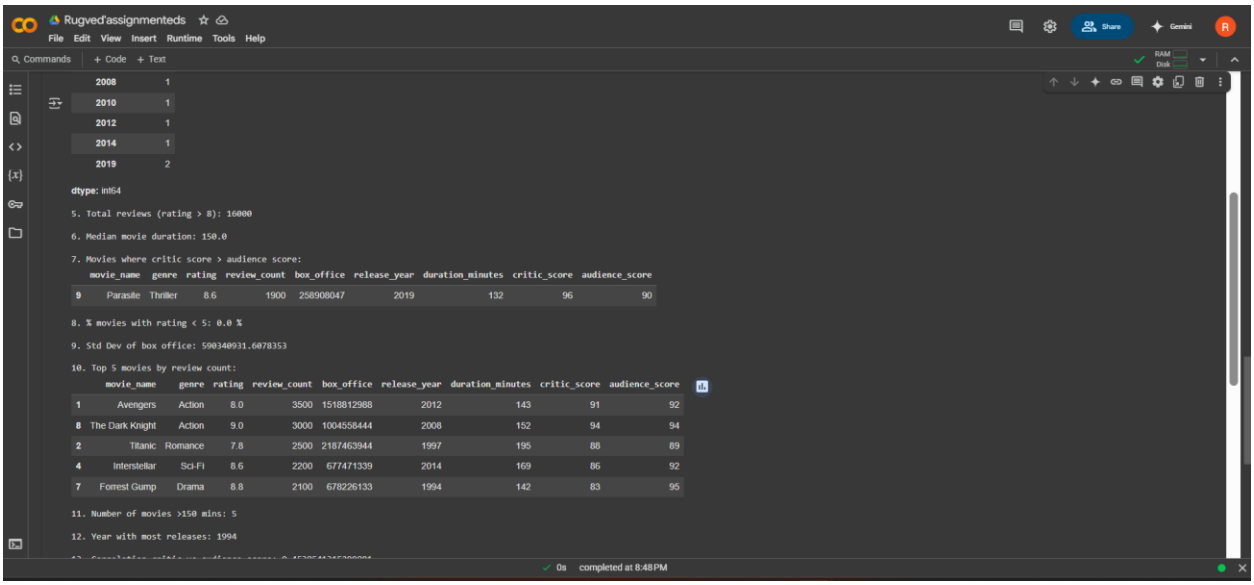


```
# 4. Number of movies released each year
print("\n4. Movies released per year:")
display(df['release_year'].value_counts().sort_index())

# 5. Total reviews for movies with rating > 8
print("\n5. Total reviews (rating > 8):", df[df['rating'] > 8]['review_count'].sum())

# 6. Median duration of movies
print("\n6. Median movie duration:", np.median(df['duration_minutes']))

# 7. Movies where critic score > audience score
print("\n7. Movies where critic score > audience score:")
display(df[df['critic_score'] > df['audience_score']])

# 8. Percentage of movies with rating < 5
print("\n8. % movies with rating < 5:", (len(df[df['rating'] < 5]) / len(df)) * 100, "%")

# 9. Standard deviation of box office collections
print("\n9. Std Dev of box office:", np.std(df['box_office']))

# 10. Top 5 movies by review count
print("\n10. Top 5 movies by review count:")
display(df.nlargest(5, 'review_count'))

# 11. Movies longer than 2.5 hours
print("\n11. Number of movies >150 mins:", len(df[df['duration_minutes'] > 150]))

# 12. Year with most movie releases
print("\n12. Year with most releases:", df['release_year'].mode()[0])

# 13. Correlation between critic score and audience score
print("\n13. Correlation critic vs audience score:", df['critic_score'].corr(df['audience_score']))

# 14. Average rating per genre
print("\n14. Avg rating per genre:")
display(df.groupby('genre')['rating'].mean())

# 15. Add 'success' column (box office > 100 million)
df['success'] = np.where(df['box_office'] > 100e6, 'Yes', 'No')
```

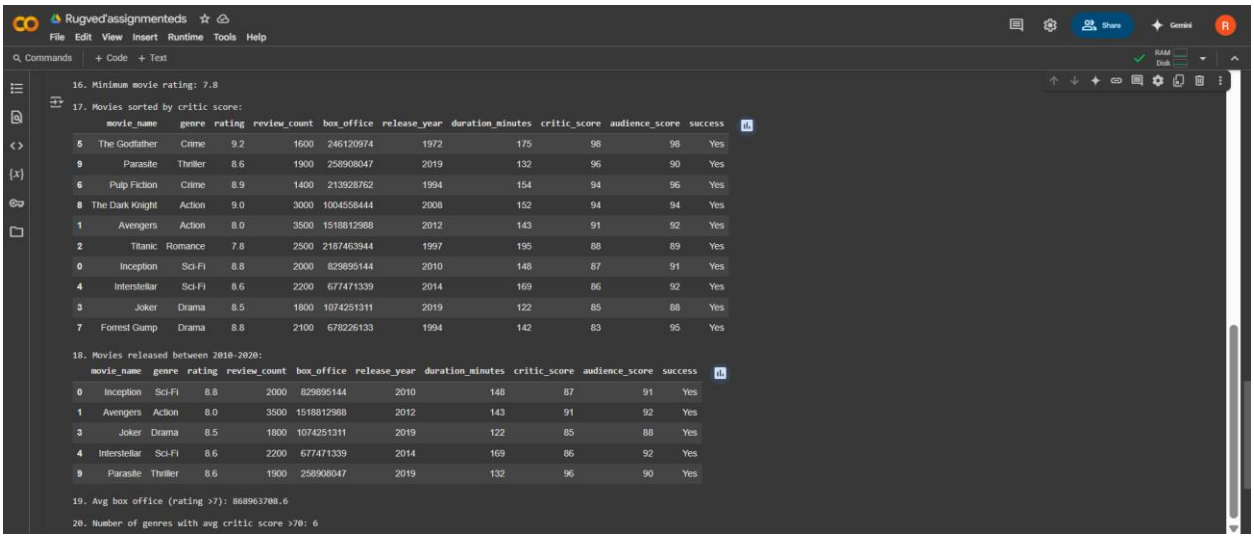
0s completed at 8:48PM


Rugved'assignmenteds
File Edit View Insert Runtime Tools Help
Commands + Code + Text
display(df.groupby('genre')['rating'].mean())
# 15. Add 'success' column (box office > 100 million)
df['success'] = np.where(df['box_office'] > 100e6, 'Yes', 'No')
print("\n15. Added 'success' column based on box office >100M:")
display(df[['movie_name', 'box_office', 'success']])
# 16. Minimum movie rating
print("\n16. Minimum movie rating:", df['rating'].min())
# 17. Movies sorted by descending critic score
print("\n17. Movies sorted by critic score:")
display(df.sort_values(by='critic_score', ascending=False))
# 18. Movies released between 2010 and 2020
print("\n18. Movies released between 2010-2020:")
display(df[(df['release_year'] >= 2010) & (df['release_year'] <= 2020)])
# 19. Avg box office for movies rated >7
print("\n19. Avg box office (rating >7):", df[df['rating'] > 7]['box_office'].mean())
# 20. Genres with avg critic score >70
print("\n20. Number of genres with avg critic score >70:", df.groupby('genre')['critic_score'].mean().gt(70).sum())
Sample Movie Review Dataset:
movie_name genre rating review_count box_office release_year duration_minutes critic_score audience_score
0 Inception Sci-Fi 8.8 2000 829895144 2010 148 87 91
1 Avengers Action 8.0 3500 1518812988 2012 143 91 92
2 Titanic Romance 7.8 2500 2187463944 1997 195 88 89
3 Joker Drama 8.5 1800 1074251311 2019 122 85 88
4 Interstellar Sci-Fi 8.6 2200 677471339 2014 169 86 92
5 The Godfather Crime 9.2 1600 246120974 1972 175 98 98
6 Pulp Fiction Crime 8.9 1400 213928762 1994 154 94 96
0s completed at 8:48PM


Rugved'assignmenteds
Saving...
File Edit View Insert Runtime Tools Help
Commands + Code + Text
7 Forrest Gump Drama 8.8 2100 678226133 1994 142 83 95
8 The Dark Knight Action 9.0 3000 1004558444 2008 152 94 94
9 Parasite Thriller 8.6 1900 258908047 2019 132 96 90
1. Average movie rating: 8.62
2. Movie with highest box office:
2
movie_name Titanic
genre Romance
rating 7.8
review_count 2500
box_office 2187463944
release_year 1997
duration_minutes 195
critic_score 88
audience_score 89
dtype: object
3. Unique genres: ['Sci-Fi' 'Action' 'Romance' 'Drama' 'Crime' 'Thriller']
4. Movies released per year:
count
release_year
1972 1
1994 2
1997 1
0s completed at 8:48PM


