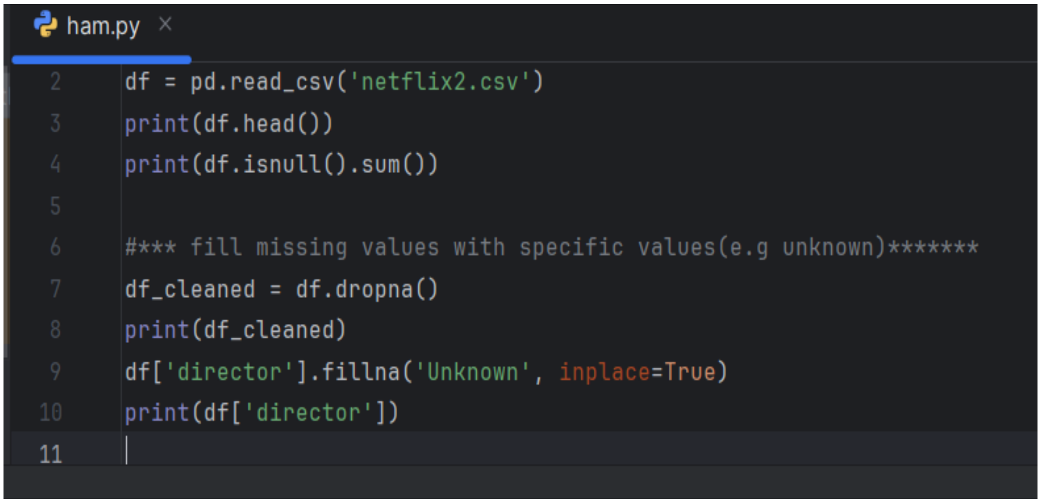
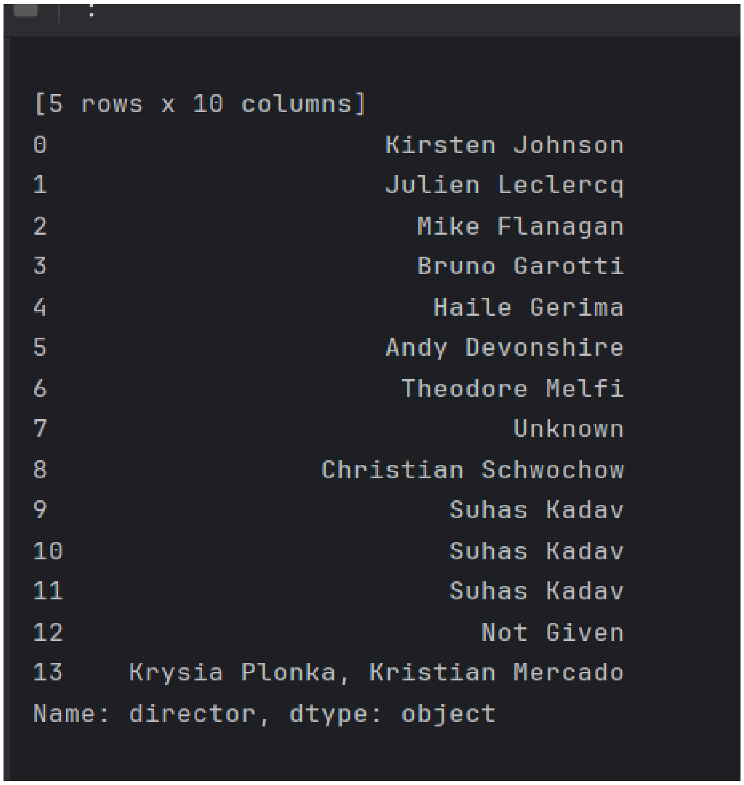
Netflix Tasks

1.



