Assignment pandas

In [1]:
```
import pandas as pd
print(pd.__version__)
```

```
2.2.2
```

In [2]:
```
pwd
```

Out[2]:
```
'C:\\Users\\ankit\\Downloads'
```

1. Import the attached Netflix csv file in Jupyter notebook and perform following operations using Pandas:

In [3]:
```
df = pd.read_csv('source panda.csv')
df
```

Out[3]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...
...	...	...	...	...	...	...	...	...	...	...	...
8775	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers
8776	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies
8777	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies
8778	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies
8779	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals

8780 rows × 11 columns

a) Print the first 5 rows and last 5 rows of the dataframe

In [4]:
```
df.head(5)
```

Out[4]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...

In [5]:
```
df.tail(5)
```

Out[5]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
8775	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers
8776	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies
8777	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies
8778	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies
8779	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals

b) Check how many rows and columns are there using Pandas function.

In [6]:
```
# Check the number of rows and columns
num_rows, num_columns = df.shape
print(f"Number of rows: {num_rows}, Number of columns: {num_columns}")
```

```
Number of rows: 8780, Number of columns: 11
```

c) Print all the column names

In [7]:
```
df.columns
```

Out[7]:
```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in'],
      dtype='object')
```

d) Calculate the descriptive statistics of all the variables(integer/float/object etc).

In [8]:
```
df.describe()
```

Out[8]:

	release_year
count	8780.000000
mean	2014.178474
std	8.827938
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

e) Check the number of unique values for each column

In [9]:
```
data=df.nunique()
print("Number of Unique Values for Each Column:")
print(data)
```

```
Number of Unique Values for Each Column:
show_id         8780
type               2
title           8780
director        4516
cast            7671
country          745
date_added      1765
release_year      74
rating            14
duration         220
listed_in        514
dtype: int64
```

f) Check the percentage of missing values for each column.

In [10]:

```
data =df.isnull().sum()
data
```

Out[10]:

```
show_id            0
type               0
title              0
director        2630
cast             824
country          828
date_added        10
release_year       0
rating             4
duration           0
listed_in          0
dtype: int64
```

In [11]:

```
# Calculate percentage of missing values for each column
missing_percentage = (df.isna().sum() / len(df)) * 100
print(missing_percentage)
```

```
show_id          0.000000
type             0.000000
title            0.000000
director        29.954442
cast             9.384966
country          9.430524
date_added       0.113895
release_year     0.000000
rating           0.045558
duration         0.000000
listed_in        0.000000
dtype: float64
```

g) Delete all the rows where Director column has missing values.

In [12]:

```
data = df.dropna(axis=0, how='all')
data
```

Out[12]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...
...	...	...	...	...	...	...	...	...	...	...	...
8775	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers
8776	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies
8777	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies
8778	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies
8779	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals

8780 rows × 11 columns

h) Print all the records where country has Germany value (including West Germany). If any other country is there along with Germany, then that row should also come in output

In [13]:

```
import numpy as np
import pandas as pd
```

In [14]:

```
df = pd.read_csv('source panda.csv')
df
```

Out[14]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...
...	...	...	...	...	...	...	...	...	...	...	...
8775	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers
8776	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies
8777	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies
8778	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies
8779	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals

8780 rows × 11 columns

In [15]:

```
op = ['Germany', 'West Germany']
condition = data['country'].isin(op) | data['country'].str.contains('Germany')
filtered = data[condition]
filtered
```

