In [1]:

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

Data Frame Operations

(1) Working With Operators On a Series

In [2]:

```
movie = pd.read_csv('movie.csv')
movie
```

Out[2]:

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_lil
0	Color	James Cameron	723.0	178.0	0.0	85
1	Color	Gore Verbinski	302.0	169.0	563.0	100
2	Color	Sam Mendes	602.0	148.0	0.0	16
3	Color	Christopher Nolan	813.0	164.0	22000.0	2300
4	NaN	Doug Walker	NaN	NaN	131.0	N
...	...	...	...	...	...	
4911	Color	Scott Smith	1.0	87.0	2.0	31
4912	Color	NaN	43.0	43.0	NaN	31
4913	Color	Benjamin Roberds	13.0	76.0	0.0	
4914	Color	Daniel Hsia	14.0	100.0	0.0	48
4915	Color	Jon Gunn	43.0	90.0	16.0	1

4916 rows × 28 columns

In [3]:

```
imdb_score = movie['imdb_score']
imdb_score
```

Out[3]:

```
0       7.9
1       7.1
2       6.8
3       8.5
4       7.1
       ...
4911    7.7
```

```
4912    7.5
4913    6.3
4914    6.3
4915    6.6
Name: imdb_score, Length: 4916, dtype: float64
```

In [4]:

```
imdb_score * 2.5
```

Out[4]:

```
0       19.75
1       17.75
2       17.00
3       21.25
4       17.75
        ...  
4911    19.25
4912    18.75
4913    15.75
4914    15.75
4915    16.50
Name: imdb_score, Length: 4916, dtype: float64
```

In [5]:

```
imdb_score // 7
```

Out[5]:

```
0       1.0
1       1.0
2       0.0
3       1.0
4       1.0
       ... 
4911    1.0
4912    1.0
4913    0.0
4914    0.0
4915    0.0
Name: imdb_score, Length: 4916, dtype: float64
```

In [6]:

```
imdb_score > 7
```

Out[6]:

```
0        True
1        True
2       False
3        True
4        True
        ...  
4911     True
4912     True
4913    False
4914    False
4915    False
Name: imdb_score, Length: 4916, dtype: bool
```

In [7]:

```
imdb_score
```

Out[7]:

```
0       7.9
1       7.1
2       6.8
3       8.5
4       7.1
       ... 
4911    7.7
```

```
4912    7.5
4913    6.3
4914    6.3
4915    6.6
Name: imdb_score, Length: 4916, dtype: float64
```

(2) Working With Operators On a Data Frame

In [8]:

```
college = pd.read_csv('college.csv')
college
```

Out[8]:

	INSTNM	CITY	STABBR	HBCU	MENONLY	WOMENONLY	RELAFFIL	SATVRMID	SATM
0	Alabama A & M University	Normal	AL	1.0	0.0	0.0	0	424.0	
1	University of Alabama at Birmingham	Birmingham	AL	0.0	0.0	0.0	0	570.0	
2	Amridge University	Montgomery	AL	0.0	0.0	0.0	1	NaN	
3	University of Alabama in Huntsville	Huntsville	AL	0.0	0.0	0.0	0	595.0	
4	Alabama State University	Montgomery	AL	1.0	0.0	0.0	0	425.0	
...	...	...	...	...	...	...	...	...	
7530	SAE Institute of Technology San Francisco	Emeryville	CA	NaN	NaN	NaN	1	NaN	
7531	Rasmussen College - Overland Park	Overland Park	KS	NaN	NaN	NaN	1	NaN	
7532	National Personal Training Institute of Cleveland	Highland Heights	OH	NaN	NaN	NaN	1	NaN	
7533	Bay Area Medical Academy - San Jose Satellite ...	San Jose	CA	NaN	NaN	NaN	1	NaN	
7534	Excel Learning	San Antonio	TX	NaN	NaN	NaN	1	NaN	

INSTNM	CITY	STABBR	HBCU	MENONLY	WOMENONLY	RELAFFIL	SATVRMID	SATM
Center-San Antonio South								

