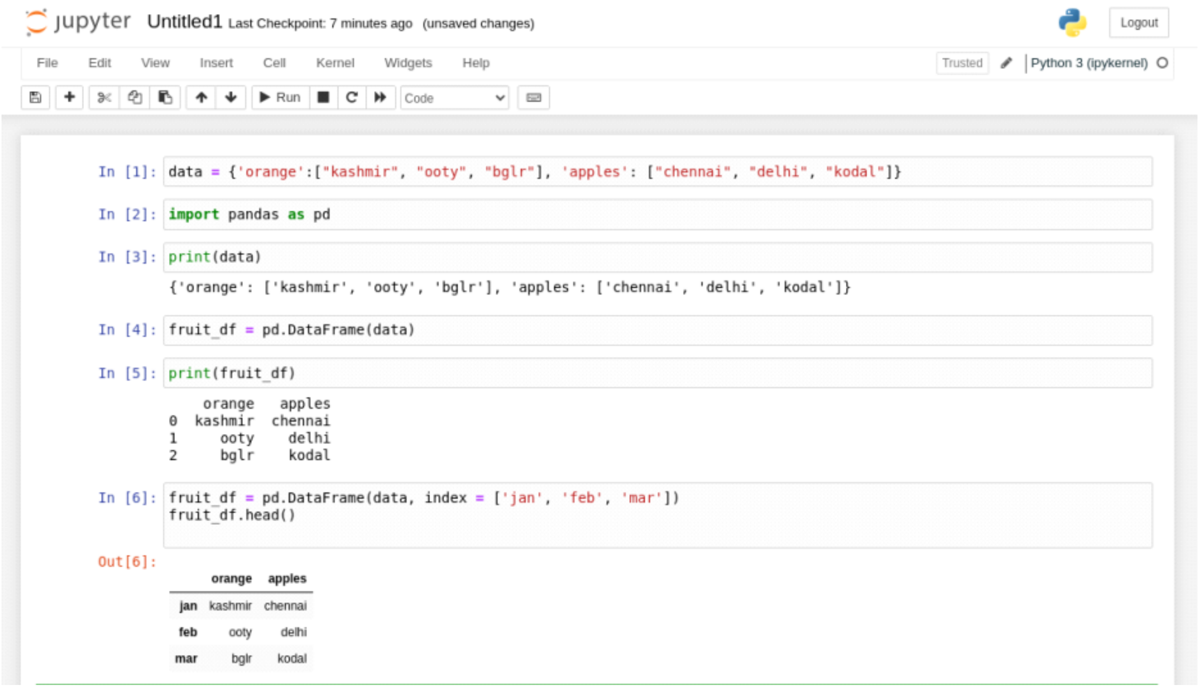
Pandas

Tuesday, September 19, 2023 3:34 PM



```
In [1]: data = {'orange':["kashmir", "ooty", "bglr"], 'apples': ["chennai", "delhi", "kodal"]}
```

```
In [2]: import pandas as pd
```

```
In [3]: print(data)
```

```
{'orange': ['kashmir', 'ooty', 'bglr'], 'apples': ['chennai', 'delhi', 'kodal']}
```

```
In [4]: fruit_df = pd.DataFrame(data)
```

