Entrance Test for Machine Learning, Ai and advanced python.

'Classroom Student' | aniket mukherjee | 8910265146

Group A : Basic Data Representation about the data set

1. Import the modules

In [1]:

```
import pandas as pd
```

2. Read the data set

In [2]:

```
a=pd.read_csv('netflix.csv')
```

3. Print the data set

In [3]:

```
a
```

2	81213894	The Zoya Factor	Abhishek Sharma	Sonam Kapoor, Dulquer Salmaan, Sanjay Kapoor, ...	India	30-Nov-19	2019	TV-14	135 min
3	81082007	Atlantics	Mati Diop	Mama Sane, Amadou Mbow, Ibrahima Traore, Nicol...	France, Senegal, Belgium	29-Nov-19	2019	TV-14	106 min
				Abigail Oliver,	Canada				

4. Show the Index of the data set

In [4]:

```
a.index
```

Out[4]:

```
RangeIndex(start=0, stop=5837, step=1)
```

5. Show the size of the dataset

In [5]:

```
a.size
```

Out[5]:

```
70044
```

6. Show all the row no and the column no.

In [6]:

```
a.shape
```

Out[6]:

```
(5837, 12)
```

7. Show all the columns name.

In [7]:

```
a.columns
```

Out[7]:

```
Index(['show_id', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description',
       'type'],
      dtype='object')
```

In [8]:

```
for i in range (len(a.columns)):
    print(a.columns[i])
```

```
show_id
title
director
cast
country
date_added
release_year
rating
duration
listed_in
description
type
```

8. Show the total no of columns.

In [9]:

```
len(a.columns)
```

Out[9]:

```
12
```

9. Show the memory consumptions for all the columns

In [10]:

```
a.memory_usage()
```

Out[10]:

```
Index            128
show_id        46696
title          46696
director       46696
cast           46696
country        46696
date_added     46696
release_year   46696
rating         46696
duration       46696
listed_in      46696
description    46696
type           46696
dtype: int64
```

10. Show the dimension of the data set.

In [11]:

```
a.ndim
```

Out[11]:

```
2
```

11. Show the column names, (null/not-null), data types all the information at a glance.

In [12]:

```
a.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5837 entries, 0 to 5836
Data columns (total 12 columns):
show_id         5837 non-null int64
title           5837 non-null object
director        3936 non-null object
cast            5281 non-null object
country         5410 non-null object
date_added      5195 non-null object
release_year    5837 non-null int64
rating          5827 non-null object
duration        5837 non-null object
listed_in       5837 non-null object
description     5837 non-null object
type            5837 non-null object
dtypes: int64(2), object(10)
memory usage: 547.3+ KB
```

12. Show the first 8 rows of the dataset.

In [13]:

```
a.head(8)
```

Out[13]:

	show_id	title	director	cast	country	date_added	release_year	rating	dura
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	30-Nov-19	2019	TV-14	Sea
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	30-Nov-19	2019	TV-G	67
2	81213894	The Zoya Factor	Abhishek Sharma	Sonam Kapoor, Dulquer Salmaan, Sanjay Kapoor, ...	India	30-Nov-19	2019	TV-14	135
3	81082007	Atlantics	Mati Diop	Mama Sane, Amadou Mbow, Ibrahima Traore, Nicol...	France, Senegal, Belgium	29-Nov-19	2019	TV-14	106
4	80213643	Chip and Potato	NaN	Abigail Oliver, Andrea Libman, Briana Buckmast...	Canada, United Kingdom	NaN	2019	TV-Y	Seas
5	81172754	Crazy people	Moses Inwang	Ramsey Nouah, Chigul, Sola Sobowale, Ireti Doy...	Nigeria	29-Nov-19	2018	TV-14	107
6	81120982	I Lost My Body	Jérémy Clapin	Hakim Faris, Victoire Du Bois, Patrick d'Assum...	France	29-Nov-19	2019	TV-MA	81
7	81227195	Kalushi: The Story of Solomon Mahlangu	Mandla Dube	Thabo Rametsi, Thabo Malema, Welile Nzuza, Jaf...	South Africa	29-Nov-19	2016	TV-MA	107