7535 rows × 27 columns

In [9]:

```
college.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7535 entries, 0 to 7534
Data columns (total 27 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   INSTNM              7535 non-null   object 
 1   CITY                7535 non-null   object 
 2   STABBR              7535 non-null   object 
 3   HBCU                7164 non-null   float64
 4   MENONLY             7164 non-null   float64
 5   WOMENONLY           7164 non-null   float64
 6   RELAFFIL            7535 non-null   int64  
 7   SATVRMID            1185 non-null   float64
 8   SATMTMID            1196 non-null   float64
 9   DISTANCEONLY        7164 non-null   float64
 10  UGDS                6874 non-null   float64
 11  UGDS_WHITE          6874 non-null   float64
 12  UGDS_BLACK          6874 non-null   float64
 13  UGDS_HISP           6874 non-null   float64
 14  UGDS_ASIAN          6874 non-null   float64
 15  UGDS_AIAN           6874 non-null   float64
 16  UGDS_NHPI           6874 non-null   float64
 17  UGDS_2MOR           6874 non-null   float64
 18  UGDS_NRA            6874 non-null   float64
 19  UGDS_UNKN           6874 non-null   float64
 20  PPTUG_EF            6853 non-null   float64
 21  CURROPER            7535 non-null   int64  
 22  PCTPELL             6849 non-null   float64
 23  PCTFLOAN            6849 non-null   float64
 24  UG25ABV             6718 non-null   float64
 25  MD_EARN_WNE_P10     6413 non-null   object 
 26  GRAD_DEBT_MDN_SUPP  7503 non-null   object 
dtypes: float64(20), int64(2), object(5)
memory usage: 1.6+ MB
```

In [10]:

```
college.columns
```

Out[10]:

```
Index(['INSTNM', 'CITY', 'STABBR', 'HBCU', 'MENONLY', 'WOMENONLY', 'RELAFFIL',
       'SATVRMID', 'SATMTMID', 'DISTANCEONLY', 'UGDS', 'UGDS_WHITE',
       'UGDS_BLACK', 'UGDS_HISP', 'UGDS_ASIAN', 'UGDS_AIAN', 'UGDS_NHPI',
       'UGDS_2MOR', 'UGDS_NRA', 'UGDS_UNKN', 'PPTUG_EF', 'CURROPER', 'PCTPELL',
       'PCTFLOAN', 'UG25ABV', 'MD_EARN_WNE_P10', 'GRAD_DEBT_MDN_SUPP'],
      dtype='object')
```

In [11]:

```
college = pd.read_csv('college.csv', index_col = 'INSTNM')
college_ugds_ = college.filter(like = 'UGDS_')
college_ugds_
```

Out[11]:

INSTNM	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2M
Alabama A & M University	0.0333	0.9353	0.0055	0.0019	0.0024	0.0019	0.0
University of Alabama at Birmingham	0.5922	0.2600	0.0283	0.0518	0.0022	0.0007	0.0
Amridge University	0.2990	0.4192	0.0069	0.0034	0.0000	0.0000	0.0
University of Alabama in Huntsville	0.6988	0.1255	0.0382	0.0376	0.0143	0.0002	0.0
Alabama State University	0.0158	0.9208	0.0121	0.0019	0.0010	0.0006	0.0
...	...	...	...	...	...	...	
SAE Institute of Technology San Francisco	NaN	NaN	NaN	NaN	NaN	NaN	
Rasmussen College - Overland Park	NaN	NaN	NaN	NaN	NaN	NaN	
National Personal Training Institute of Cleveland	NaN	NaN	NaN	NaN	NaN	NaN	
Bay Area Medical Academy - San Jose Satellite Location	NaN	NaN	NaN	NaN	NaN	NaN	
Excel Learning Center-San Antonio South	NaN	NaN	NaN	NaN	NaN	NaN	

7535 rows × 9 columns

In [12]:

```
college_ugds_round_per = college_ugds_.round(2) * 100
college_ugds_round_per
```

Out[12]:

INSTNM	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2M
Alabama A & M University	3.0	94.0	1.0	0.0	0.0	0.0	
University of Alabama at Birmingham	59.0	26.0	3.0	5.0	0.0	0.0	
Amridge University	30.0	42.0	1.0	0.0	0.0	0.0	
University of Alabama in Huntsville	70.0	13.0	4.0	4.0	1.0	0.0	
Alabama State University	2.0	92.0	1.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	
SAE Institute of Technology San Francisco	NaN	NaN	NaN	NaN	NaN	NaN	
Rasmussen College - Overland Park	NaN	NaN	NaN	NaN	NaN	NaN	
National Personal Training Institute of Cleveland	NaN	NaN	NaN	NaN	NaN	NaN	
Bay Area Medical Academy - San Jose Satellite Location	NaN	NaN	NaN	NaN	NaN	NaN	
Excel Learning Center-San Antonio South	NaN	NaN	NaN	NaN	NaN	NaN	

7535 rows × 9 columns

In [ ]:

```
college_ugds_round_per.sort_values('UGDS_ASIAN', ascending = False)
```

(3) Count a Number of Values Using value_counts()

In [13]:

```
movie = pd.read_csv('movie.csv')
```

In [14]:

```
director = movie['director_name']
actor_1_fb_likes = movie['actor_1_facebook_likes']
```

In [15]:

```
director
```

Out[15]:

```
0           James Cameron
1          Gore Verbinski
2              Sam Mendes
3       Christopher Nolan
4             Doug Walker
              ...
4911          Scott Smith
4912                  NaN
4913     Benjamin Roberds
4914          Daniel Hsia
4915             Jon Gunn
Name: director_name, Length: 4916, dtype: object
```

In [16]:

```
director.value_counts()
```

Out[16]:

```
Steven Spielberg    26
Woody Allen         22
Martin Scorsese     20
Clint Eastwood      20
Spike Lee           16
                    ..
Thea Sharrock        1
Gary Chapman         1
Fred Savage          1
Robert M. Young      1
Paul Abascal         1
Name: director_name, Length: 2397, dtype: int64
```

In [17]:

```
director.value_counts(normalize = True)
```

Out[17]:

```
Steven Spielberg    0.005401
Woody Allen         0.004570
Martin Scorsese     0.004155
Clint Eastwood      0.004155
Spike Lee           0.003324
                      ...
Thea Sharrock       0.000208
```

```
Gary Chapman        0.000208
Fred Savage         0.000208
Robert M. Young     0.000208
Paul Abascal        0.000208
Name: director_name, Length: 2397, dtype: float64
```

In [18]:

```
actor_1_fb_likes
```

Out[18]:

```
0        1000.0
1       40000.0
2       11000.0
3       27000.0
4         131.0
         ...
4911      637.0
4912      841.0
4913        0.0
4914      946.0
4915       86.0
Name: actor_1_facebook_likes, Length: 4916, dtype: float64
```

In [19]:

```
actor_1_fb_likes.value_counts()
```

Out[19]:

```
1000.0     436
11000.0    206
2000.0     189
3000.0     150
12000.0    131
          ...
564.0        1
46000.0      1
49.0         1
447.0        1
161.0        1
Name: actor_1_facebook_likes, Length: 877, dtype: int64
```

(4) Descriptive Statistics

In [20]:

```
movie = pd.read_csv('movie.csv')
director = movie['director_name']
```

In [21]:

```
director.describe()
```

Out[21]:

```
count                 4814
unique                2397
top       Steven Spielberg
freq                    26
Name: director_name, dtype: object
```

In [22]:

```
actor_1_fb_likes.describe()
```

Out[22]:

```
count     4909.000000
mean      6494.488491
std      15106.986884
min          0.000000
25%        607.000000
```

```
50%         982.000000
75%       11000.000000
max      640000.000000
Name: actor_1_facebook_likes, dtype: float64
```

In [23]:

```
actor_1_fb_likes.min()
```

Out[23]:

```
0.0
```

In [24]:

```
actor_1_fb_likes.max()
```

Out[24]:

```
640000.0
```

In [25]:

```
actor_1_fb_likes.mean()
```

Out[25]:

```
6494.488490527602
```

In [26]:

```
actor_1_fb_likes.std()
```

Out[26]:

```
15106.986883848309
```

In [27]:

```
actor_1_fb_likes.median()
```

Out[27]:

```
982.0
```

In [28]:

```
actor_1_fb_likes.sum()
```

Out[28]:

```
31881444.0
```

In [29]:

```
actor_1_fb_likes.quantile()
```

Out[29]:

```
982.0
```

In [30]:

```
actor_1_fb_likes.quantile(0.5)
```

Out[30]:

```
982.0
```

In [31]:

```
actor_1_fb_likes.quantile(0.2)
```

Out[31]:

```
510.0
```

In [32]:

```
actor_1_fb_likes.quantile([0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9])
```