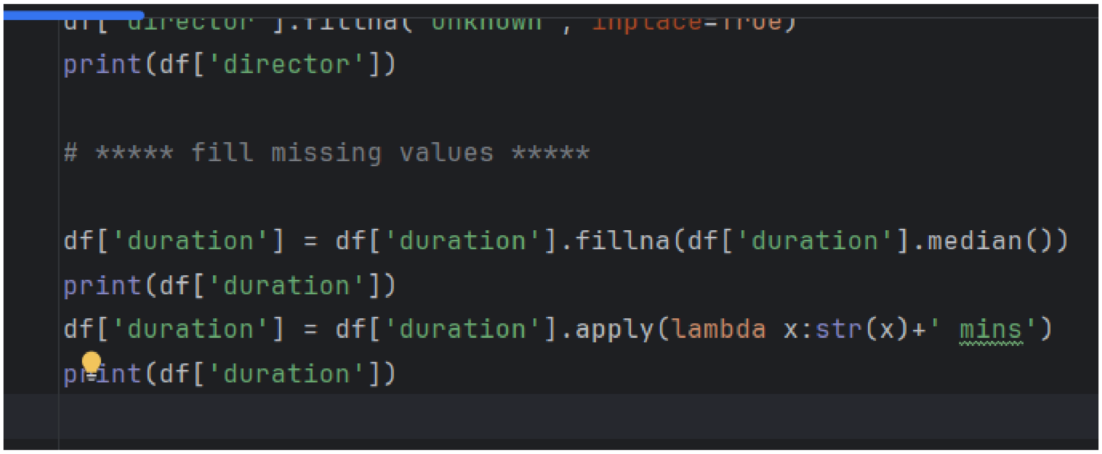
```
df = pd.read_csv('netflix2.csv')
print(df.head())
print(df.isnull().sum())

#*** fill missing values with specific values(e.g unknown)*******
df_cleaned = df.dropna()
print(df_cleaned)
df['director'].fillna('Unknown', inplace=True)
print(df['director'])

```

Output:



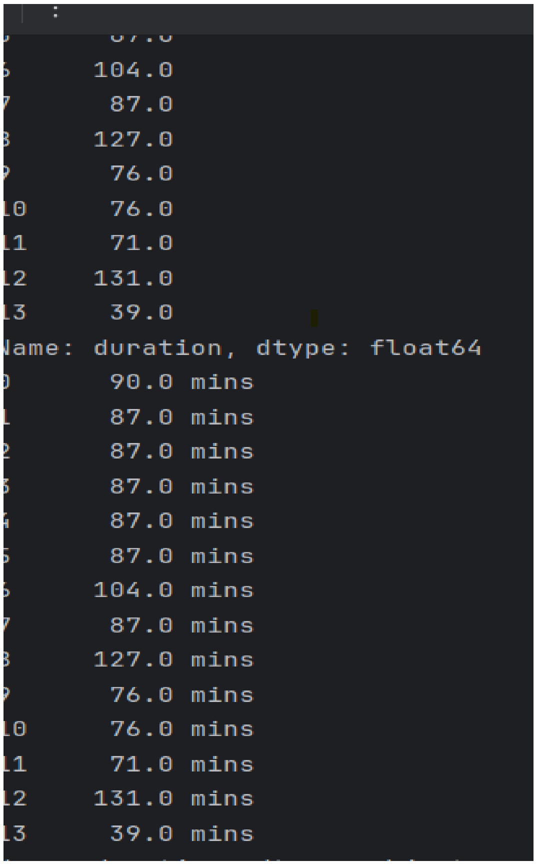
```
[5 rows x 10 columns]
0                      Kirsten Johnson
1                      Julien Leclercq
2                        Mike Flanagan
3                        Bruno Garotti
4                         Haile Gerima
5                      Andy Devonshire
6                       Theodore Melfi
7                              Unknown
8                  Christian Schwochow
9                          Suhas Kadav
10                         Suhas Kadav
11                         Suhas Kadav
12                           Not Given
13    Krysia Plonka, Kristian Mercado
Name: director, dtype: object
```

2.