13. Show the last 8 rows of the dataset.

In [21]:

```
a.tail(8)
```

Out[21]:

	show_id	title	director	cast	country	date_added	release_year	rating	d
5829	70206826	Victim of Beauty	Roger Young	William Devane, Jeri Ryan, Michele Abrams, Nic...	United States	01-Oct-11	1991	NR	
5830	60003155	Joseph: King of Dreams	Rob LaDuca, Robert C. Ramirez	Ben Affleck, Mark Hamill, Richard Herd, Mauree...	United States	27-Sep-11	2000	TV-PG	
5831	70154110	Even the Rain	Icíar Bollaín	Luis Tosar, Gael García Bernal, Juan Carlos Ad...	Spain, Mexico, France	17-May-11	2010	NR	
5832	70141644	Mad Ron's Prevues from Hell	Jim Monaco	Nick Pawlow, Jordu Schell, Jay Kushwara, Micha...	United States	01-Nov-10	1987	NR	
5833	70127998	Splatter	Joe Dante	Corey Feldman, Tony Todd, Tara Leigh, Erin Way...	United States	18-Nov-09	2009	TV-14	
5834	70084180	Just Another Love Story	Ole Bornedal	Anders W. Berthelsen, Rebecka Hemse, Nikolaj L...	Denmark	05-May-09	2007	NR	
5835	70157452	Dinner for Five	NaN	NaN	United States	04-Feb-08	2007	TV-MA	
5836	70053412	To and From New York	Sorin Dan Mihalcescu	Barbara King, Shaana Diya, John Krisiukenas, Y...	United States	01-Jan-08	2006	NR	

14. Show the rows between 78 to 87 rows. As index starts from 0 here, so you have to show 77 to 86.

In [14]:

```
a.head(87).tail(10)
```

Out[14]:

	show_id	title	director	cast	country	date_added	release_year	rating
77	80028357	Love, Rosie	Christian Ditter	Lily Collins, Sam Claflin, Christian Cooke, Ja...	Germany, United Kingdom	20-Nov-19	2014	R
78	81217739	Mallesham	Raj R	Jhansi, Priyadarshi Pullikonda, Ananya Nagalla	India	20-Nov-19	2019	TV-PG
79	60031884	Once Upon a Time in the West	Sergio Leone	Henry Fonda, Charles Bronson, Claudia Cardinal...	Italy, United States	20-Nov-19	1968	PG-13
80	70117289	She's Out of My League	Jim Field Smith	Jay Baruchel, Alice Eve, T.J. Miller, Mike Vog...	United States	20-Nov-19	2010	R
81	28631236	Superstar	Bruce McCulloch	Molly Shannon, Will Ferrell, Elaine Hendrix, H...	United States	20-Nov-19	1999	PG-13
82	70121502	The Adventures of Tintin	Steven Spielberg	Jamie Bell, Andy Serkis, Daniel Craig, Nick Fr...	United States, New Zealand, United Kingdom	20-Nov-19	2011	PG
83	70215455	The Devil Inside	William Brent Bell	Fernanda Andrade, Simon Quarterman, Evan Helmu...	United States	20-Nov-19	2012	R
84	506464	The First Wives Club	Hugh WIlson	Bette Midler, Goldie Hawn, Diane Keaton, Maggi...	United States	20-Nov-19	1996	PG
85	60003508	The Gift	Sam Raimi	Cate Blanchett, Giovanni Ribisi, Keanu Reeves,...	United States	20-Nov-19	2000	R

	show_id	title	director	cast	country	date_added	release_year	rating
86	70105132	The Goods: Live Hard, Sell Hard	Neal Brennan	Jeremy Piven, Ving Rhames, James Brolin, David...	United States	20-Nov-19	2009	R

15.Show all the unique release_year from the data set.

In [15]:

```
import numpy as np
np.unique(a['release_year'])
```

Out[15]:

```
array([1925, 1942, 1943, 1944, 1945, 1946, 1947, 1954, 1955, 1956, 1958,
       1959, 1960, 1962, 1963, 1965, 1966, 1967, 1968, 1969, 1970, 1971,
       1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982,
       1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993,
       1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004,
       2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015,
       2016, 2017, 2018, 2019, 2020], dtype=int64)
```