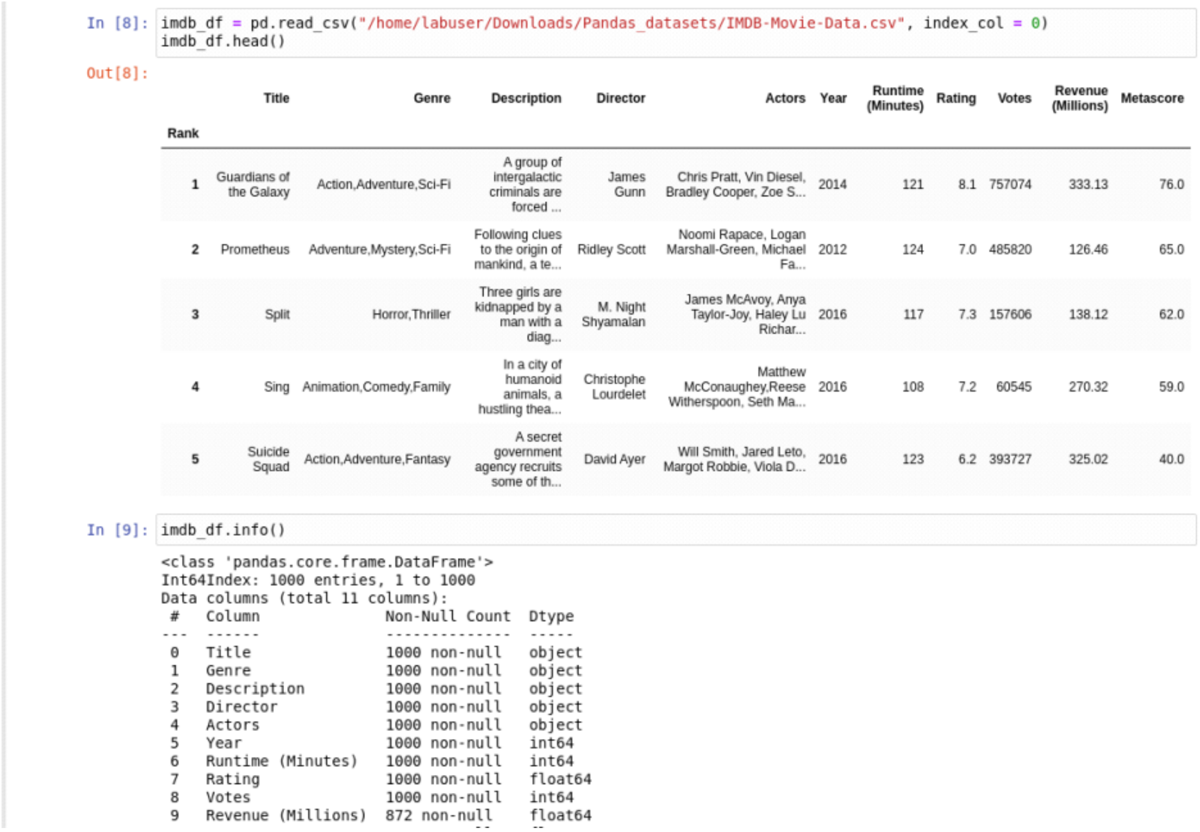
```
In [5]: print(fruit_df)
```

```
    orange   apples
0  kashmir  chennai
1     ooty    delhi
2     bglr    kodal
```

```
In [6]: fruit_df = pd.DataFrame(data, index = ['jan', 'feb', 'mar'])
        fruit_df.head()
```

Out[6]:

	orange	apples
jan	kashmir	chennai
feb	ooty	delhi
mar	bglr	kodal



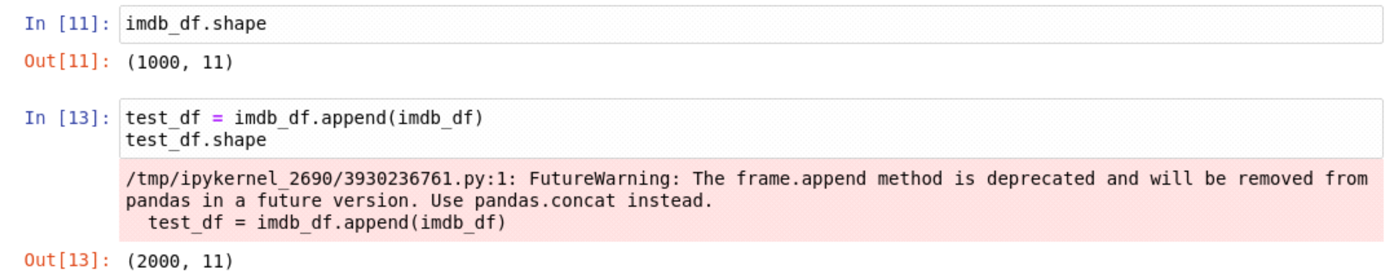
```
In [8]: imdb_df = pd.read_csv("/home/labuser/Downloads/Pandas_datasets/IMDB-Movie-Data.csv", index_col = 0)
        imdb_df.head()
```