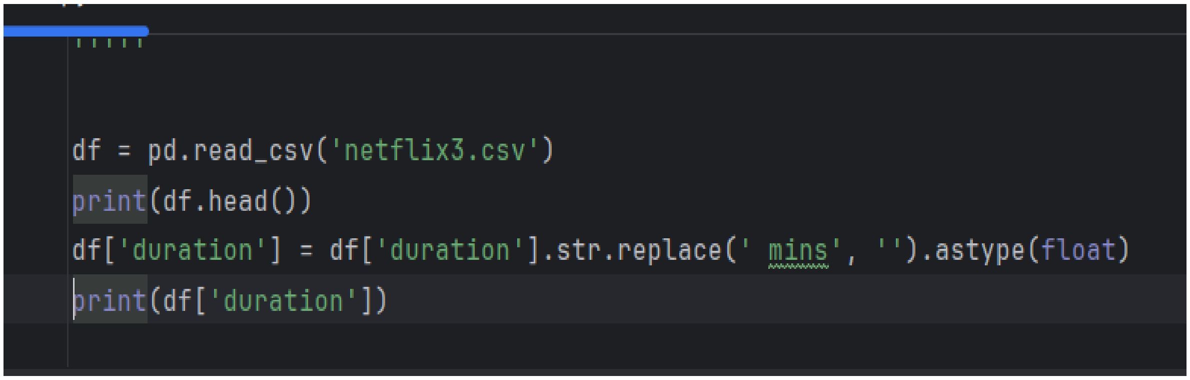
```
df['director'].fillna('unknown', inplace=True)
print(df['director'])

# ***** fill missing values *****

df['duration'] = df['duration'].fillna(df['duration'].median())
print(df['duration'])
df['duration'] = df['duration'].apply(lambda x:str(x)+' mins')
print(df['duration'])
```

Output:



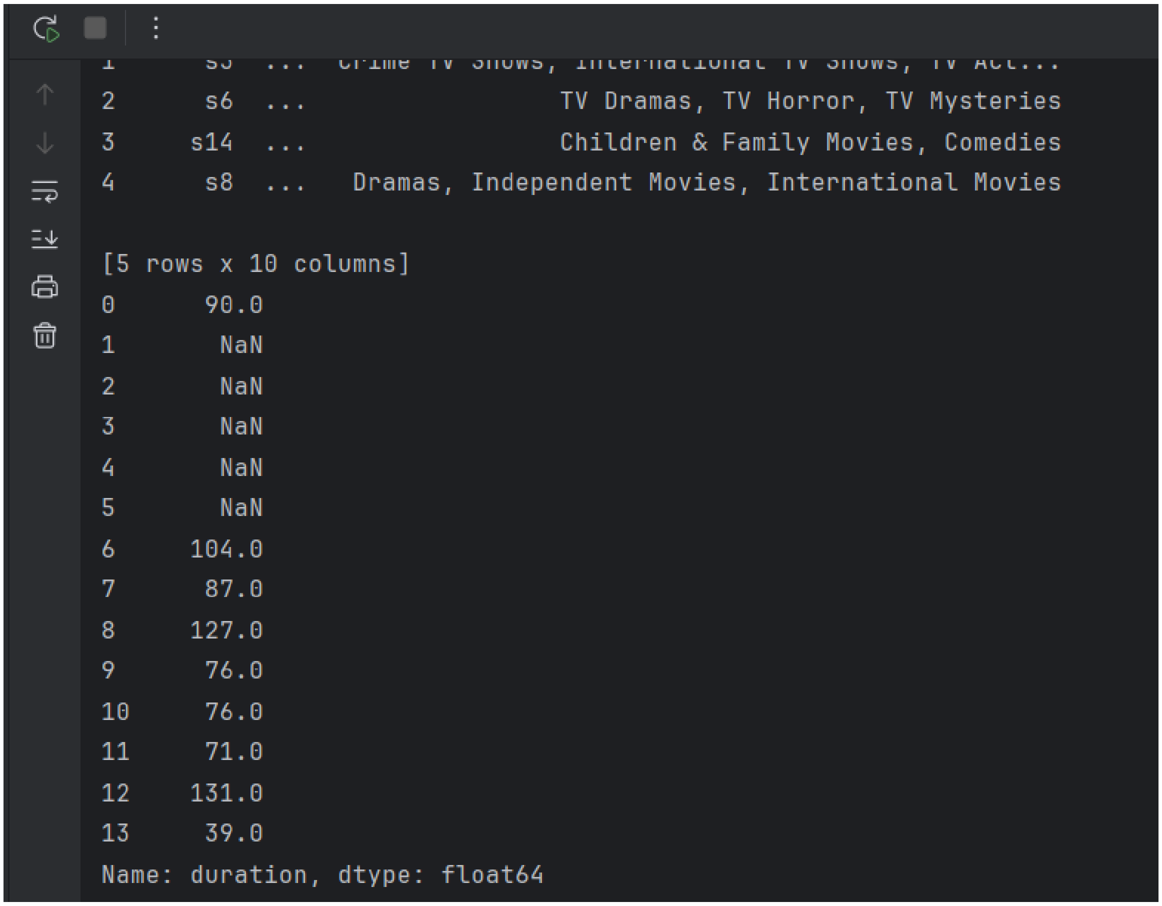
```
5      87.0
6     104.0
7      87.0
8     127.0
9      76.0
10     76.0
11     71.0
12    131.0
13     39.0
Name: duration, dtype: float64
0      90.0 mins
1      87.0 mins
2      87.0 mins
3      87.0 mins
4      87.0 mins
5      87.0 mins
6     104.0 mins
7      87.0 mins
8     127.0 mins
9      76.0 mins
10     76.0 mins
11     71.0 mins
12    131.0 mins
13     39.0 mins
```

