Assignment-2_Pandas_Kaustubh_Prakash_Deolase

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

In [51]:
```
import pandas as pd
```

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

In [127…
```
df = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\netflix_titles.csv")
print(df.head(5))
```

```
  show_id     type                  title         director  \
0      s1    Movie   Dick Johnson Is Dead  Kirsten Johnson   
1      s2  TV Show          Blood & Water              NaN   
2      s3  TV Show              Ganglands  Julien Leclercq   
3      s4  TV Show  Jailbirds New Orleans              NaN   
4      s5  TV Show           Kota Factory              NaN   

                                                cast        country  \
0                                                NaN  United States   
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa   
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN   
3                                                NaN            NaN   
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India   

          date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min   
1  September 24, 2021          2021  TV-MA  2 Seasons   
2  September 24, 2021          2021  TV-MA   1 Season   
3  September 24, 2021          2021  TV-MA   1 Season   
4  September 24, 2021          2021  TV-MA  2 Seasons   

                                     listed_in  
0                                Documentaries  
1   International TV Shows, TV Dramas, TV Mysteries  
2  Crime TV Shows, International TV Shows, TV Act...  
3                            Docuseries, Reality TV  
4  International TV Shows, Romantic TV Shows, TV ...  
```

In [129…
```
print(df.tail(5))
```

```
     show_id     type        title         director  \
8775   s8803    Movie       Zodiac    David Fincher
8776   s8804  TV Show  Zombie Dumb              NaN
8777   s8805    Movie   Zombieland  Ruben Fleischer
8778   s8806    Movie         Zoom     Peter Hewitt
8779   s8807    Movie       Zubaan      Mozez Singh

                                                   cast        country  \
8775  Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...  United States
8776                                                NaN            NaN
8777  Jesse Eisenberg, Woody Harrelson, Emma Stone, ...  United States
8778  Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...  United States
8779  Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...          India

            date_added  release_year rating   duration  \
8775  November 20, 2019          2007      R    158 min
8776       July 1, 2019          2018  TV-Y7  2 Seasons
8777   November 1, 2019          2009      R     88 min
8778   January 11, 2020          2006     PG     88 min
8779      March 2, 2019          2015  TV-14    111 min

                                      listed_in
8775                Cult Movies, Dramas, Thrillers
8776         Kids' TV, Korean TV Shows, TV Comedies
8777                        Comedies, Horror Movies
8778            Children & Family Movies, Comedies
8779  Dramas, International Movies, Music & Musicals
```

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

In [56]:
```
import pandas as pd
df = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\netflix_titles.csv")

#================================#
#Funtion & Output:

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8780 entries, 0 to 8779
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8780 non-null   object
 1   type          8780 non-null   object
 2   title         8780 non-null   object
 3   director      6150 non-null   object
 4   cast          7956 non-null   object
 5   country       7952 non-null   object
 6   date_added    8770 non-null   object
 7   release_year  8780 non-null   int64 
 8   rating        8776 non-null   object
 9   duration      8780 non-null   object
 10  listed_in     8780 non-null   object
dtypes: int64(1), object(10)
memory usage: 754.7+ KB
```

c) Print all the column names.

In [58]:
```
import pandas as pd
df = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\netflix_titles.csv")

#================================#
#Funtion & Output:

print(df.columns)
```

```
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 [60]:
```
import pandas as pd
df = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\netflix_titles.csv")

#================================#
#Funtion & Output:

df.describe()
```

Out[60]:

	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 [62]:
```
import pandas as pd
df = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\netflix_titles.csv")

#================================#
#Funtion & Output:

df.nunique()
```

Out[62]:
```
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 [64]:
```
import pandas as pd
df = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\netflix_titles.csv")

#================================#
#Funtion & Output:

missing_report = df.isnull().sum().reset_index()
missing_report.columns = ['Column', 'Missing Count']
missing_report['Missing Percentage'] = (missing_report['Missing Count'] / len(df))

# Print report
print(missing_report)
```

```
          Column  Missing Count  Missing Percentage
0        show_id              0            0.000000
1           type              0            0.000000
2          title              0            0.000000
3       director           2630           29.954442
4           cast            824            9.384966
5        country            828            9.430524
6     date_added             10            0.113895
7   release_year              0            0.000000
8         rating              4            0.045558
9       duration              0            0.000000
10     listed_in              0            0.000000
```