Out[8]:

	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
Rank											
1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13	76.0
2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46	65.0
3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12	62.0
4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32	59.0
5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02	40.0

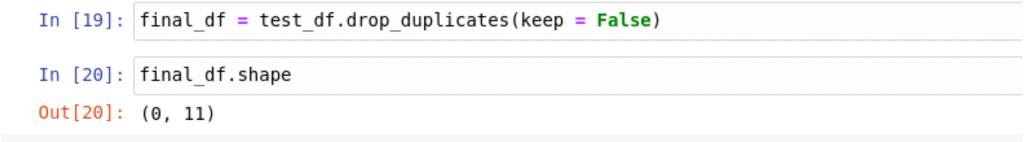
```
In [9]: imdb_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1000 entries, 1 to 1000
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Title               1000 non-null   object
 1   Genre               1000 non-null   object
 2   Description         1000 non-null   object
 3   Director            1000 non-null   object
 4   Actors              1000 non-null   object
 5   Year                1000 non-null   int64
 6   Runtime (Minutes)   1000 non-null   int64
 7   Rating              1000 non-null   float64
 8   Votes               1000 non-null   int64
 9   Revenue (Millions)  872 non-null    float64
```

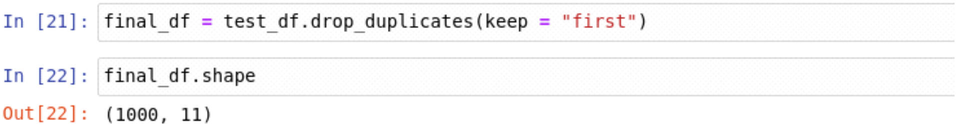


```
In [11]: imdb_df.shape

Out[11]: (1000, 11)

In [13]: test_df = imdb_df.append(imdb_df)
         test_df.shape

/tmp/ipykernel_2690/3930236761.py:1: FutureWarning: The frame.append method is deprecated and will be removed from
pandas in a future version. Use pandas.concat instead.
  test_df = imdb_df.append(imdb_df)

Out[13]: (2000, 11)
```

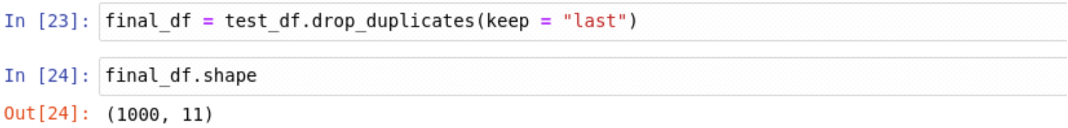
Imdb_df rows are appended to imdb_df, therefore 2000 rows.
Now, we will do duplicate handling.
Using drop_duplicates, it has three types, keep = first, keep = last, keep = false, default is keep = first.
Keep = false deletes both copies, first keeps the first copy only, keep = last keeps the last duplicate only.



```
In [19]: final_df = test_df.drop_duplicates(keep = False)

In [20]: final_df.shape

Out[20]: (0, 11)
```