16. Show the first releasing year of the show.

In [16]:

```
np.unique(a['release_year'])[0]
```

Out[16]:

```
1925
```

17. Show the last releasing year of the show.

In [17]:

```
np.unique(a['release_year'])[-1]
```

Out[17]:

```
2020
```

18. Show the details of the 1st row where release_year is 1988.

In [18]:

```
a.loc[1988,:]
```

Out[18]:

```
show_id                                              80145141
title                                        Black Earth Rising
director                                                    NaN
cast          Michaela Coel, John Goodman, Abena Ayivor, Nom...
country                                           United Kingdom
date_added                                            25-Jan-19
release_year                                               2018
rating                                                    TV-MA
duration                                              1 Season
listed_in     British TV Shows, International TV Shows, TV D...
description   Adopted by a human rights attorney after the R...
type                                                    TV Show
Name: 1988, dtype: object
```

19. Show the data types of all the columns

In [19]:

```
a.dtypes
```

Out[19]:

```
show_id          int64
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
duration        object
listed_in       object
description     object
type            object
dtype: object
```

20. Show the data types of rating column.

In [20]:

```
a['rating'].dtypes
```

Out[20]:

```
dtype('O')
```

21. Show the total nos of every category.

In [21]:

```
a.dtypes.value_counts()
```

Out[21]:

```
object    10
int64      2
dtype: int64
```

22. Show the total no of objects.

In [54]:

```
c.dtypes.value_counts()[0]
```

Out[54]:

```
10
```

23.Select the data set that contains all the data types except object.

In [22]:

```
a.select_dtypes(exclude=['object'])
```

Out[22]:

	show_id	release_year
0	81193313	2019
1	81197050	2019
2	81213894	2019
3	81082007	2019
4	80213643	2019
...	...	...
5832	70141644	1987
5833	70127998	2009
5834	70084180	2007
5835	70157452	2007
5836	70053412	2006

5837 rows × 2 columns

24. Print all the names of the title.

In [23]:

```
np.unique(a['title'])
```

Out[23]:

```
array(['#Roxy', '#Rucker50', '#Selfie', ..., '마녀사냥', '반드시 잡는다',
       '최강전사 미니특공대 : 영웅의 탄생'], dtype=object)
```

25. Check every cell is Null or not.

In [24]:

```
a.isnull()
```

Out[24]:

	show_id	title	director	cast	country	date_added	release_year	rating	duration	liste
0	False	False	True	False	False	False	False	False	False	F
1	False	False	False	False	True	False	False	False	False	F
2	False	False	False	False	False	False	False	False	False	F
3	False	False	False	False	False	False	False	False	False	F
4	False	False	True	False	False	True	False	False	False	F
...	...	...	...	...	...	...	...	...	...	
5832	False	False	False	False	False	False	False	False	False	F
5833	False	False	False	False	False	False	False	False	False	F
5834	False	False	False	False	False	False	False	False	False	F
5835	False	False	True	True	False	False	False	False	False	F
5836	False	False	False	False	False	False	False	False	False	F

5837 rows × 12 columns

26. Count the sum of all the columns.

In [25]:

```
a.isnull().sum()
```

Out[25]:

```
show_id            0
title              0
director        1901
cast             556
country          427
date_added       642
release_year       0
rating            10
duration           0
listed_in          0
description        0
type               0
dtype: int64
```

27. Check the data types of the above 5 columns which contains null values.

In [26]:

```
a['director'].dtypes
```

Out[26]:

```
dtype('O')
```

In [27]:

```
a['cast'].dtypes
```

Out[27]:

```
dtype('O')
```

In [28]:

```
a['country'].dtypes
```

Out[28]:

```
dtype('O')
```

In [29]:

```
a['date_added'].dtypes
```

Out[29]:

```
dtype('O')
```

In [30]:

```
a['rating'].dtypes
```

Out[30]:

```
dtype('O')
```

28. Show the count as per category from the type column.

In [31]:

```
a['type'].value_counts()
```

Out[31]:

```
Movie      3939
TV Show    1898
Name: type, dtype: int64
```

29. Create a copy of the above dataset.

In [32]:

```
c=a.copy()
```

30. Print the data set of the copied data set.

In [33]:

```
c
```

Out[33]:

	show_id	title	director	cast	country	date_added	release_year	rati
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	30-Nov-19	2019	TV-
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	30-Nov-19	2019	TV
2	81213894	The Zoya Factor	Abhishek Sharma	Sonam Kapoor, Dulquer Salmaan, Sanjay Kapoor, ...	India	30-Nov-19	2019	TV-
3	81082007	Atlantics	Mati Diop	Mama Sane, Amadou Mbow, Ibrahima Traore, Nicol...	France, Senegal, Belgium	29-Nov-19	2019	TV-
4	80213643	Chip and Potato	NaN	Abigail Oliver, Andrea Libman, Briana Buckmast...	Canada, United Kingdom	NaN	2019	TV
...	...	...	...	...	...	...	...	
5832	70141644	Mad Ron's Prevues from Hell	Jim Monaco	Nick Pawlow, Jordu Schell, Jay Kushwara, Micha...	United States	01-Nov-10	1987	I
5833	70127998	Splatter	Joe Dante	Corey Feldman, Tony Todd, Tara Leigh, Erin Way...	United States	18-Nov-09	2009	TV-
5834	70084180	Just Another Love Story	Ole Bornedal	Anders W. Berthelsen, Rebecka Hemse, Nikolaj L...	Denmark	05-May-09	2007	I
5835	70157452	Dinner for Five	NaN	NaN	United States	04-Feb-08	2007	T M

	show_id	title	director	cast	country	date_added	release_year	rati
5836	70053412	To and From New York	Sorin Dan Mihalcescu	Barbara King, Shaana Diya, John Krisiukenas,	United States	01-Jan-08	2006	I

31. Create a two way table of type and rating dropping the nan values.

In [34]:

```
pd.crosstab(index=a['type'],columns=a['rating'], dropna=True)
```

Out[34]:

rating	G	NC-17	NR	PG	PG-13	R	TV-14	TV-G	TV-MA	TV-PG	TV-Y	TV-Y7	TV-Y7-FV	UR
type														
Movie	31	2	202	160	227	437	955	79	1288	413	41	62	27	7
TV Show	1	0	16	0	0	2	638	68	649	265	98	94	65	0

32. Create a bar plot of type and rating using seaborn module.

In [35]:

```
import seaborn as sns
```

In [36]:

```
c
```

Out[36]:

	show_id	title	director	cast	country	date_added	release_year	rating
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	30-Nov-19	2019	TV-14
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	30-Nov-19	2019	TV-G
2	81213894	The Zoya Factor	Abhishek Sharma	Sonam Kapoor, Dulquer Salmaan, Sanjay Kapoor, ...	India	30-Nov-19	2019	TV-14
3	81082007	Atlantics	Mati Diop	Mama Sane, Amadou Mbow, Ibrahima Traore, Nicol...	France, Senegal, Belgium	29-Nov-19	2019	TV-14
4	80213643	Chip and Potato	NaN	Abigail Oliver, Andrea Libman, Briana Buckmast...	Canada, United Kingdom	NaN	2019	TV-Y
...	...	...	...	...	...	...	...	...
5832	70141644	Mad Ron's Prevues from Hell	Jim Monaco	Nick Pawlow, Jordu Schell, Jay Kushwara, Micha...	United States	01-Nov-10	1987	NR
5833	70127998	Splatter	Joe Dante	Corey Feldman, Tony Todd, Tara Leigh, Erin Way...	United States	18-Nov-09	2009	TV-14
5834	70084180	Just Another Love Story	Ole Bornedal	Anders W. Berthelsen, Rebecka Hemse, Nikolaj L...	Denmark	05-May-09	2007	NR
5835	70157452	Dinner for Five	NaN	NaN	United States	04-Feb-08	2007	TV-MA

	show_id	title	director	cast	country	date_added	release_year	rating
5836	70053412	To and From New York	Sorin Dan Mihalcescu	Barbara King, Shaana Diya, John Krisiukenas, Y...	United States	01-Jan-08	2006	NR