Out[15]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies
12	s13	Movie	Je Suis Karl	Christian Schwochow	Luna Wedler, Jannis Niewöhner, Milan Peschel, ...	Germany, Czech Republic	September 23, 2021	2021	TV-MA	127 min	Dramas, International Movies
129	s130	Movie	An Unfinished Life	Lasse Hallström	Robert Redford, Jennifer Lopez, Morgan Freeman...	Germany, United States	September 1, 2021	2005	PG-13	108 min	Dramas
142	s143	Movie	Freedom Writers	Richard LaGravenese	Hilary Swank, Patrick Dempsey, Scott Glenn, Im...	Germany, United States	September 1, 2021	2007	PG-13	124 min	Dramas
172	s173	Movie	School of Rock	Richard Linklater	Jack Black, Joan Cusack, Mike White, Sarah Sil...	United States, Germany	September 1, 2021	2003	PG-13	110 min	Comedies, Music & Musicals
...	...	...	...	...	...	...	...	...	...	...	...
8590	s8618	Movie	Trash	Stephen Daldry	Wagner Moura, Martin Sheen, Rooney Mara, Selto...	United Kingdom, Brazil, Germany	January 1, 2019	2014	R	114 min	Dramas, Independent Movies, Thrillers
8634	s8662	Movie	Unfinished Song	Paul Andrew Williams	Terence Stamp, Gemma Arterton, Christopher Ecc...	United Kingdom, Germany	July 22, 2019	2012	PG-13	94 min	Comedies, Dramas, Independent Movies
8641	s8669	Movie	V for Vendetta	James McTeigue	Natalie Portman, Hugo Weaving, Stephen Rea, St...	United States, United Kingdom, Germany	October 1, 2018	2005	R	132 min	Action & Adventure, Dramas, Sci-Fi & Fantasy
8702	s8730	Movie	Where the Money Is	Marek Kanievska	Paul Newman, Linda Fiorentino, Dermot Mulroney...	Germany, United States, United Kingdom, Canada	January 15, 2020	2000	PG-13	89 min	Action & Adventure, Comedies, Dramas
8718	s8746	Movie	Willy Wonka & the Chocolate Factory	Mel Stuart	Gene Wilder, Jack Albertson, Peter Ostrum, Roy...	United States, East Germany, West Germany	January 1, 2020	1971	G	100 min	Children & Family Movies, Classic Movies, Come...

231 rows × 11 columns

i) Expand Duration column into 2 separate columns – First column having the numeric value and other having String. Eg: 3 seasons should be split in 2 columns having 3 in 1st column and seasons in 2nd column

In [16]:

```
# Expand the duration column into numeric and string columns
df[['Duration_Value', 'Duration_Unit']] = df['duration'].str.extract(r'(\d+)\s*(\w+)')

# Convert Duration_Value to numeric type
df['Duration_Value'] = pd.to_numeric(df['Duration_Value'])

# Display the updated DataFrame
df
```

Out[16]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	Duration_Value	Duration_Unit
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	90	min
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	2	Seasons
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	1	Season
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	1	Season
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	2	Seasons
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8775	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	158	min
8776	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	2	Seasons
8777	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	88	min
8778	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	88	min
8779	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	111	min

8780 rows × 13 columns

j) Split Date added into 3 separate columns having date value in 1st column, month value in 2nd column and year value in 3rd.

In [17]:

```
df = pd.read_csv('source panda.csv')
df
```

Out[17]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...
...	...	...	...	...	...	...	...	...	...	...	...
8775	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers
8776	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies
8777	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies
8778	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies
8779	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals

8780 rows × 11 columns

In [18]:

```
# Extract day, month, and year from the date_added column
```

In [19]:

```
data=df[['month', 'day', 'year']] = df['date_added'].str.extract(r'(\w+)\s+(\d{1,2}),\s+(\d{4})')
```

In [20]:

```
# Display the updated DataFrame
data
```

Out[20]:

	0	1	2
0	September	25	2021
1	September	24	2021
2	September	24	2021
3	September	24	2021
4	September	24	2021
...	...	...	...
8775	November	20	2019
8776	July	1	2019
8777	November	1	2019
8778	January	11	2020
8779	March	2	2019

8780 rows × 3 columns

k) Print the number of TV shows/Movies released in each year.

In [ ]:

l) Rename the column title with movie_title

In [21]:

```python
import pandas as pd
```

In [22]:

```python
df = pd.read_csv('source panda.csv')
df
```

Out[22]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...
...	...	...	...	...	...	...	...	...	...	...	...
8775	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers
8776	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies
8777	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies
8778	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies
8779	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals

8780 rows × 11 columns

In [23]:

```python
df.rename(columns={'title': 'movie_title'}, inplace=True)
df
```

Out[23]:

	show_id	type	movie_title	director	cast	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...
...	...	...	...	...	...	...	...	...	...	...	...
8775	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers
8776	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies
8777	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies
8778	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies
8779	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals

8780 rows × 11 columns

m) Split Listed_in column into 3 different columns with col name (Genre1, Genre2, Genre3). Split the column based on comma.