Out[32]:

```
0.1      240.0
0.2      510.0
0.3      694.0
0.4      854.0
```

```
0.5      982.0
0.6     1000.0
0.7     8000.0
0.8    13000.0
0.9    18000.0
Name: actor_1_facebook_likes, dtype: float64
```

In [34]:

```
movie.describe()
```

Out[34]:

	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_1_facebo
count	4867.000000	4901.000000	4814.000000	4893.000000	490
mean	137.988905	107.090798	691.014541	631.276313	649
std	120.239379	25.286015	2832.954125	1625.874802	1510
min	1.000000	7.000000	0.000000	0.000000	(
25%	49.000000	93.000000	7.000000	132.000000	60
50%	108.000000	103.000000	48.000000	366.000000	98
75%	191.000000	118.000000	189.750000	633.000000	1100
max	813.000000	511.000000	23000.000000	23000.000000	64000

In [35]:

```
movie.min()
```

Out[35]:

```
num_critic_for_reviews                                                 1.0
duration                                                               7.0
director_facebook_likes                                                0.0
actor_3_facebook_likes                                                 0.0
actor_1_facebook_likes                                                 0.0
gross                                                                162.0
genres                                                              Action
movie_title                                                        #Horror
num_voted_users                                                          5
cast_total_facebook_likes                                                0
facenumber_in_poster                                                   0.0
movie_imdb_link            http://www.imdb.com/title/tt0006864/?ref_=fn_t...
num_user_for_reviews                                                   1.0
budget                                                               218.0
title_year                                                          1916.0
actor_2_facebook_likes                                                 0.0
imdb_score                                                             1.6
aspect_ratio                                                          1.18
movie_facebook_likes                                                     0
dtype: object
```

(5) Handling Null Values

In [36]:

```
movie = pd.read_csv('movie.csv')
director = movie['director_name']
```

In [37]:

```
director
```

Out[37]:
```
0           James Cameron
1          Gore Verbinski
2              Sam Mendes
3       Christopher Nolan
4             Doug Walker
              ...        
4911          Scott Smith
4912                  NaN
4913     Benjamin Roberds
4914          Daniel Hsia
4915             Jon Gunn
Name: director_name, Length: 4916, dtype: object
```

In [38]:
```
director.hasnans
```

Out[38]:
```
True
```

In [39]:
```
director.isnull()
# director.notnull()
```

Out[39]:
```
0       False
1       False
2       False
3       False
4       False
        ...  
4911    False
4912     True
4913    False
4914    False
4915    False
Name: director_name, Length: 4916, dtype: bool
```

In [40]:
```
director.isnull().any()
```

Out[40]:
```
True
```

In [41]:
```
director.isnull().sum()
```

Out[41]:
```
102
```

In [42]:
```
actor_1_fb_likes.isnull()
```

Out[42]:
```
0       False
1       False
2       False
3       False
4       False
        ...  
4911    False
4912    False
4913    False
4914    False
4915    False
Name: actor_1_facebook_likes, Length: 4916, dtype: bool
```

In [43]:

```
actor_1_fb_likes.isnull().sum()
```

Out[43]:

```
7
```

In [44]:

```
actor_1_fb_likes_filled = actor_1_fb_likes.fillna(0)
len(actor_1_fb_likes_filled)
# actor_1_fb_likes_filled.size
# actor_1_fb_likes_filled.count()
```

Out[44]:

```
4916
```

In [45]:

```
actor_1_fb_likes_dropped = actor_1_fb_likes.dropna()
len(actor_1_fb_likes_dropped)
# actor_1_fb_likes_dropped.size
# actor_1_fb_likes_dropped.count()
```

Out[45]:

```
4909
```

(6) Transposing the Direction of a Data Frame Operation

In [46]:

```
college = pd.read_csv('college.csv', index_col = 'INSTNM')
college_ugds_ = college.filter(like = 'UGDS_')
college_ugds_
```