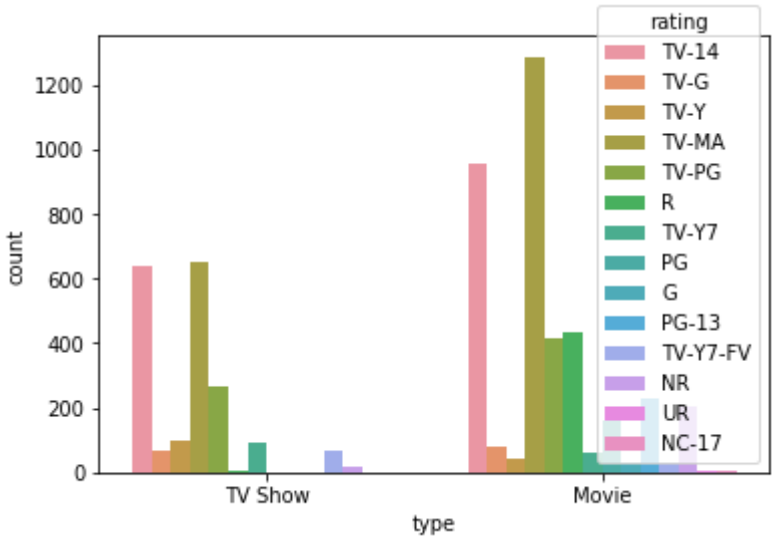
5837 rows × 12 columns

In [37]:

```
sns.countplot(x="type",data=c,hue="rating")
```

Out[37]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1db696e7688>
```



33. Show the correlation matrix

In [38]:

```
q=a.corr()
q
```

Out[38]:

	show_id	release_year
show_id	1.000000	0.536742
release_year	0.536742	1.000000

34. Show the description of the data set.

In [39]:

```
c.describe()
```

Out[39]:

	show_id	release_year
count	5.837000e+03	5837.000000
mean	7.730079e+07	2013.688539
std	9.479777e+06	8.419088
min	2.698800e+05	1925.000000
25%	8.004520e+07	2013.000000
50%	8.016353e+07	2016.000000
75%	8.024188e+07	2018.000000
max	8.122720e+07	2020.000000

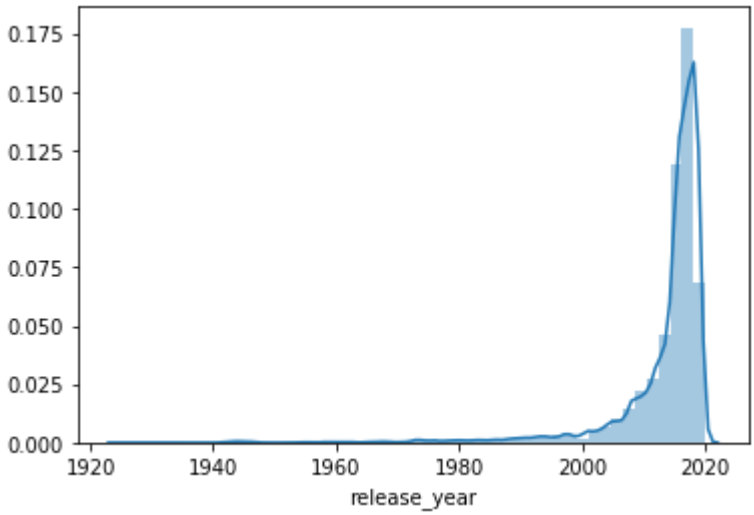
35. Show the histogram of release_year with the kernel density estimate.

In [41]:

```
sns.distplot(c["release_year"])
```

Out[41]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1db6c1da808>
```



36. Show the mean of the data set

In [42]:

```
c.mean()
```

Out[42]:

```
show_id         7.730079e+07
release_year    2.013689e+03
dtype: float64
```

37. Check the sumation of null values one more times.

In [43]:

```
c.isna().sum()
```

Out[43]:

```
show_id            0
title              0
director        1901
cast             556
country          427
date_added       642
release_year       0
rating            10
duration           0
listed_in          0
description        0
type               0
dtype: int64
```

38. Create a copy d2 of the data set.

In [44]:

```
c1=c.copy()
```

39. Show the value which has the highest counts in the column director

In [45]:

```
c1["director"].mode()
```

Out[45]:

```
0    Raúl Campos, Jan Suter
dtype: object
```

40. Replace the missing values of the director column by the value which has the highest counts in the column director

In [46]:

```
c1["director"].fillna(c1["director"].mode()[0],inplace=True)
```

41. Check the null value sum

In [48]:

```
c1.isna().sum()
```

Out[48]:

```
show_id          0
title            0
director         0
cast           556
country        427
date_added     642
release_year     0
rating          10
duration         0
listed_in        0
description      0
type             0
dtype: int64
```

42. Simlarly change the other columns which have the missing values with the values(Highest occurance).

In [50]:

```
c1=c1.apply(lambda x:x.fillna(x.mean()) if x.dtype=='float' else x.fillna(x.value_counts().
```

43. Check the null value sum

In [51]:

```
c1.isna().sum()
```

Out[51]:

```
show_id         0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
description     0
type            0
dtype: int64
```

In [ ]:

```

```

In [ ]:

```

```