In [24]:

```python
# Split the listed_in column into three new columns
df[['Genre1', 'Genre2', 'Genre3']] = df['listed_in'].str.split(',', expand=True)

# Display the updated DataFrame
df
```

Out[24]:

	show_id	type	movie_title	director	cast	country	date_added	release_year	rating	duration	listed_in	Genre1	Genre2	Genre3
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	Documentaries	None	None
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	International TV Shows	TV Dramas	TV Mysteries
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	Crime TV Shows	International TV Shows	TV Action & Adventure
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Docuseries	Reality TV	None
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	International TV Shows	Romantic TV Shows	TV Comedies
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8775	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	Cult Movies	Dramas	Thrillers
8776	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	Kids' TV	Korean TV Shows	TV Comedies
8777	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Comedies	Horror Movies	None
8778	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Children & Family Movies	Comedies	None
8779	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	Dramas	International Movies	Music & Musicals

8780 rows × 14 columns

In [ ]:

In [ ]:

In [ ]:

In [ ]:

import student.csv and marks.csv

In [25]:

```
import numpy as np
```

In [26]:

```
import pandas as pd
```

In [27]:

```
pwd
```

Out[27]:

```
'C:\\Users\\ankit\\Downloads'
```

In [28]:

```
df1=pd.read_csv('marks.csv')
df1
```

Out[28]:

	Student_id	Mark	City
0	1	95	Chennai
1	2	70	Delhi
2	3	98	Mumbai
3	4	75	Pune
4	5	89	Kochi
...	...	...	...
224	228	99	Pune
225	229	70	Chennai
226	230	55	Delhi
227	231	97	Mumbai
228	232	59	Pune

229 rows × 3 columns

In [29]:

```
df2=pd.read_csv('student_id panda.csv')
df2
```

Out[29]:

	Student_id	Age	Gender	Grade	Employed
0	1	19	Male	1st Class	yes
1	2	20	Female	2nd Class	no
2	3	18	Male	1st Class	no
3	4	21	Female	2nd Class	no
4	5	19	Male	1st Class	no
...	...	...	...	...	...
227	228	21	Female	1st Class	no
228	229	20	Male	2nd Class	no
229	230	20	Male	3rd Class	yes
230	231	19	Female	1st Class	yes
231	232	20	Male	3rd Class	yes

232 rows × 5 columns

Q2

a) Combine both the dataframes into single dataframe which will have all the records from both the tables.

In [30]:

```
df1 = pd.DataFrame(df1)
df2 = pd.DataFrame(df2)
# Combine both DataFrames
combined_df = pd.concat([df1, df2], ignore_index=True)
```

In [31]:

```
# Display the combined DataFrame
print(combined_df)
```

```
     Student_id  Mark     City   Age  Gender      Grade Employed
0             1  95.0  Chennai   NaN     NaN        NaN      NaN
1             2  70.0    Delhi   NaN     NaN        NaN      NaN
2             3  98.0   Mumbai   NaN     NaN        NaN      NaN
3             4  75.0     Pune   NaN     NaN        NaN      NaN
4             5  89.0    Kochi   NaN     NaN        NaN      NaN
..          ...   ...      ...   ...     ...        ...      ...
456         228   NaN      NaN  21.0  Female  1st Class       no
457         229   NaN      NaN  20.0    Male  2nd Class       no
458         230   NaN      NaN  20.0    Male  3rd Class      yes
459         231   NaN      NaN  19.0  Female  1st Class      yes
460         232   NaN      NaN  20.0    Male  3rd Class      yes

[461 rows x 7 columns]
```

b) Print the maximum and minimum marks Gender wise.