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

In [131…
```
import pandas as pd
df = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\netflix_titles.csv")

#================================#
#Funtion & Output:

print(df[df['director'].notna()])
```

```
     show_id     type                             title  \
0         s1    Movie              Dick Johnson Is Dead
2         s3  TV Show                         Ganglands
5         s6  TV Show                     Midnight Mass
6         s7    Movie  My Little Pony: A New Generation
7         s8    Movie                           Sankofa
...      ...      ...                               ...
8774   s8802    Movie                          Zinzana
8775   s8803    Movie                           Zodiac
8777   s8805    Movie                       Zombieland
8778   s8806    Movie                             Zoom
8779   s8807    Movie                           Zubaan

                          director  \
0                  Kirsten Johnson
2                  Julien Leclercq
5                    Mike Flanagan
6     Robert Cullen, JosÃ© Luis Ucha
7                     Haile Gerima
...                            ...
8774               Majid Al Ansari
8775                 David Fincher
8777               Ruben Fleischer
8778                  Peter Hewitt
8779                   Mozez Singh

                                                   cast  \
0                                                   NaN
2     Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...
5     Kate Siegel, Zach Gilford, Hamish Linklater, H...
6     Vanessa Hudgens, Kimiko Glenn, James Marsden, ...
7     Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...
...                                                 ...
8774  Ali Suliman, Saleh Bakri, Yasa, Ali Al-Jabri, ...
8775  Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...
8777  Jesse Eisenberg, Woody Harrelson, Emma Stone, ...
8778  Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...
8779  Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...

                                            country          date_added  \
0                                     United States  September 25, 2021
2                                               NaN  September 24, 2021
5                                               NaN  September 24, 2021
6                                               NaN  September 24, 2021
7     United States, Ghana, Burkina Faso, United Kin...  September 24, 2021
...                                             ...                 ...
8774                United Arab Emirates, Jordan       March 9, 2016
8775                                United States   November 20, 2019
8777                                United States    November 1, 2019
8778                                United States    January 11, 2020
8779                                        India       March 2, 2019

      release_year rating  duration  \
0             2020  PG-13    90 min
2             2021  TV-MA  1 Season
5             2021  TV-MA  1 Season
```

```
6             2021     PG    91 min
7             1993  TV-MA   125 min
...            ...    ...       ...
8774          2015  TV-MA    96 min
8775          2007      R   158 min
8777          2009      R    88 min
8778          2006     PG    88 min
8779          2015  TV-14   111 min

                                              listed_in
0                                          Documentaries
2     Crime TV Shows, International TV Shows, TV Act...
5                   TV Dramas, TV Horror, TV Mysteries
6                             Children & Family Movies
7      Dramas, Independent Movies, International Movies
...                                                 ...
8774           Dramas, International Movies, Thrillers
8775                    Cult Movies, Dramas, Thrillers
8777                          Comedies, Horror Movies
8778                Children & Family Movies, Comedies
8779     Dramas, International Movies, Music & Musicals

[6150 rows x 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 [133…

```
import pandas as pd
df = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\netflix_titles.csv")

#================================#
#Funtion & Output:

print(df[df['country'].str.contains('Germany|West Germany', case=False, na=False)])
```

```
     show_id   type                                title  \
7         s8  Movie                              Sankofa
12       s13  Movie                         Je Suis Karl
129     s130  Movie                   An Unfinished Life
142     s143  Movie                      Freedom Writers
172     s173  Movie                       School of Rock
...      ...    ...                                  ...
8590   s8618  Movie                                Trash
8634   s8662  Movie                      Unfinished Song
8641   s8669  Movie                       V for Vendetta
8702   s8730  Movie                    Where the Money Is
8718   s8746  Movie  Willy Wonka & the Chocolate Factory

                director                                               cast  \
7            Haile Gerima  Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...
12    Christian Schwochow  Luna Wedler, Jannis NiewÃ¶hner, Milan Peschel,...
129      Lasse HallstrÃ¶m  Robert Redford, Jennifer Lopez, Morgan Freeman...
142   Richard LaGravenese  Hilary Swank, Patrick Dempsey, Scott Glenn, Im...
172     Richard Linklater  Jack Black, Joan Cusack, Mike White, Sarah Sil...
...                   ...                                                ...
8590       Stephen Daldry  Wagner Moura, Martin Sheen, Rooney Mara, Selto...
8634  Paul Andrew Williams  Terence Stamp, Gemma Arterton, Christopher Ecc...
8641       James McTeigue  Natalie Portman, Hugo Weaving, Stephen Rea, St...
8702      Marek Kanievska  Paul Newman, Linda Fiorentino, Dermot Mulroney...
8718           Mel Stuart  Gene Wilder, Jack Albertson, Peter Ostrum, Roy...