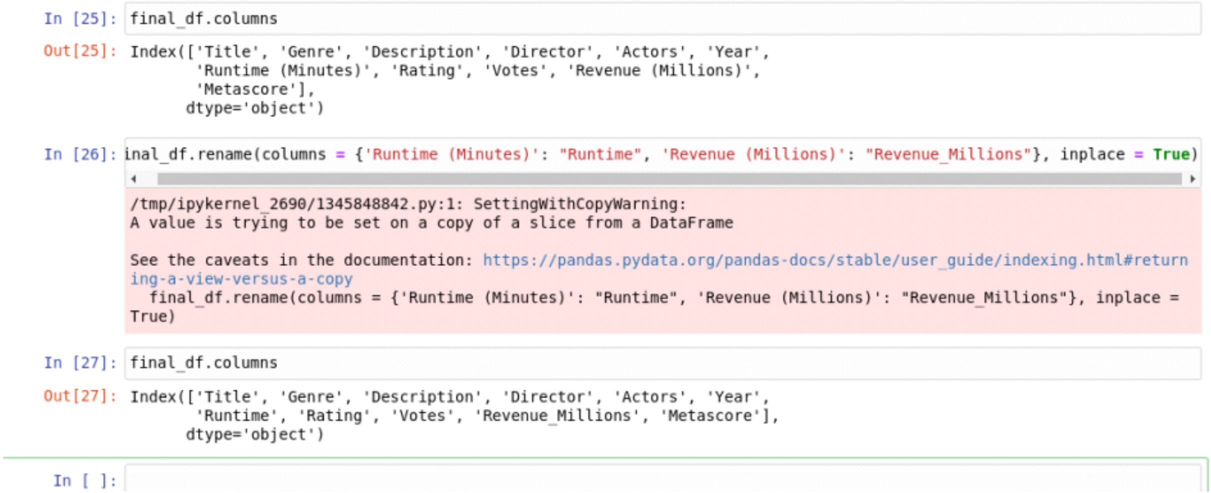


```
In [21]: final_df = test_df.drop_duplicates(keep = "first")

In [22]: final_df.shape

Out[22]: (1000, 11)
```

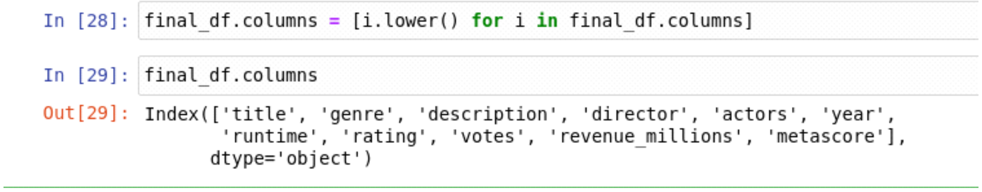


```
In [23]: final_df = test_df.drop_duplicates(keep = "last")

In [24]: final_df.shape

Out[24]: (1000, 11)
```