Out[46]:

	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2N
INSTNM							
Alabama A & M University	0.0333	0.9353	0.0055	0.0019	0.0024	0.0019	0.0
University of Alabama at Birmingham	0.5922	0.2600	0.0283	0.0518	0.0022	0.0007	0.0
Amridge University	0.2990	0.4192	0.0069	0.0034	0.0000	0.0000	0.0
University of Alabama in Huntsville	0.6988	0.1255	0.0382	0.0376	0.0143	0.0002	0.0
Alabama State University	0.0158	0.9208	0.0121	0.0019	0.0010	0.0006	0.0
...	...	...	...	...	...	...	
SAE Institute of	NaN	NaN	NaN	NaN	NaN	NaN	

	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2M
INSTNM							
Technology San Francisco							
Rasmussen College - Overland Park	NaN	NaN	NaN	NaN	NaN	NaN	
National Personal Training Institute of Cleveland	NaN	NaN	NaN	NaN	NaN	NaN	
Bay Area Medical Academy - San Jose Satellite Location	NaN	NaN	NaN	NaN	NaN	NaN	
Excel Learning Center-San Antonio South	NaN	NaN	NaN	NaN	NaN	NaN	

7535 rows × 9 columns

In [47]:

```
college_ugds_.count()
```

Out[47]:

```
UGDS_WHITE    6874
UGDS_BLACK    6874
UGDS_HISP     6874
UGDS_ASIAN    6874
UGDS_AIAN     6874
UGDS_NHPI     6874
UGDS_2MOR     6874
UGDS_NRA      6874
UGDS_UNKN     6874
dtype: int64
```

In [48]:

```
college_ugds_.count(axis = 'index')
```

Out[48]:

```
UGDS_WHITE    6874
UGDS_BLACK    6874
UGDS_HISP     6874
UGDS_ASIAN    6874
UGDS_AIAN     6874
UGDS_NHPI     6874
UGDS_2MOR     6874
UGDS_NRA      6874
UGDS_UNKN     6874
dtype: int64
```

In [49]:

```
college_ugds_.count(axis = 'columns')
```

Out[49]:

```
INSTNM
Alabama A & M University                                  9
University of Alabama at Birmingham                       9
Amridge University                                        9
University of Alabama in Huntsville                       9
Alabama State University                                  9
                                                         ..
SAE Institute of Technology  San Francisco                0
Rasmussen College - Overland Park                         0
National Personal Training Institute of Cleveland         0
Bay Area Medical Academy - San Jose Satellite Location    0
Excel Learning Center-San Antonio South                   0
Length: 7535, dtype: int64
```

In [50]:

```
college_ugds_.count(axis = 'columns').sort_values()
```

Out[50]:

```
INSTNM
Excel Learning Center-San Antonio South                     0
Albany Law School                                           0
Albany Medical College                                      0
Institute for the Psychological Sciences                    0
Forest Institute of Professional Psychology                 0
                                                           ..
Farmingdale State College                                   9
SUNY College of Agriculture and Technology at Cobleskill    9
SUNY College of Technology at Delhi                         9
SUNY College of Technology at Alfred                        9
The University of Texas at Austin                           9
Length: 7535, dtype: int64
```

In [51]:

```
college_ugds_ = college_ugds_.dropna(how = 'all')
```

In [52]:

```
college_ugds_.count(axis = 'columns').sort_values()
```

Out[52]:

```
INSTNM
Alabama A & M University                        9
Pike County Joint Vocational School District    9
South Texas College                             9
Professional Technical Institution Inc          9
Franklin Technology-MSSU                        9
                                               ..
CUNY Graduate School and University Center      9
CUNY City College                               9
College of Staten Island CUNY                   9
CUNY Bronx Community College                    9
Coastal Pines Technical College                 9
Length: 6874, dtype: int64
```

In [53]:

```
college_ugds_.count()
```

Out[53]:

```
UGDS_WHITE    6874
UGDS_BLACK    6874
UGDS_HISP     6874
UGDS_ASIAN    6874
UGDS_AIAN     6874
```

```
UGDS_NHPI     6874
UGDS_2MOR     6874
UGDS_NRA      6874
UGDS_UNKN     6874
dtype: int64
```