                                              country          date_added  \
7     United States, Ghana, Burkina Faso, United Kin...  September 24, 2021
12                            Germany, Czech Republic  September 23, 2021
129                            Germany, United States   September 1, 2021
142                            Germany, United States   September 1, 2021
172                            United States, Germany   September 1, 2021
...                                               ...                ...
8590                   United Kingdom, Brazil, Germany     January 1, 2019
8634                           United Kingdom, Germany       July 22, 2019
8641            United States, United Kingdom, Germany     October 1, 2018
8702    Germany, United States, United Kingdom, Canada    January 15, 2020
8718          United States, East Germany, West Germany     January 1, 2020

      release_year rating duration  \
7             1993  TV-MA  125 min
12            2021  TV-MA  127 min
129           2005  PG-13  108 min
142           2007  PG-13  124 min
172           2003  PG-13  110 min
...            ...    ...      ...
8590          2014      R  114 min
8634          2012  PG-13   94 min
8641          2005      R  132 min
8702          2000  PG-13   89 min
8718          1971      G  100 min

                                         listed_in
7      Dramas, Independent Movies, International Movies
12                        Dramas, International Movies
129                                         Dramas
```

```
142                                                Dramas
172                            Comedies, Music & Musicals
...                                                   ...
8590               Dramas, Independent Movies, Thrillers
8634                Comedies, Dramas, Independent Movies
8641       Action & Adventure, Dramas, Sci-Fi & Fantasy
8702                Action & Adventure, Comedies, Dramas
8718  Children & Family Movies, Classic Movies, Come...

[231 rows x 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 [135…

```
import pandas as pd
df = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\netflix_titles.csv")

#================================#
#Funtion & Output:

df['Duration_Value'] = df['duration'].apply(lambda x: int(''.join(filter(str.isdigi
df['Duration_Unit'] = df['duration'].apply(lambda x: ''.join(filter(lambda c: not c

print(df)
```

```
     show_id     type                  title         director  \
0         s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1         s2  TV Show          Blood & Water              NaN
2         s3  TV Show              Ganglands  Julien Leclercq
3         s4  TV Show  Jailbirds New Orleans              NaN
4         s5  TV Show           Kota Factory              NaN
...      ...      ...                   ...              ...
8775   s8803    Movie                Zodiac    David Fincher
8776   s8804  TV Show            Zombie Dumb              NaN
8777   s8805    Movie             Zombieland  Ruben Fleischer
8778   s8806    Movie                  Zoom     Peter Hewitt
8779   s8807    Movie                Zubaan      Mozez Singh

                                                   cast        country  \
0                                                   NaN  United States
1     Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2     Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                   NaN            NaN
4     Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India
...                                                 ...            ...
8775  Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...  United States
8776                                                NaN            NaN
8777  Jesse Eisenberg, Woody Harrelson, Emma Stone, ...  United States
8778  Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...  United States
8779  Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...          India

              date_added  release_year rating   duration  \
0     September 25, 2021          2020  PG-13     90 min
1     September 24, 2021          2021  TV-MA  2 Seasons
2     September 24, 2021          2021  TV-MA   1 Season
3     September 24, 2021          2021  TV-MA   1 Season
4     September 24, 2021          2021  TV-MA  2 Seasons
...                  ...           ...    ...        ...
8775   November 20, 2019          2007      R    158 min
8776        July 1, 2019          2018  TV-Y7  2 Seasons
8777    November 1, 2019          2009      R     88 min
8778    January 11, 2020          2006     PG     88 min
8779       March 2, 2019          2015  TV-14    111 min

                                          listed_in  Duration_Value  \
0                                     Documentaries              90
1       International TV Shows, TV Dramas, TV Mysteries               2
2     Crime TV Shows, International TV Shows, TV Act...               1
3                             Docuseries, Reality TV                1
4     International TV Shows, Romantic TV Shows, TV ...               2
...                                                 ...             ...
8775                    Cult Movies, Dramas, Thrillers             158
8776            Kids' TV, Korean TV Shows, TV Comedies               2
8777                           Comedies, Horror Movies              88
8778                Children & Family Movies, Comedies              88
8779    Dramas, International Movies, Music & Musicals             111

     Duration_Unit
0              min
1          Seasons
2           Season
```

```
3           Season
4          Seasons
...            ...
8775           min
8776       Seasons
8777           min
8778           min
8779           min

[8780 rows x 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 [73]:
```
import pandas as pd
df = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\netflix_titles.csv")