3.



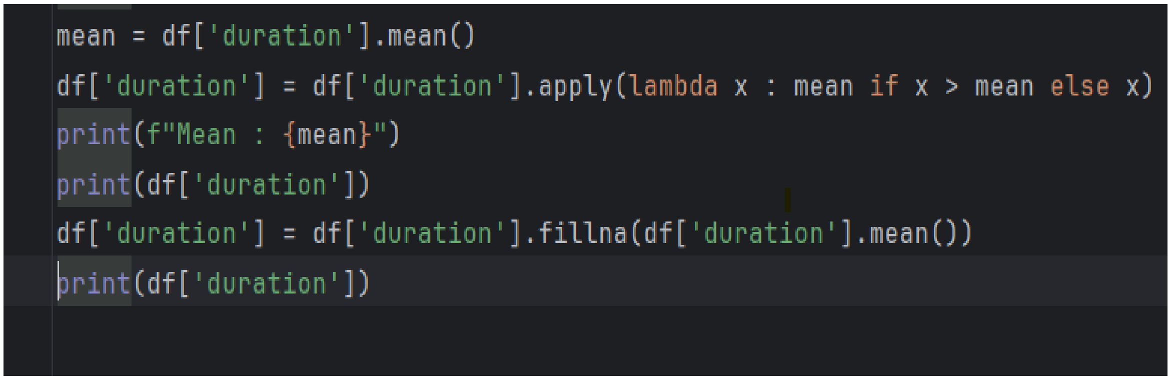
```
'''


df = pd.read_csv('netflix3.csv')
print(df.head())
df['duration'] = df['duration'].str.replace(' mins', '').astype(float)
print(df['duration'])
```

Output:



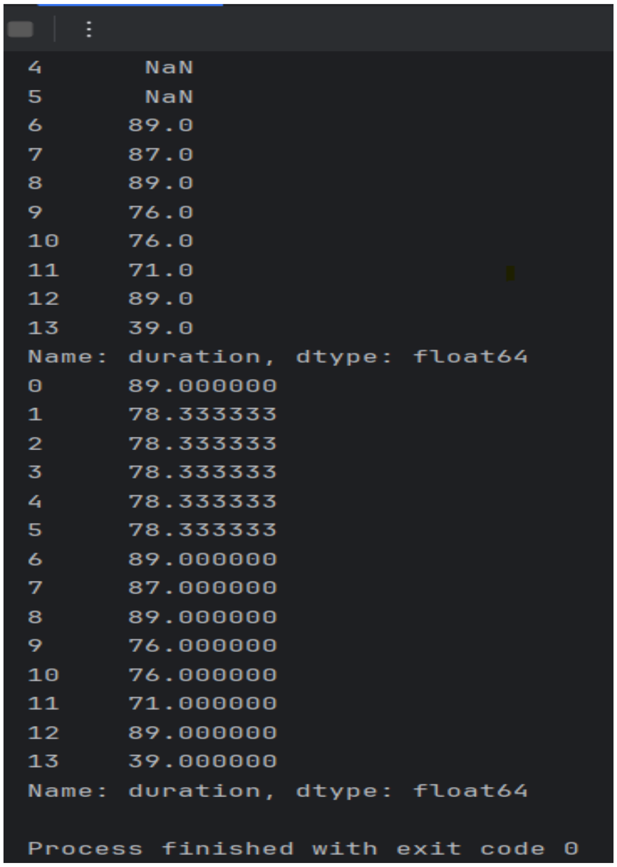
```
1      s3  ...  Crime TV Shows, International TV Shows, TV Act...
2      s6  ...                 TV Dramas, TV Horror, TV Mysteries
3     s14  ...                 Children & Family Movies, Comedies
4      s8  ...    Dramas, Independent Movies, International Movies

[5 rows x 10 columns]
0      90.0
1       NaN
2       NaN
3       NaN
4       NaN
5       NaN
6     104.0
7      87.0
8     127.0
9      76.0
10     76.0
11     71.0
12    131.0
13     39.0
Name: duration, dtype: float64
```