Rugved'assignmenteds
File Edit View Insert Runtime Tools Help
Commands + Code + Text
2008 1
2010 1
2012 1
2014 1
2019 2
dtype: int64
5. Total reviews (rating > 8): 16000
6. Median movie duration: 150.0
7. Movies where critic score > audience score:
movie_name genre rating review_count box_office release_year duration_minutes critic_score audience_score
9 Parasite Thriller 8.6 1900 258908047 2019 132 96 90
8. % movies with rating < 5: 0.0 %
9. Std Dev of box office: 590340931.6078353
10. Top 5 movies by review count:
movie_name genre rating review_count box_office release_year duration_minutes critic_score audience_score
1 Avengers Action 8.0 3500 1518812988 2012 143 91 92
8 The Dark Knight Action 9.0 3000 1004558444 2008 152 94 94
2 Titanic Romance 7.8 2500 2187463944 1997 195 88 89
4 Interstellar Sci-Fi 8.6 2200 677471339 2014 169 86 92
7 Forrest Gump Drama 8.8 2100 678226133 1994 142 83 95
11. Number of movies >150 mins: 5
12. Year with most releases: 1994
0s completed at 8:48PM


Rugved'assignmenteds
File Edit View Insert Runtime Tools Help
Commands + Code + Text
13. Correlation critic vs audience score: 0.4538541315299801
14. Avg rating per genre:
rating
genre
Action 8.50
Crime 9.05
Drama 8.65
Romance 7.80
Sci-Fi 8.70
Thriller 8.60
dtype: float64
15. Added 'success' column based on box office >100M:
movie_name box_office success
0 Inception 829895144 Yes
1 Avengers 1518812988 Yes
2 Titanic 2187463944 Yes
3 Joker 1074251311 Yes
4 Interstellar 677471339 Yes
5 The Godfather 246120974 Yes
6 Pulp Fiction 213928762 Yes
7 Forrest Gump 678226133 Yes
8 The Dark Knight 1004558444 Yes
9 Parasite 258908047 Yes
0s completed at 8:48PM


Rugved'assignmenteds
File Edit View Insert Runtime Tools Help
Commands + Code + Text
16. Minimum movie rating: 7.8
17. Movies sorted by critic score:
movie_name genre rating review_count box_office release_year duration_minutes critic_score audience_score success
5 The Godfather Crime 9.2 1600 246120974 1972 175 98 98 Yes
9 Parasite Thriller 8.6 1900 258908047 2019 132 96 90 Yes
6 Pulp Fiction Crime 8.9 1400 213928762 1994 154 94 96 Yes
8 The Dark Knight Action 9.0 3000 1004558444 2008 152 94 94 Yes
1 Avengers Action 8.0 3500 1518812988 2012 143 91 92 Yes
2 Titanic Romance 7.8 2500 2187463944 1997 195 88 89 Yes
0 Inception Sci-Fi 8.8 2000 829895144 2010 148 87 91 Yes
4 Interstellar Sci-Fi 8.6 2200 677471339 2014 169 86 92 Yes
3 Joker Drama 8.5 1800 1074251311 2019 122 85 88 Yes
7 Forrest Gump Drama 8.8 2100 678226133 1994 142 83 95 Yes
18. Movies released between 2010-2020:
movie_name genre rating review_count box_office release_year duration_minutes critic_score audience_score success
0 Inception Sci-Fi 8.8 2000 829895144 2010 148 87 91 Yes
1 Avengers Action 8.0 3500 1518812988 2012 143 91 92 Yes
3 Joker Drama 8.5 1800 1074251311 2019 122 85 88 Yes
4 Interstellar Sci-Fi 8.6 2200 677471339 2014 169 86 92 Yes
9 Parasite Thriller 8.6 1900 258908047 2019 132 96 90 Yes
19. Avg box office (rating >7): 868963708.6
20. Number of genres with avg critic score >70: 6