#================================#
#Funtion & Output:

df['Date added'] = pd.to_datetime(df['date_added'], errors='coerce', format='mixed'

df['Month'] = df['Date added'].dt.month_name()
df['Date'] = df['Date added'].dt.day
df['Year'] = df['Date added'].dt.year
print(df)
```

```
     show_id     type                  title         director  \
0         s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1         s2  TV Show          Blood & Water              NaN
2         s3  TV Show              Ganglands  Julien Leclercq
3         s4  TV Show  Jailbirds New Orleans              NaN
4         s5  TV Show           Kota Factory              NaN
...      ...      ...                   ...              ...
8775   s8803    Movie                Zodiac    David Fincher
8776   s8804  TV Show            Zombie Dumb              NaN
8777   s8805    Movie             Zombieland  Ruben Fleischer
8778   s8806    Movie                  Zoom     Peter Hewitt
8779   s8807    Movie                Zubaan      Mozez Singh

                                                  cast        country  \
0                                                  NaN  United States
1     Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2     Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                  NaN            NaN
4     Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India
...                                                 ...            ...
8775  Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...  United States
8776                                               NaN            NaN
8777  Jesse Eisenberg, Woody Harrelson, Emma Stone, ...  United States
8778  Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...  United States
8779  Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...          India

              date_added  release_year rating   duration  \
0     September 25, 2021          2020  PG-13     90 min
1     September 24, 2021          2021  TV-MA  2 Seasons
2     September 24, 2021          2021  TV-MA   1 Season
3     September 24, 2021          2021  TV-MA   1 Season
4     September 24, 2021          2021  TV-MA  2 Seasons
...                  ...           ...    ...        ...
8775   November 20, 2019          2007      R    158 min
8776        July 1, 2019          2018  TV-Y7  2 Seasons
8777    November 1, 2019          2009      R     88 min
8778    January 11, 2020          2006     PG     88 min
8779       March 2, 2019          2015  TV-14    111 min

                                          listed_in Date added      Month  \
0                                     Documentaries 2021-09-25  September
1       International TV Shows, TV Dramas, TV Mysteries 2021-09-24  September
2     Crime TV Shows, International TV Shows, TV Act... 2021-09-24  September
3                             Docuseries, Reality TV 2021-09-24  September
4     International TV Shows, Romantic TV Shows, TV ... 2021-09-24  September
...                                                ...        ...        ...
8775                  Cult Movies, Dramas, Thrillers 2019-11-20   November
8776            Kids' TV, Korean TV Shows, TV Comedies 2019-07-01       July
8777                        Comedies, Horror Movies 2019-11-01   November
8778               Children & Family Movies, Comedies 2020-01-11    January
8779    Dramas, International Movies, Music & Musicals 2019-03-02      March

      Date    Year
0     25.0  2021.0
1     24.0  2021.0
2     24.0  2021.0
```

```
3     24.0  2021.0
4     24.0  2021.0
...    ...     ...
8775  20.0  2019.0
8776   1.0  2019.0
8777   1.0  2019.0
8778  11.0  2020.0
8779   2.0  2019.0

[8780 rows x 15 columns]
```

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

In [75]:
```
import pandas as pd

df = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\netflix_titles.csv")
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce', format='mixed'

#================================#
#Funtion & Output:

yearly_releases = df.groupby(df['date_added'].dt.year).size().reset_index(name='Num

print(yearly_releases)
```

```
    date_added  Number of Releases
0       2008.0                   2
1       2009.0                   2
2       2010.0                   1
3       2011.0                  13
4       2012.0                   3
5       2013.0                  11
6       2014.0                  24
7       2015.0                  82
8       2016.0                 426
9       2017.0                1183
10      2018.0                1645
11      2019.0                2007
12      2020.0                1875
13      2021.0                1496
```

l) Rename the column title with movie_title.

In [77]:
```
import pandas as pd
df = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\netflix_titles.csv")

#================================#
#Funtion & Output:

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

```
     show_id     type            movie_title         director  \
0         s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1         s2  TV Show          Blood & Water              NaN
2         s3  TV Show              Ganglands  Julien Leclercq
3         s4  TV Show  Jailbirds New Orleans              NaN
4         s5  TV Show           Kota Factory              NaN
...      ...      ...                    ...              ...
8775   s8803    Movie                Zodiac    David Fincher
8776   s8804  TV Show            Zombie Dumb              NaN
8777   s8805    Movie             Zombieland  Ruben Fleischer
8778   s8806    Movie                  Zoom     Peter Hewitt
8779   s8807    Movie                Zubaan      Mozez Singh

                                                  cast        country  \
0                                                  NaN  United States
1     Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2     Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                  NaN            NaN
4     Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India
...                                                 ...            ...
8775  Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...  United States
8776                                               NaN            NaN
8777  Jesse Eisenberg, Woody Harrelson, Emma Stone, ...  United States
8778  Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...  United States
8779  Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...          India

              date_added  release_year rating   duration  \
0     September 25, 2021          2020  PG-13     90 min
1     September 24, 2021          2021  TV-MA  2 Seasons
2     September 24, 2021          2021  TV-MA   1 Season
3     September 24, 2021          2021  TV-MA   1 Season
4     September 24, 2021          2021  TV-MA  2 Seasons
...                  ...           ...    ...        ...
8775   November 20, 2019          2007      R    158 min
8776        July 1, 2019          2018  TV-Y7  2 Seasons
8777    November 1, 2019          2009      R     88 min
8778    January 11, 2020          2006     PG     88 min
8779       March 2, 2019          2015  TV-14    111 min

                                          listed_in
0                                     Documentaries
1       International TV Shows, TV Dramas, TV Mysteries
2     Crime TV Shows, International TV Shows, TV Act...
3                             Docuseries, Reality TV
4     International TV Shows, Romantic TV Shows, TV ...
...                                                 ...
8775                    Cult Movies, Dramas, Thrillers
8776             Kids' TV, Korean TV Shows, TV Comedies
8777                           Comedies, Horror Movies
8778                Children & Family Movies, Comedies
8779    Dramas, International Movies, Music & Musicals

[8780 rows x 11 columns]
```

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

In [79]:
```
import pandas as pd
df = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\netflix_titles.csv")

#================================#
#Funtion & Output:

df[['Genre1', 'Genre2', 'Genre3']] = df['listed_in'].str.split(',', n=2, expand=Tru
print(df)
```

```
     show_id     type                  title         director  \
0         s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1         s2  TV Show          Blood & Water              NaN
2         s3  TV Show              Ganglands  Julien Leclercq
3         s4  TV Show  Jailbirds New Orleans              NaN
4         s5  TV Show           Kota Factory              NaN
...      ...      ...                    ...              ...
8775   s8803    Movie                Zodiac    David Fincher
8776   s8804  TV Show            Zombie Dumb              NaN
8777   s8805    Movie             Zombieland  Ruben Fleischer
8778   s8806    Movie                  Zoom     Peter Hewitt
8779   s8807    Movie                Zubaan      Mozez Singh

                                                  cast        country  \
0                                                  NaN  United States
1     Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2     Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                  NaN            NaN
4     Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India
...                                                 ...            ...
8775  Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...  United States
8776                                               NaN            NaN
8777  Jesse Eisenberg, Woody Harrelson, Emma Stone, ...  United States
8778  Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...  United States
8779  Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...          India

              date_added  release_year rating   duration  \
0     September 25, 2021          2020  PG-13     90 min
1     September 24, 2021          2021  TV-MA  2 Seasons
2     September 24, 2021          2021  TV-MA   1 Season
3     September 24, 2021          2021  TV-MA   1 Season
4     September 24, 2021          2021  TV-MA  2 Seasons
...                  ...           ...    ...        ...
8775   November 20, 2019          2007      R    158 min
8776        July 1, 2019          2018  TV-Y7  2 Seasons
8777    November 1, 2019          2009      R     88 min
8778    January 11, 2020          2006     PG     88 min
8779       March 2, 2019          2015  TV-14    111 min

                                             listed_in  \
0                                       Documentaries
1       International TV Shows, TV Dramas, TV Mysteries
2     Crime TV Shows, International TV Shows, TV Act...
3                               Docuseries, Reality TV
4     International TV Shows, Romantic TV Shows, TV ...
...                                                 ...
8775                      Cult Movies, Dramas, Thrillers
8776             Kids' TV, Korean TV Shows, TV Comedies
8777                             Comedies, Horror Movies
8778                 Children & Family Movies, Comedies
8779     Dramas, International Movies, Music & Musicals

                         Genre1                    Genre2  \
0                 Documentaries                      None
1       International TV Shows                 TV Dramas
2                Crime TV Shows   International TV Shows
```

```
3                  Docuseries              Reality TV
4       International TV Shows        Romantic TV Shows
...                        ...                     ...
8775              Cult Movies                  Dramas
8776                 Kids' TV          Korean TV Shows
8777                 Comedies            Horror Movies
8778  Children & Family Movies                Comedies
8779                   Dramas     International Movies

                     Genre3
0                      None
1              TV Mysteries
2      TV Action & Adventure
3                      None
4               TV Comedies
...                     ...
8775              Thrillers
8776            TV Comedies
8777                   None
8778                   None
8779        Music & Musicals

[8780 rows x 14 columns]
```