4.



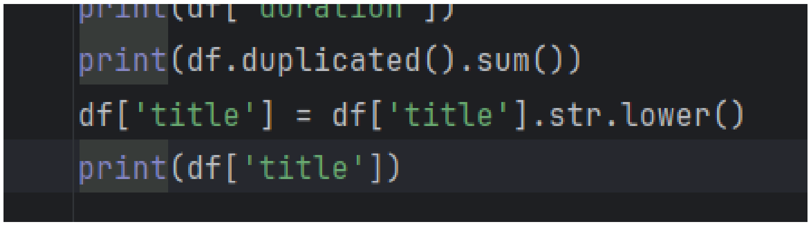
```
mean = df['duration'].mean()
df['duration'] = df['duration'].apply(lambda x : mean if x > mean else x)
print(f"Mean : {mean}")
print(df['duration'])
df['duration'] = df['duration'].fillna(df['duration'].mean())
print(df['duration'])
```

Output:



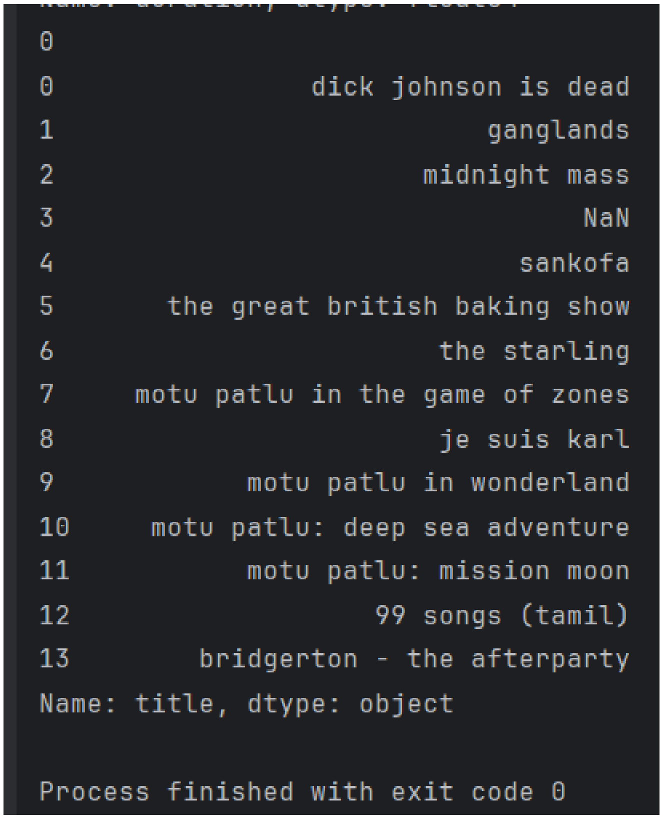
```
4       NaN
5       NaN
6      89.0
7      87.0
8      89.0
9      76.0
10     76.0
11     71.0
12     89.0
13     39.0
Name: duration, dtype: float64
0      89.000000
1      78.333333
2      78.333333
3      78.333333
4      78.333333
5      78.333333
6      89.000000
7      87.000000
8      89.000000
9      76.000000
10     76.000000
11     71.000000
12     89.000000
13     39.000000
Name: duration, dtype: float64

Process finished with exit code 0
```

5.