In [32]:
```
student=df1
marks=df2
```

In [33]:
```
df1=pd.read_csv('marks.csv')
df2=pd.read_csv('student_id panda.csv')
```

In [34]:
```
df=pd.merge(student,marks,how="right",on="Student_id")
```

In [35]:
```
df.groupby("Gender")["Mark"].max()
```

Out[35]:
```
Gender
Female     99.0
Male      100.0
Name: Mark, dtype: float64
```

In [ ]:

c) Print all the students IDs and their marks who have scored more than the average marks of the class

In [36]:
```
# Calculate the average marks of the class
```

In [37]:
```
df1=pd.merge(student,marks,how="right",on="Student_id")
df1
```

Out[37]:

	Student_id	Mark	City	Age	Gender	Grade	Employed
0	1	95.0	Chennai	19	Male	1st Class	yes
1	2	70.0	Delhi	20	Female	2nd Class	no
2	3	98.0	Mumbai	18	Male	1st Class	no
3	4	75.0	Pune	21	Female	2nd Class	no
4	5	89.0	Kochi	19	Male	1st Class	no
...	...	...	...	...	...	...	...
227	228	99.0	Pune	21	Female	1st Class	no
228	229	70.0	Chennai	20	Male	2nd Class	no
229	230	55.0	Delhi	20	Male	3rd Class	yes
230	231	97.0	Mumbai	19	Female	1st Class	yes
231	232	59.0	Pune	20	Male	3rd Class	yes

232 rows × 7 columns

d) Print the dataframe who are Males and are Employed.

In [38]:
```
# Filter for Males who are Employed
```

In [39]:
```
print("Original DataFrame:")
print(df)
```

```
Original DataFrame:
     Student_id  Mark     City  Age  Gender      Grade Employed
0             1  95.0  Chennai   19    Male  1st Class      yes
1             2  70.0    Delhi   20  Female  2nd Class       no
2             3  98.0   Mumbai   18    Male  1st Class       no
3             4  75.0     Pune   21  Female  2nd Class       no
4             5  89.0    Kochi   19    Male  1st Class       no
..          ...   ...      ...  ...     ...        ...      ...
227         228  99.0     Pune   21  Female  1st Class       no
228         229  70.0  Chennai   20    Male  2nd Class       no
229         230  55.0    Delhi   20    Male  3rd Class      yes
230         231  97.0   Mumbai   19  Female  1st Class      yes
231         232  59.0     Pune   20    Male  3rd Class      yes

[232 rows x 7 columns]
```

In [64]:
```
import pandas as pd
student=pd.read_csv('student_id panda.csv')
marks=pd.read_csv('marks.csv')

male_emp=student[(student['Gender']=='Male') & (student['Employed']=='yes')]
print(male_emp)
```

```
     Student_id  Age Gender      Grade Employed
0             1   19    Male  1st Class       yes
5             6   20    Male  2nd Class       yes
7             8   21    Male  3rd Class       yes
12           13   19    Male  1st Class       yes
14           15   19    Male  1st Class       yes
16           17   20    Male  2nd Class       yes
18           19   21    Male  2nd Class       yes
29           30   19    Male  1st Class       yes
34           35   20    Male  2nd Class       yes
36           37   21    Male  3rd Class       yes
41           42   19    Male  1st Class       yes
43           44   19    Male  1st Class       yes
45           46   20    Male  2nd Class       yes
47           48   21    Male  2nd Class       yes
58           59   19    Male  1st Class       yes
63           64   20    Male  2nd Class       yes
65           66   21    Male  3rd Class       yes
70           71   19    Male  1st Class       yes
72           73   19    Male  1st Class       yes
74           75   20    Male  2nd Class       yes
76           77   21    Male  2nd Class       yes
87           88   19    Male  1st Class       yes
92           93   20    Male  2nd Class       yes
94           95   21    Male  3rd Class       yes
99          100   19    Male  1st Class       yes
101         102   19    Male  1st Class       yes
103         104   20    Male  2nd Class       yes
105         106   21    Male  2nd Class       yes
116         117   19    Male  1st Class       yes
121         122   20    Male  2nd Class       yes
123         124   21    Male  3rd Class       yes
128         129   19    Male  1st Class       yes
130         131   19    Male  1st Class       yes
132         133   20    Male  2nd Class       yes
134         135   21    Male  2nd Class       yes
145         146   19    Male  1st Class       yes
150         151   20    Male  2nd Class       yes
152         153   21    Male  3rd Class       yes
157         158   19    Male  1st Class       yes
159         160   19    Male  1st Class       yes
161         162   20    Male  2nd Class       yes
163         164   21    Male  2nd Class       yes
174         175   19    Male  1st Class       yes
179         180   20    Male  2nd Class       yes
181         182   21    Male  3rd Class       yes
186         187   19    Male  1st Class       yes
188         189   19    Male  1st Class       yes
190         191   20    Male  2nd Class       yes
192         193   21    Male  2nd Class       yes
203         204   19    Male  1st Class       yes
208         209   20    Male  2nd Class       yes
210         211   21    Male  3rd Class       yes
215         216   19    Male  1st Class       yes
217         218   19    Male  1st Class       yes
219         220   20    Male  2nd Class       yes
221         222   21    Male  2nd Class       yes
229         230   20    Male  3rd Class       yes
231         232   20    Male  3rd Class       yes
```

e) Create a new Column ‘IQ_level’ which will have 3 values (Intelligent, Mediocre, weak). If student scored than 80 then Tag him as Intelligent, if student scored between 50-80, then Mediocre, else weak.

In [44]:

```python
import pandas as pd
```

In [46]:

```python
def IQ(Mark):
    if Mark > 80:
        return 'Intelligent'
    elif Mark >= 50:
        return 'Mediocre'
    else:
        return 'Weak'
```

In [48]:

```python
df['IQ_level'] = df['Mark'].apply(iq)

df
```

Out[48]:

	Student_id	Mark	City	Age	Gender	Grade	Employed	IQ_level
0	1	95.0	Chennai	19	Male	1st Class	yes	Intelligent
1	2	70.0	Delhi	20	Female	2nd Class	no	Mediocre
2	3	98.0	Mumbai	18	Male	1st Class	no	Intelligent
3	4	75.0	Pune	21	Female	2nd Class	no	Mediocre
4	5	89.0	Kochi	19	Male	1st Class	no	Intelligent
...	...	...	...	...	...	...	...	...
227	228	99.0	Pune	21	Female	1st Class	no	Intelligent
228	229	70.0	Chennai	20	Male	2nd Class	no	Mediocre
229	230	55.0	Delhi	20	Male	3rd Class	yes	Mediocre
230	231	97.0	Mumbai	19	Female	1st Class	yes	Intelligent
231	232	59.0	Pune	20	Male	3rd Class	yes	Mediocre

232 rows × 8 columns

f) Count the number of males and females from each city

In [65]:

```python
import pandas as pd
```

In [70]:

```python
df1=pd.read_csv('marks.csv')
```

In [73]:

```python
df2=pd.read_csv('student_id panda.csv')
```

In [74]:

```python
combined_df
```

Out[74]:

	Student_id	Mark	City	Age	Gender	Grade	Employed
0	1	95.0	Chennai	NaN	NaN	NaN	NaN
1	2	70.0	Delhi	NaN	NaN	NaN	NaN
2	3	98.0	Mumbai	NaN	NaN	NaN	NaN
3	4	75.0	Pune	NaN	NaN	NaN	NaN
4	5	89.0	Kochi	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...
456	228	NaN	NaN	21.0	Female	1st Class	no
457	229	NaN	NaN	20.0	Male	2nd Class	no
458	230	NaN	NaN	20.0	Male	3rd Class	yes
459	231	NaN	NaN	19.0	Female	1st Class	yes
460	232	NaN	NaN	20.0	Male	3rd Class	yes

461 rows × 7 columns

In [ ]:

In [ ]:

g) Print the top 5 Male scorers.

In [88]:

```
import pandas as pd
```

In [92]:

```
df2=pd.read_csv('student_id panda.csv')
```

In [95]:

```
duplicates = df[df.duplicated(subset='student_ID', keep=False)]
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_57444\2403296154.py in ?()
----> 1 duplicates = df[df.duplicated(subset='student_ID', keep=True)]

C:\ProgramData\anaconda3\Lib\site-packages\pandas\core\frame.py in ?(self, subset, keep)
   6946         # Otherwise, raise a KeyError, same as if you try to __getitem__ with a
   6947         # key that doesn't exist.
   6948         diff = set(subset) - set(self.columns)
   6949         if diff:
-> 6950             raise KeyError(Index(diff))
   6951 
   6952         if len(subset) == 1 and self.columns.is_unique:
   6953             # GH#45236 This is faster than get_group_index below

KeyError: Index(['student_ID'], dtype='object')
```

In [ ]:

h) Replace the Male value with M and Female value with F and export this dataframe to excel file in D: (D drive) and name the file as test.csv.

In [ ]:

i) Check if any student_ID is duplicated

In [ ]:

In [ ]:

j) Create a separate dataframe which will have all the Integer/Float variables.

In [96]:

```
print("Original DataFrame:")
print(df)
```

```
Original DataFrame:
     Student_id  Mark     City
0             1    95  Chennai
1             2    70    Delhi
2             3    98   Mumbai
3             4    75     Pune
4             5    89    Kochi
..          ...   ...      ...
224         228    99     Pune
225         229    70  Chennai
226         230    55    Delhi
227         231    97   Mumbai
228         232    59     Pune

[229 rows x 3 columns]
```

In [97]:

```
numeric_df = df.select_dtypes(include=['int64', 'float64'])
```

In [98]:

```
print("\nDataFrame with Integer and Float Variables:")
print(numeric_df)
```

```
DataFrame with Integer and Float Variables:
     Student_id  Mark
0             1    95
1             2    70
2             3    98
3             4    75
4             5    89
..          ...   ...
224         228    99
225         229    70
226         230    55
227         231    97
228         232    59

[229 rows x 2 columns]
```

In [ ]:

k) Get those Student_IDs which are present in Students table but not in Marks table

In [100…

```
import pandas as pd
```

In [101…

```
df2=pd.read_csv('student_id panda.csv')
```

In [104…

```
df2
```

Out[104…

	Student_id	Age	Gender	Grade	Employed
0	1	19	Male	1st Class	yes
1	2	20	Female	2nd Class	no
2	3	18	Male	1st Class	no
3	4	21	Female	2nd Class	no
4	5	19	Male	1st Class	no
...	...	...	...	...	...
227	228	21	Female	1st Class	no
228	229	20	Male	2nd Class	no
229	230	20	Male	3rd Class	yes
230	231	19	Female	1st Class	yes
231	232	20	Male	3rd Class	yes

232 rows × 5 columns

In [105…

```
not_in_marks = df2.merge(marks_df, on='Student_ID', how='left', indicator=True)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[105], line 1
----> 1 not_in_marks = df2.merge(marks_df, on='Student_ID', how='left', indicator=True)

NameError: name 'marks_df' is not defined
```

In [ ]:

In [ ]:

In [ ]:

3. Explain the concept of missing values. How can you identify the missing values in a Pandas DataFrame

?### What are the different ways of treating/Imputing/Deleting the missing values### . Explain with example>

<strike style="color: rgb(255, 255, 255);">explore more</strike><br> seats are full!!   <br>
Missing values refer to the absence of data in a dataset. They can occur for various reasons, such as: Data not being collected or recorded. Errors during data entry. Data being filtered out or excluded. Differences in data sources or formats. Missing values can be represented in various ways, including: NaN (Not a Number) in numerical data. None in Python. Blank entries or placeholders like -999 or "NULL". Handling missing values is crucial because they can significantly impact statistical analyses, machine learning models, and overall data quality.
##

In [106…

```
# 1. Removing Missing Values
# a. Drop Rows with Missing Values
# You can remove rows that contain any missing values using the dropna() method.

# python
# Copy code
import pandas as pd

# Sample DataFrame
data = {
    'A': [1, 2, None],
    'B': [4, None, 6],
    'C': [None, None, 9]
}
df = pd.DataFrame(data)

# Drop rows with any missing values
cleaned_df = df.dropna()
print("DataFrame after dropping rows with missing values:")
print(cleaned_df)
```

```
DataFrame after dropping rows with missing values:
Empty DataFrame
Columns: [A, B, C]
Index: []
```

In [107…

```
# 2. Imputing Missing Values
 # a. Fill with a Constant Value
# You can fill missing values with a specific constant, such as 0 or -1.

# python
# Copy code
# Fill missing values with a constant (e.g., 0)
filled_df = df.fillna(0)
print("\nDataFrame after filling missing values with 0:")
print(filled_df)
```

```
DataFrame after filling missing values with 0:
     A    B    C
0  1.0  4.0  0.0
1  2.0  0.0  0.0
2  0.0  6.0  9.0
```

In [ ]:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js