2. Import both the attached files (student.csv and marks.csv) in Jupyter notebook and perform following operations:

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

In [82]:
```
import pandas as pd
df_student = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\student.csv")
df_mark = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\mark.csv")

#================================#
#Funtion & Output:

df_combined = pd.concat([df_student, df_mark], ignore_index=True)
print(df_combined)
```

```
    Student_id   Age  Gender      Grade Employed  Mark     City
0            1  19.0    Male  1st Class      yes   NaN      NaN
1            2  20.0  Female  2nd Class       no   NaN      NaN
2            3  18.0    Male  1st Class       no   NaN      NaN
3            4  21.0  Female  2nd Class       no   NaN      NaN
4            5  19.0    Male  1st Class       no   NaN      NaN
..         ...   ...     ...        ...      ...   ...      ...
456        228   NaN     NaN        NaN      NaN  99.0     Pune
457        229   NaN     NaN        NaN      NaN  70.0  Chennai
458        230   NaN     NaN        NaN      NaN  55.0    Delhi
459        231   NaN     NaN        NaN      NaN  97.0   Mumbai
460        232   NaN     NaN        NaN      NaN  59.0     Pune

[461 rows x 7 columns]
```

In [83]:
```
df_combined.head(200)
```

Out[83]:

	Student_id	Age	Gender	Grade	Employed	Mark	City
0	1	19.0	Male	1st Class	yes	NaN	NaN
1	2	20.0	Female	2nd Class	no	NaN	NaN
2	3	18.0	Male	1st Class	no	NaN	NaN
3	4	21.0	Female	2nd Class	no	NaN	NaN
4	5	19.0	Male	1st Class	no	NaN	NaN
...	...	...	...	...	...	...	...
195	196	19.0	Male	2nd Class	no	NaN	NaN
196	197	21.0	Female	2nd Class	yes	NaN	NaN
197	198	22.0	Male	3rd Class	no	NaN	NaN
198	199	21.0	Female	1st Class	no	NaN	NaN
199	200	20.0	Male	2nd Class	no	NaN	NaN

200 rows × 7 columns

In [84]:
```
df_combined.tail(200)
```

Out[84]:

	Student_id	Age	Gender	Grade	Employed	Mark	City
261	32	NaN	NaN	NaN	NaN	82.0	Mumbai
262	33	NaN	NaN	NaN	NaN	78.0	Pune
263	34	NaN	NaN	NaN	NaN	83.0	Kochi
264	35	NaN	NaN	NaN	NaN	79.0	Gwalior
265	36	NaN	NaN	NaN	NaN	55.0	Bhopal
...	...	...	...	...	...	...	...
456	228	NaN	NaN	NaN	NaN	99.0	Pune
457	229	NaN	NaN	NaN	NaN	70.0	Chennai
458	230	NaN	NaN	NaN	NaN	55.0	Delhi
459	231	NaN	NaN	NaN	NaN	97.0	Mumbai
460	232	NaN	NaN	NaN	NaN	59.0	Pune

200 rows × 7 columns

b) Print the maximum and minimum marks Gender wise.

