

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN



BÁO CÁO LAB 2: SPARK

GVHD: CN. Nguyễn Hiếu Nghĩa

Lớp: IE212.Q11

Sinh viên thực hiện:

Họ và tên: Trương Hoài Bảo

MSSV: 22520126

☞ Tp. Hồ Chí Minh, 11/2025 ☞

[illegible]

(Ký tên và ghi rõ họ tên)

MỤC LỤC

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN	ii
DANH MỤC HÌNH ẢNH.....	iv
NỘI DUNG THỰC HÀNH.....	5
1.1 Bài 1: Tính Điểm Đánh Giá Trung Bình và Tổng Số Lượt Đánh Giá Cho Mỗi Phim.....	5
1.2 Bài 2: Phân Tích Đánh Giá Theo Thể Loại	5
1.3 Bài 3: Phân Tích Đánh Giá Theo Giới Tính.....	6
1.4 Bài 4 : Phân Tích Đánh Giá Theo Nhóm Tuổi	6
1.5 Bài 5: Phân Tích Đánh Giá Theo Occupation (Nghề nghiệp) Của Người Dùng.....	7
1.6 Bài 6: Phân Tích Đánh Giá Theo Thời Gian	7
TÀI LIỆU THAM KHẢO	8

DANH MỤC HÌNH ẢNH

Hình 1.1-1 Kết quả câu 1.....5

Hình 1.2-1 Kết quả câu 2.....5

Hình 1.3-1 Kết quả câu 3.....6

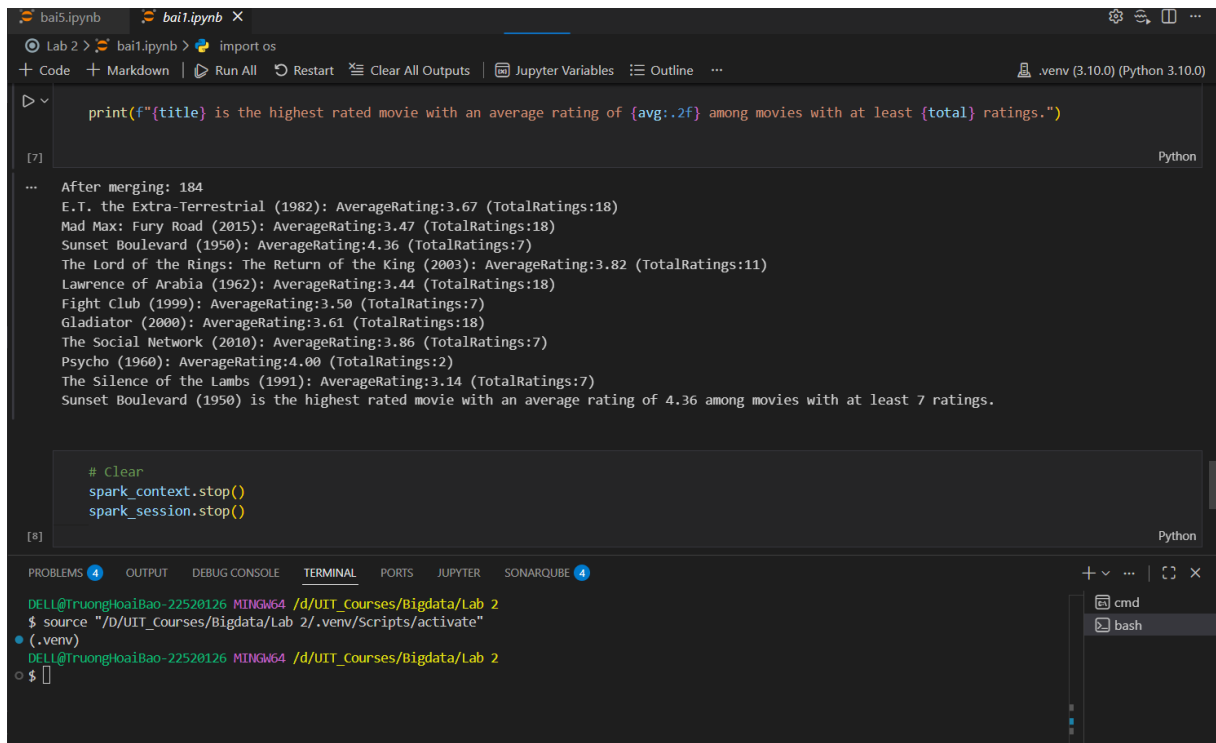
Hình 1.4-1 Kết quả câu 4.....6

Hình 1.5-1 Kết quả câu 5.....7

Hình 1.6-1 Kết quả câu 6.....7

NỘI DUNG THỰC HÀNH

1.1 Bài 1: Tính Điểm Đánh Giá Trung Bình và Tổng Số Lượt Đánh Giá Cho Mỗi Phim



The screenshot shows a Jupyter Notebook interface with two cells. The first cell contains a Python script that prints the highest rated movie with an average rating of at least 4.36 among movies with at least 7 ratings. The second cell contains a Python script that prints the average rating and total ratings for each genre.

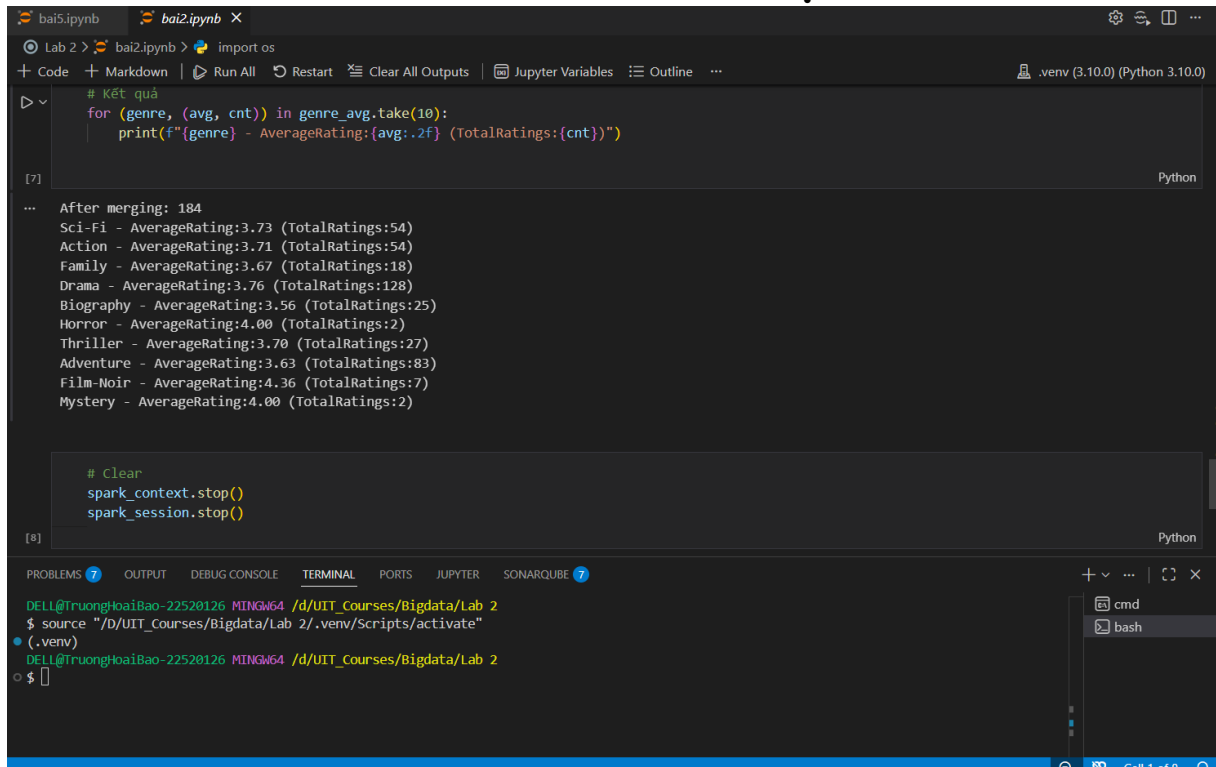
```
print(f"{title} is the highest rated movie with an average rating of {avg:.2f} among movies with at least {total} ratings.")
```

```
After merging: 184
E.T. the Extra-Terrestrial (1982): AverageRating:3.67 (TotalRatings:18)
Mad Max: Fury Road (2015): AverageRating:3.47 (TotalRatings:18)
Sunset Boulevard (1950): AverageRating:4.36 (TotalRatings:7)
The Lord of the Rings: The Return of the King (2003): AverageRating:3.82 (TotalRatings:11)
Lawrence of Arabia (1962): AverageRating:3.44 (TotalRatings:18)
Fight Club (1999): AverageRating:3.50 (TotalRatings:7)
Gladiator (2000): AverageRating:3.61 (TotalRatings:18)
The Social Network (2010): AverageRating:3.86 (TotalRatings:7)
Psycho (1960): AverageRating:4.00 (TotalRatings:2)
The Silence of the Lambs (1991): AverageRating:3.14 (TotalRatings:7)
Sunset Boulevard (1950) is the highest rated movie with an average rating of 4.36 among movies with at least 7 ratings.
```

```
# Clear
spark_context.stop()
spark_session.stop()
```

Hình 1.1-1 Kết quả câu 1

1.2 Bài 2: Phân Tích Đánh Giá Theo Thể Loại



The screenshot shows a Jupyter Notebook interface with two cells. The first cell contains a Python script that prints the average rating and total ratings for each genre. The second cell contains a Python script that prints the average rating and total ratings for each genre.

```
# Kết quả
for (genre, (avg, cnt)) in genre_avg.take(10):
    print(f"{genre} - AverageRating:{avg:.2f} (TotalRatings:{cnt})")
```

```
After merging: 184
Sci-Fi - AverageRating:3.73 (TotalRatings:54)
Action - AverageRating:3.71 (TotalRatings:54)
Family - AverageRating:3.67 (TotalRatings:18)
Drama - AverageRating:3.76 (TotalRatings:128)
Biography - AverageRating:3.56 (TotalRatings:25)
Horror - AverageRating:4.00 (TotalRatings:2)
Thriller - AverageRating:3.70 (TotalRatings:27)
Adventure - AverageRating:3.63 (TotalRatings:83)
Film-Noir - AverageRating:4.36 (TotalRatings:7)
Mystery - AverageRating:4.00 (TotalRatings:2)
```

```
# Clear
spark_context.stop()
spark_session.stop()
```

Hình 1.2-1 Kết quả câu 2

1.3 Bài 3: Phân Tích Đánh Giá Theo Giới Tính

```

print(f"{title} - Male_Avg: {formatter(male_avg)}, Female_Avg: {formatter(female_avg)}")

```

After merging: 184

Movie Title	Male Avg	Female Avg
Gladiator (2000)	3.59	3.64
The Terminator (1984)	3.93	4.14
Lawrence of Arabia (1962)	3.55	3.31
Mad Max: Fury Road (2015)	4.00	3.32
No Country for Old Men (2007)	3.92	3.83
Psycho (1960)	NA	4.00
E.T. the Extra-Terrestrial (1982)	3.81	3.55
Fight Club (1999)	3.50	3.50
The Godfather: Part II (1974)	4.06	3.94
The Lord of the Rings: The Fellowship of the Ring (2001)	4.00	3.80
The Silence of the Lambs (1991)	3.33	3.00
Sunset Boulevard (1950)	4.33	4.50
The Lord of the Rings: The Return of the King (2003)	3.75	3.90
The Social Network (2010)	4.00	3.67

```

# Clear
spark_context.stop()

```

DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2
\$ source "/d/UIT_Courses/Bigdata/Lab 2/.venv/Scripts/activate"
(.venv)
DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2
\$

Hình 1.3-1 Kết quả câu 3

1.4 Bài 4 : Phân Tích Đánh Giá Theo Nhóm Tuổi

```

for _, (title, age_dict) in movie_age_group.take(10):
    print(
        f"{title} - "
        f"[0-18: {formatter(age_dict.get('0-18'))}, "
        f"[18-35: {formatter(age_dict.get('18-35'))}, "
        f"[35-50: {formatter(age_dict.get('35-50'))}, "
        f"[50+: {formatter(age_dict.get('50+'))}]"
    )

```

After merging: 184

Movie Title	0-18	18-35	35-50	50+
Gladiator (2000)	NA	3.44	3.81	3.50
The Terminator (1984)	NA	4.17	4.05	3.75
Lawrence of Arabia (1962)	NA	3.60	3.29	4.50
Mad Max: Fury Road (2015)	NA	3.36	3.64	NA
No Country for Old Men (2007)	NA	3.81	3.94	4.00
Psycho (1960)	NA	4.50	3.50	NA
E.T. the Extra-Terrestrial (1982)	NA	3.56	3.83	3.00
Fight Club (1999)	NA	3.50	3.50	3.50
The Godfather: Part II (1974)	NA	3.78	4.25	NA
The Lord of the Rings: The Fellowship of the Ring (2001)	NA	4.00	3.83	NA

```

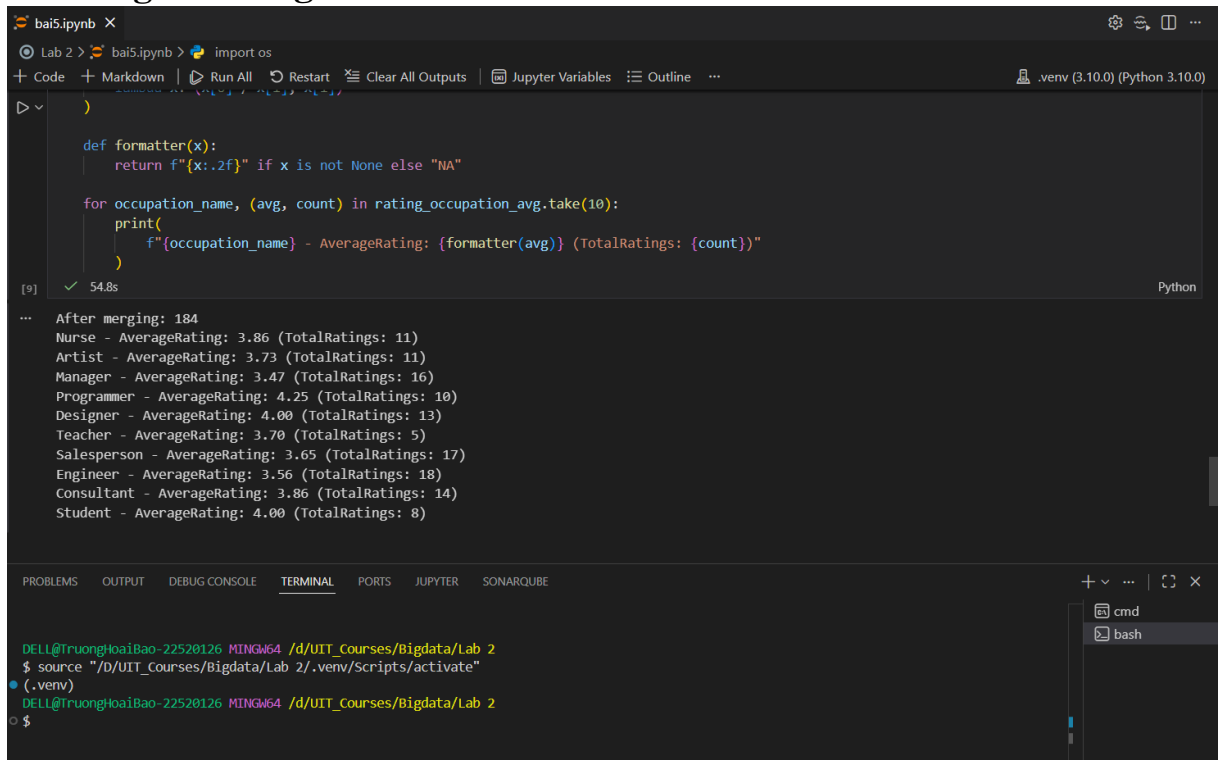
import os

```

DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2
\$ source "/d/UIT_Courses/Bigdata/Lab 2/.venv/Scripts/activate"
(.venv)
DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2
\$

Hình 1.4-1 Kết quả câu 4

1.5 Bài 5: Phân Tích Đánh Giá Theo Occupation (Nghề nghiệp) Của Người Dùng



```
def formatter(x):
    return f"{x:.2f}" if x is not None else "NA"

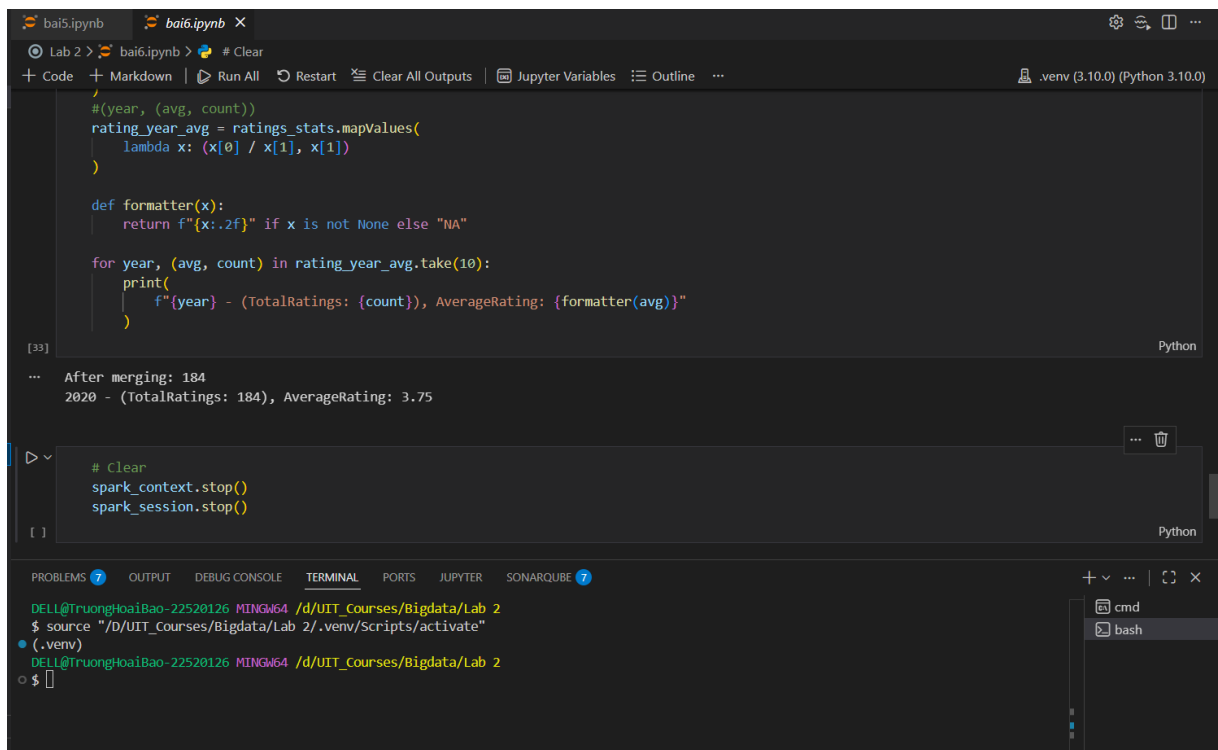
for occupation_name, (avg, count) in rating_occupation_avg.take(10):
    print(
        f"{occupation_name} - AverageRating: {formatter(avg)} (TotalRatings: {count})"
    )
```

After merging: 184

Occupation	AverageRating	TotalRatings
Nurse	3.86	11
Artist	3.73	11
Manager	3.47	16
Programmer	4.25	10
Designer	4.00	13
Teacher	3.70	5
Salesperson	3.65	17
Engineer	3.56	18
Consultant	3.86	14
Student	4.00	8

Hình 1.5-1 Kết quả câu 5

1.6 Bài 6: Phân Tích Đánh Giá Theo Thời Gian



```
#(year, (avg, count))
rating_year_avg = ratings_stats.mapValues(
    lambda x: (x[0] / x[1], x[1])
)

def formatter(x):
    return f"{x:.2f}" if x is not None else "NA"

for year, (avg, count) in rating_year_avg.take(10):
    print(
        f"{year} - (TotalRatings: {count}), AverageRating: {formatter(avg)}"
    )
```

After merging: 184

Year	TotalRatings	AverageRating
2020	184	3.75

Hình 1.6-1 Kết quả câu 6

TÀI LIỆU THAM KHẢO