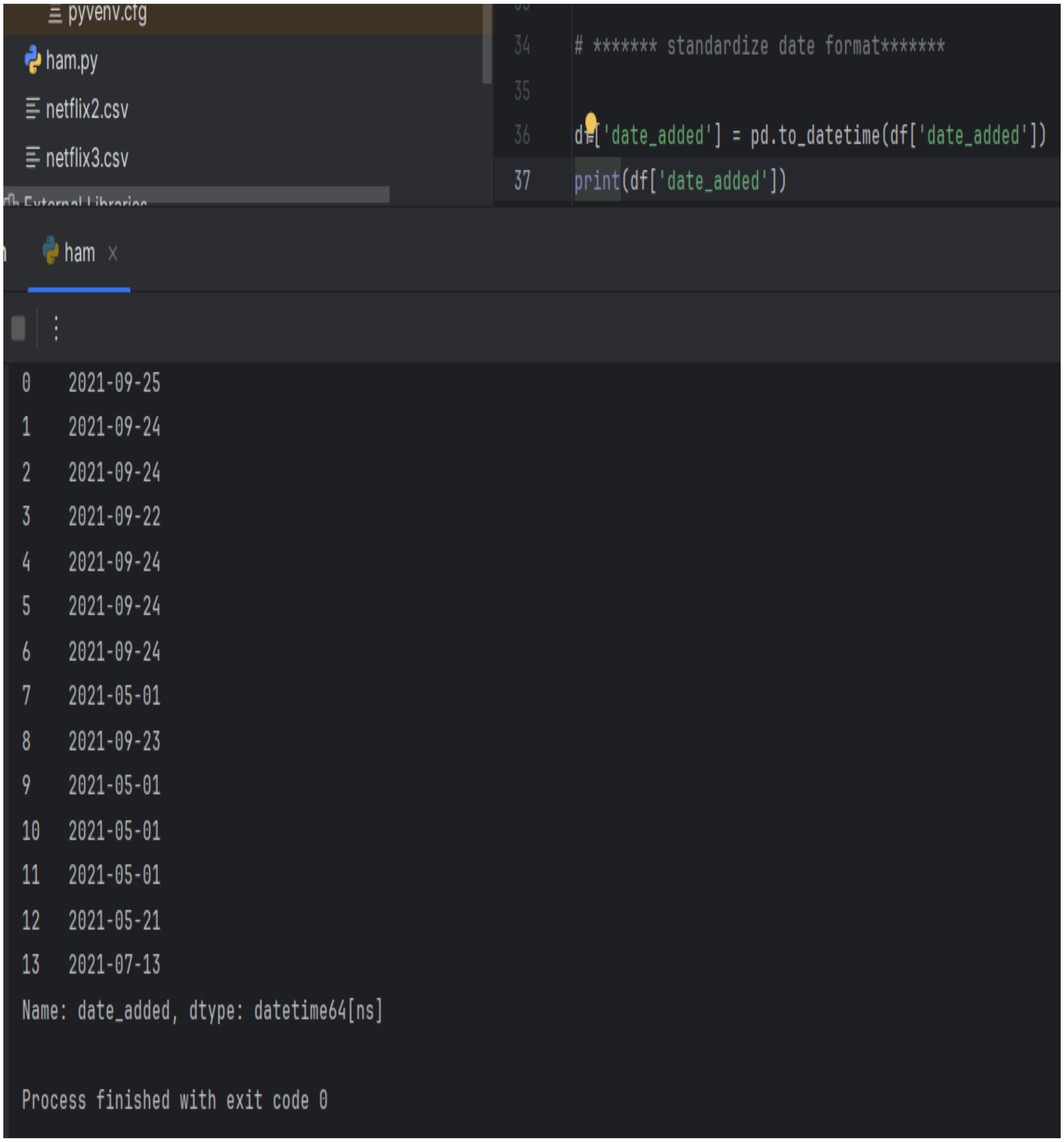
```
print(df['duration'])
print(df.duplicated().sum())
df['title'] = df['title'].str.lower()
print(df['title'])
```

Output:



```
Name: duration, dtype: float64
0
0                  dick johnson is dead
1                             ganglands
2                         midnight mass
3                                   NaN
4                               sankofa
5          the great british baking show
6                          the starling
7        motu patlu in the game of zones
8                          je suis karl
9               motu patlu in wonderland
10        motu patlu: deep sea adventure
11              motu patlu: mission moon
12                      99 songs (tamil)
13           bridgerton - the afterparty
Name: title, dtype: object

Process finished with exit code 0
```

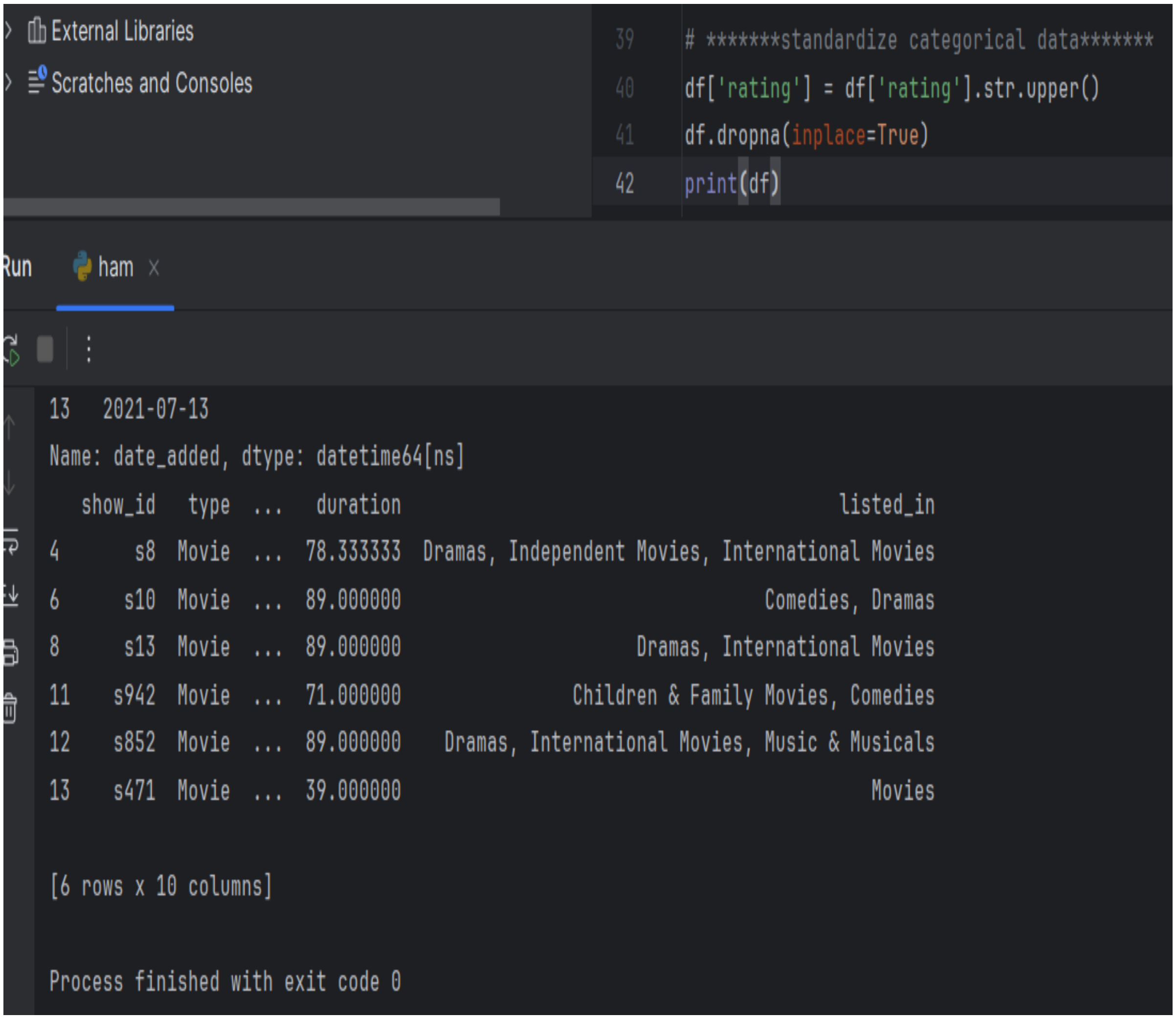
6.



```
# ******* standardize date format*******

df['date_added'] = pd.to_datetime(df['date_added'])
print(df['date_added'])
```

```
0     2021-09-25
1     2021-09-24
2     2021-09-24
3     2021-09-22
4     2021-09-24
5     2021-09-24
6     2021-09-24
7     2021-05-01
8     2021-09-23
9     2021-05-01
10    2021-05-01
11    2021-05-01
12    2021-05-21
13    2021-07-13
Name: date_added, dtype: datetime64[ns]

Process finished with exit code 0
```