In [86]:
```
import pandas as pd
df_student = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\student.csv")
df_mark = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\mark.csv")

#================================#
#Funtion:

df_Combined = pd.merge(df_student, df_mark, on='Student_id')

df_Combined.to_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\Combined.csv", index=Fals

gender_wise_marks = df_Combined.groupby('Student_id')['Mark'].agg(['max', 'min'])

#================================#
#Output:

print(gender_wise_marks)
```

```
            max  min
Student_id
1            95   95
2            70   70
3            98   98
4            75   75
5            89   89
...         ...  ...
228          99   99
229          70   70
230          55   55
231          97   97
232          59   59

[229 rows x 2 columns]
```

In [87]:
```
df_Combined
```

Out[87]:

	Student_id	Age	Gender	Grade	Employed	Mark	City
0	1	19	Male	1st Class	yes	95	Chennai
1	2	20	Female	2nd Class	no	70	Delhi
2	3	18	Male	1st Class	no	98	Mumbai
3	4	21	Female	2nd Class	no	75	Pune
4	5	19	Male	1st Class	no	89	Kochi
...	...	...	...	...	...	...	...
224	228	21	Female	1st Class	no	99	Pune
225	229	20	Male	2nd Class	no	70	Chennai
226	230	20	Male	3rd Class	yes	55	Delhi
227	231	19	Female	1st Class	yes	97	Mumbai
228	232	20	Male	3rd Class	yes	59	Pune

229 rows × 7 columns

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

In [89]:
```
import pandas as pd
df_Combined = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\Combined.csv")

#================================#
#Funtion:

Avrg_marks = df_Combined['Mark'].mean()

Above_Avrg_marks = df_Combined[df_Combined['Mark'] > Avrg_marks]

#================================#
#Output:

print("Students with marks above average:")
print(Above_Avrg_marks[['Student_id', 'Mark']])
```

```
Students with marks above average:
     Student_id  Mark
0             1    95
2             3    98
3             4    75
4             5    89
9            12    90
..          ...   ...
218         222    74
219         223    79
221         225    72
224         228    99
227         231    97

[137 rows x 2 columns]
```

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

In [91]:
```
import pandas as pd

df_student = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\student.csv")
df_mark = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\mark.csv")

#================================#
#Funtion & Output:

Male_emp = df_student[(df_student['Gender']=='Male') & (df_student['Employed']=='ye
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 [93]:
```
import pandas as pd
import numpy as np
df_Combined = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\Combined.csv")

#================================#
#Funtion:

df_Combined['IQ_level'] = np.where(df_Combined['Mark'] > 80, 'Intelligent', np.wher

#================================#
#Output:
print(df_Combined['IQ_level'].to_frame())
```

```
       IQ_level
0    Intelligent
1       Mediocre
2    Intelligent
3       Mediocre
4    Intelligent
..           ...
224  Intelligent
225     Mediocre
226     Mediocre
227  Intelligent
228     Mediocre

[229 rows x 1 columns]
```

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

In [95]:
```
import pandas as pd

df_student = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\student.csv")
df_mark = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\mark.csv")
df_Combined = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\Combined.csv")

#================================#
#Funtion & Output:

df_Combined['Gender'].groupby(df_Combined['Gender']).count()
```

Out[95]:
```
Gender
Female     95
Male      134
Name: Gender, dtype: int64
```

g) Print the top 5 Male scorers.

In [137…
```
import pandas as pd
df_Combined = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\Combined.csv")

#================================#
#Funtion & Output:

print(df_Combined.sort_values('Mark', ascending = False).head(5))
```

```
     Student_id  Age  Gender      Grade Employed  Mark     City
146         150   19    Male  1st Class       no   100   Bhopal
224         228   21  Female  1st Class       no    99     Pune
144         148   18    Male  1st Class       no    99    Kochi
59           63   19    Male  1st Class       no    98  Chennai
2             3   18    Male  1st Class       no    98   Mumbai
```

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.

> I do not have seperate compartment for D drive, so i will keep the storage location of data based on availability.

In [100…
```
import pandas as pd
df_Combined = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\Combined.csv")
#================================#
#Input

df_Combined['Gender'] = df_Combined['Gender'].replace({'Male':'M', 'Female':'F'})
#================================#
#Output

print(df_Combined)
#================================#
#Save to Excel
df_Combined = df_Combined.to_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\test.csv")
```

```
     Student_id  Age Gender      Grade Employed  Mark     City
0             1   19      M  1st Class      yes    95  Chennai
1             2   20      F  2nd Class       no    70    Delhi
2             3   18      M  1st Class       no    98   Mumbai
3             4   21      F  2nd Class       no    75     Pune
4             5   19      M  1st Class       no    89    Kochi
..          ...  ...    ...        ...      ...   ...      ...
224         228   21      F  1st Class       no    99     Pune
225         229   20      M  2nd Class       no    70  Chennai
226         230   20      M  3rd Class      yes    55    Delhi
227         231   19      F  1st Class      yes    97   Mumbai
228         232   20      M  3rd Class      yes    59     Pune