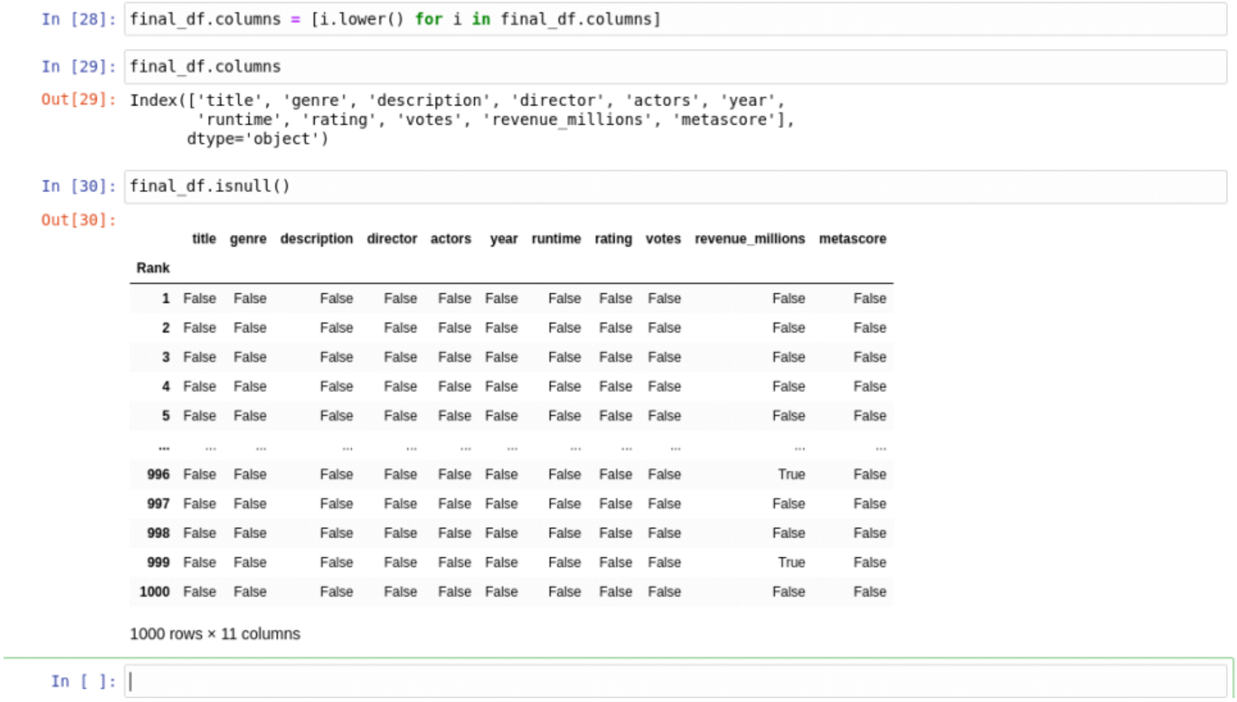
Renaming columns:



```
In [25]: final_df.columns

Out[25]: Index(['Title', 'Genre', 'Description', 'Director', 'Actors', 'Year',
                'Runtime (Minutes)', 'Rating', 'Votes', 'Revenue (Millions)',
                'Metascore'],
               dtype='object')

In [26]: inal_df.rename(columns = {'Runtime (Minutes)': "Runtime", 'Revenue (Millions)': "Revenue_Millions"}, inplace = True)

/tmp/ipykernel_2690/1345848842.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#return
ing-a-view-versus-a-copy
  final_df.rename(columns = {'Runtime (Minutes)': "Runtime", 'Revenue (Millions)': "Revenue_Millions"}, inplace =
True)

In [27]: final_df.columns

Out[27]: Index(['Title', 'Genre', 'Description', 'Director', 'Actors', 'Year',
                'Runtime', 'Rating', 'Votes', 'Revenue_Millions', 'Metascore'],
               dtype='object')

In [ ]:
```



```
In [28]: final_df.columns = [i.lower() for i in final_df.columns]

In [29]: final_df.columns

Out[29]: Index(['title', 'genre', 'description', 'director', 'actors', 'year',
                'runtime', 'rating', 'votes', 'revenue_millions', 'metascore'],
               dtype='object')
```



```
In [28]: final_df.columns = [i.lower() for i in final_df.columns]
```

```
In [29]: final_df.columns
```

```
Out[29]: Index(['title', 'genre', 'description', 'director', 'actors', 'year',
                'runtime', 'rating', 'votes', 'revenue_millions', 'metascore'],
               dtype='object')
```

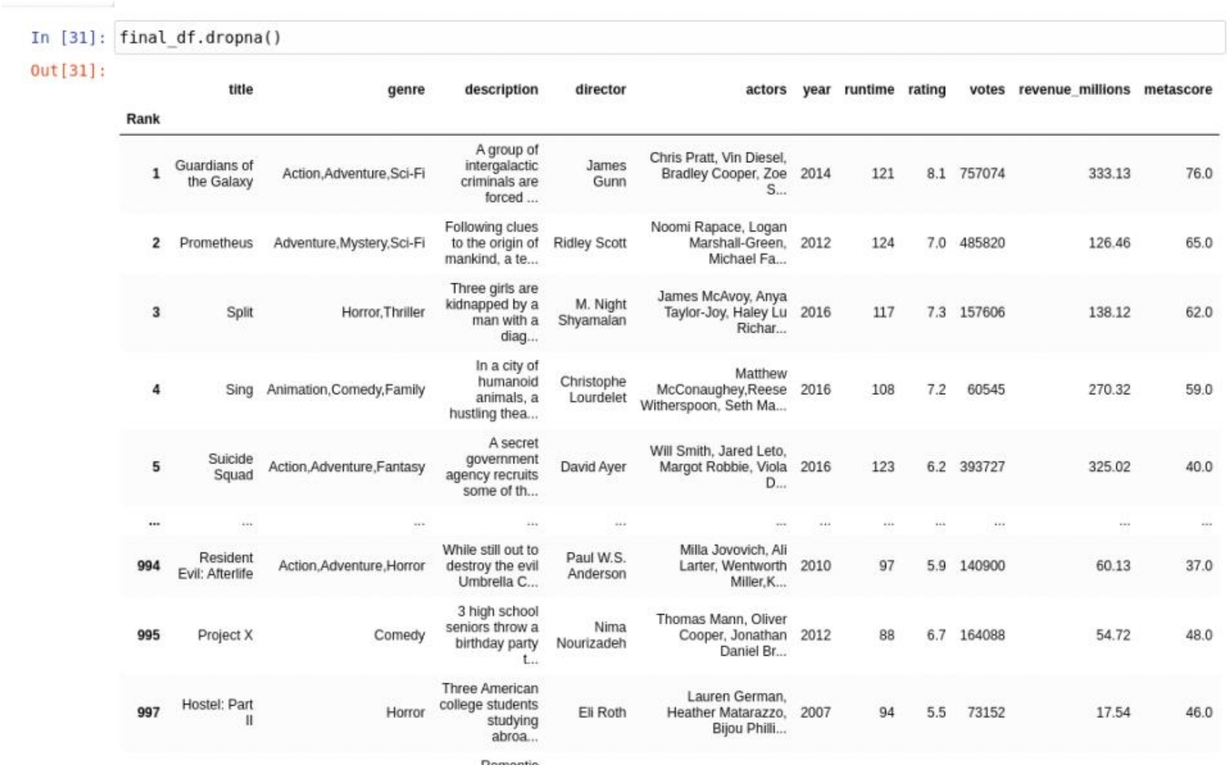
```
In [30]: final_df.isnull()
```

Out[30]:

	title	genre	description	director	actors	year	runtime	rating	votes	revenue_millions	metascore
Rank											
1	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False
5	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...
996	False	False	False	False	False	False	False	False	False	True	False
997	False	False	False	False	False	False	False	False	False	False	False
998	False	False	False	False	False	False	False	False	False	False	False
999	False	False	False	False	False	False	False	False	False	True	False
1000	False	False	False	False	False	False	False	False	False	False	False

1000 rows × 11 columns

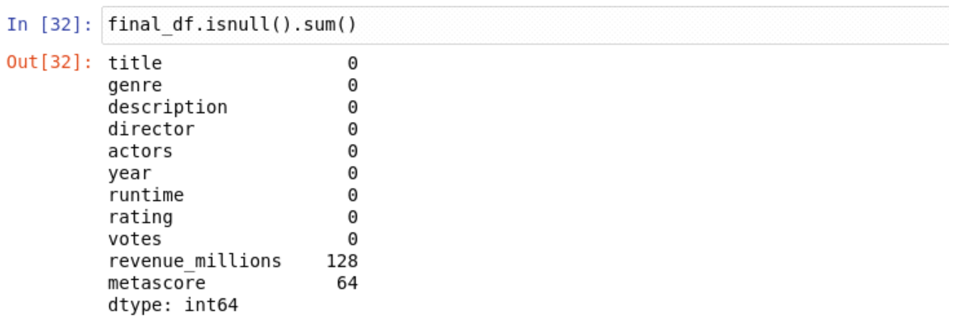
```
In [ ]: |
```



```
In [31]: final_df.dropna()
```

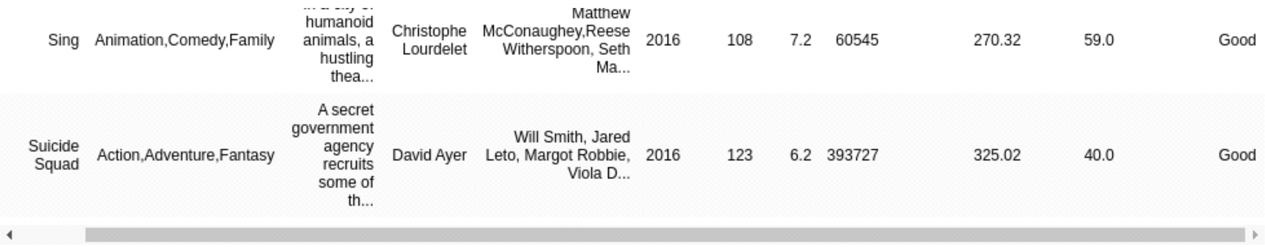
Out[31]:

	title	genre	description	director	actors	year	runtime	rating	votes	revenue_millions	metascore
Rank											
1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13	76.0
2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46	65.0
3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12	62.0
4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32	59.0
5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02	40.0
...	...	...	...	...	...	...	...	...	...	...	...
994	Resident Evil: Afterlife	Action,Adventure,Horror	While still out to destroy the evil Umbrella C...	Paul W.S. Anderson	Milla Jovovich, Ali Larter, Wentworth Miller,K...	2010	97	5.9	140900	60.13	37.0
995	Project X	Comedy	3 high school seniors throw a birthday party t...	Nima Nourizadeh	Thomas Mann, Oliver Cooper, Jonathan Daniel Br...	2012	88	6.7	164088	54.72	48.0
997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007	94	5.5	73152	17.54	46.0
			Romantic								

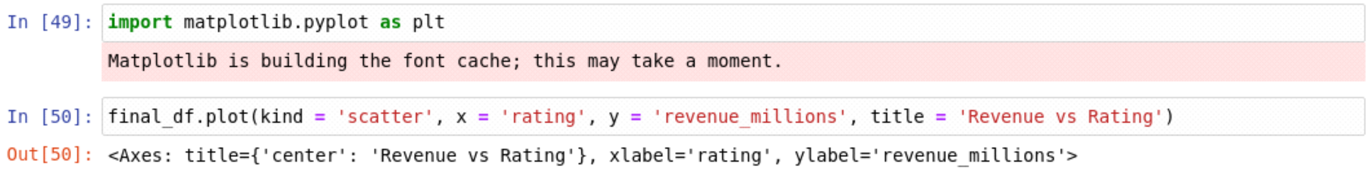


```
In [32]: final_df.isnull().sum()

Out[32]: title                 0
         genre                 0
         description           0
         director              0
         actors                0
         year                  0
         runtime               0
         rating                0
         votes                 0
         revenue_millions    128
         metascore            64
         dtype: int64
```

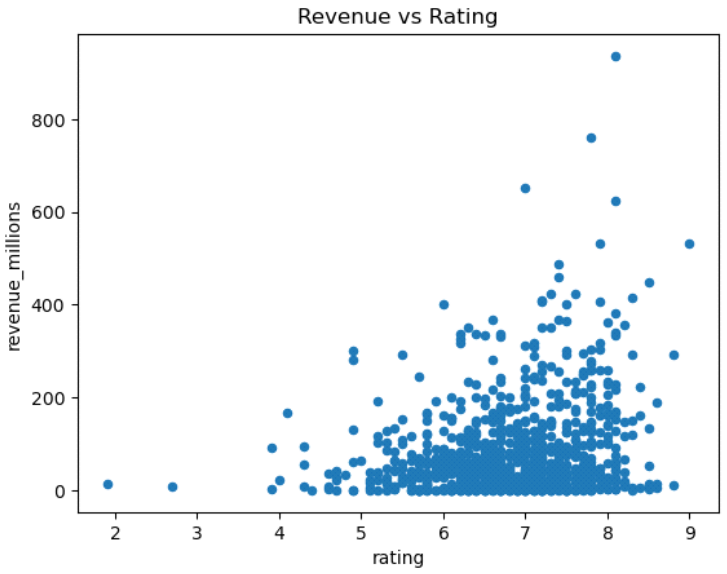
Note: Don't use dropna normally as the whole row gets deleted which contains a null value
Dropna with axis = 1 then it will drop columns instead of rows. Eg. Final_df.dropna(axis = 1)



Sing	Animation,Comedy,Family	humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32	59.0	Good
Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02	40.0	Good



```
In [49]: import matplotlib.pyplot as plt

         Matplotlib is building the font cache; this may take a moment.

In [50]: final_df.plot(kind = 'scatter', x = 'rating', y = 'revenue_millions', title = 'Revenue vs Rating')

Out[50]: <Axes: title={'center': 'Revenue vs Rating'}, xlabel='rating', ylabel='revenue_millions'>
```


Revenue vs Rating
revenue_millions
rating