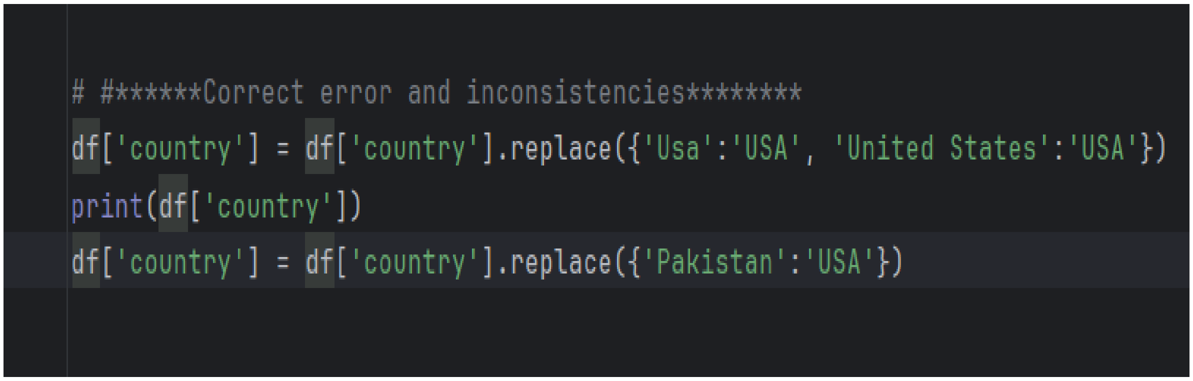
7.



```
# *******standardize categorical data*******
df['rating'] = df['rating'].str.upper()
df.dropna(inplace=True)
print(df)
```

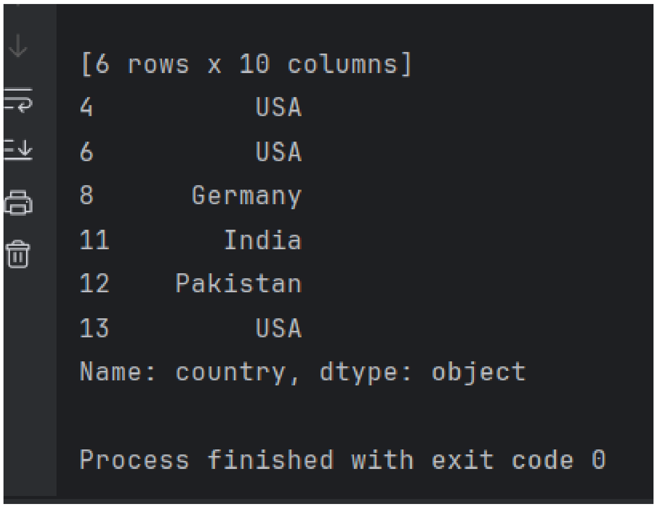
```
13   2021-07-13
Name: date_added, dtype: datetime64[ns]
   show_id   type  ...   duration                                     listed_in
4       s8  Movie  ...  78.333333  Dramas, Independent Movies, International Movies
6      s10  Movie  ...  89.000000                                Comedies, Dramas
8      s13  Movie  ...  89.000000                    Dramas, International Movies
11    s942  Movie  ...  71.000000              Children & Family Movies, Comedies
12    s852  Movie  ...  89.000000    Dramas, International Movies, Music & Musicals
13    s471  Movie  ...  39.000000                                          Movies

[6 rows x 10 columns]

Process finished with exit code 0
```

8.



```
# #******Correct error and inconsistencies********
df['country'] = df['country'].replace({'Usa':'USA', 'United States':'USA'})
print(df['country'])
df['country'] = df['country'].replace({'Pakistan':'USA'})
```

Output:



```
[6 rows x 10 columns]
4          USA
6          USA
8      Germany
11       India
12    Pakistan
13         USA
Name: country, dtype: object

Process finished with exit code 0
```