[229 rows x 7 columns]
```

In [122…
```
import glob
import pandas as pd
```

```
import os

file_path = r"C:\Users\kaust\OneDrive\Desktop\IITG\*.csv"

for file in glob.glob(file_path):
    print(os.path.basename(file))
```

```
Combined.csv
mark.csv
netflix_titles.csv
student.csv
test.csv
```

i) Check if any student_ID is duplicated.

In [120…

```
import pandas as pd
df_Combined = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\Combined.csv")
#================================#
#Function and Output:
print("="*30)
print("Duplicate Count is:")
print(df_Combined['Student_id'].duplicated().sum())
print("="*30)
print(df_Combined['Student_id'].value_counts())
print("="*30)
```

```
==============================
Duplicate Count is:
0
==============================
Student_id
1      1
148    1
150    1
151    1
152    1
      ..
84     1
85     1
86     1
87     1
232    1
Name: count, Length: 229, dtype: int64
==============================
```

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

In [109…

```
import pandas as pd

df_student = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\student.csv")
df_mark = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\mark.csv")
#================================#
#Input:

df_num = df_student.select_dtypes(include=['int64', 'float64'])
```

```
#================================#

print(df_num)
```

```
     Student_id  Age
0             1   19
1             2   20
2             3   18
3             4   21
4             5   19
..          ...  ...
227         228   21
228         229   20
229         230   20
230         231   19
231         232   20

[232 rows x 2 columns]
```

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

In [112…

```
import pandas as pd

df_student = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\student.csv")
df_mark = pd.read_csv(r"C:\Users\kaust\OneDrive\Desktop\IITG\mark.csv")
#================================#
#Input

missing_students = df_student[~df_student['Student_id'].isin(df_mark['Student_id'])
#================================#
#Output

print(missing_students)
missing_students
```

```
   Student_id  Age  Gender      Grade Employed
8           9   22  Female  3rd Class      yes
9          10   21    Male  1st Class       no
48         49   19    Male  2nd Class       no
```

Out[112…

	Student_id	Age	Gender	Grade	Employed
8	9	22	Female	3rd Class	yes
9	10	21	Male	1st Class	no
48	49	19	Male	2nd Class	no

Q3) Explain the concept of missing values. How can you identify the missing valuesin a Pandas DataFrame ?

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

Missing values, also known as null or NaN (Not a Number) values, are data points that are absent or unknown in a dataset. They can occur due to various reasons such as:

1. Non-response or incomplete surveys
2. Data entry errors
3. Technical issues during data collection
4. Intentional removal of sensitive information

> How to Find Missing Values in a Table (DataFrame)

To find missing values in a table:

1. Use isnull() or isna() to highlight missing values.
2. Check the table's info using info().
3. Look at the table's summary using describe().

> What to Do with Missing Values?

we have three options:

1. Delete: Remove rows or columns with missing values using dropna().
2. Fill: Replace missing values with:
   - A specific number using fillna().
   - The average value using mean().
   - The middle value using median().
   - The most common value using mode().
3. Predict: Use advanced methods to guess the missing values.

Example for above:

In [118…

```
import pandas as pd
import numpy as np
#================================#
# Creating a sample DataFrame with missing values

df = pd.DataFrame({
    'A': [1, 2, np.nan, 4],
    'B': [5, np.nan, 7, 8],
    'C': [np.nan, 10, 11, 12]})

#================================#
# Print the original DataFrame:

print("Original DataFrame:")
print(df)
#================================#
# Deleting rows with missing values:
```

```
df_dropped = df.dropna()
print("\nDataFrame after dropping rows with missing values:")
print(df_dropped)
#================================#
# Imputing missing values with mean:

df_imputed = df.fillna(df.mean())
print("\nDataFrame after imputing missing values with mean:")
print(df_imputed)
```

```
Original DataFrame:
     A    B     C
0  1.0  5.0   NaN
1  2.0  NaN  10.0
2  NaN  7.0  11.0
3  4.0  8.0  12.0

DataFrame after dropping rows with missing values:
     A    B     C
3  4.0  8.0  12.0

DataFrame after imputing missing values with mean:
          A         B     C
0  1.000000  5.000000  11.0
1  2.000000  6.666667  10.0
2  2.333333  7.000000  11.0
3  4.000000  8.000000  12.0
```

In [ ]: