pandas-lecture-3-dec-batch

June 8, 2023

0.1 Pandas-Lecture - 3

```
[1]: import pandas as pd
     import numpy as np
```

```
[2]: users = pd.DataFrame({'userid': [1, 2, 3], 'name': ['infini', 'kiran',␣
       ↪'sayed']})
     users
```

```
[2]:    userid    name
     0       1  infini
     1       2   kiran
     2       3   sayed
```

```
[3]: msgs = pd.DataFrame({'userid': [1, 1, 2, 4], 'msg': ['hi', 'how are you?',␣
       ↪'bye', 'nice']})
     msgs
```

```
[3]:    userid           msg
     0       1            hi
     1       1  how are you?
     2       2           bye
     3       4          nice
```

```
[4]: pd.concat([users, msgs])
```

```
[4]:    userid    name           msg
     0       1  infini           NaN
     1       2   kiran           NaN
     2       3   sayed           NaN
     0       1     NaN            hi
     1       1     NaN  how are you?
     2       2     NaN           bye
     3       4     NaN          nice
```

```
[5]: pd.concat([users, msgs], ignore_index=True)
```

[5]:
```
   userid    name           msg
0       1  infini           NaN
1       2   kiran           NaN
2       3   sayed           NaN
3       1     NaN            hi
4       1     NaN  how are you?
5       2     NaN           bye
6       4     NaN          nice
```

[6]:
```
pd.concat([users, msgs], axis=1)
```

[6]:
```
   userid    name  userid           msg
0     1.0  infini       1            hi
1     2.0   kiran       1  how are you?
2     3.0   sayed       2           bye
3     NaN     NaN       4          nice
```

[7]:
```
users.merge(msgs, on='userid')
```

[7]:
```
   userid    name           msg
0       1  infini            hi
1       1  infini  how are you?
2       2   kiran           bye
```

[8]:
```
users.merge(msgs, on='userid',  how='outer')
```

[8]:
```
   userid    name           msg
0       1  infini            hi
1       1  infini  how are you?
2       2   kiran           bye
3       3   sayed           NaN
4       4     NaN          nice
```

[9]:
```
users.merge(msgs, on='userid',  how='left')
```

[9]:
```
   userid    name           msg
0       1  infini            hi
1       1  infini  how are you?
2       2   kiran           bye
3       3   sayed           NaN
```

[10]:
```
users.merge(msgs, on='userid',  how='right')
```

[10]:
```
   userid    name           msg
0       1  infini            hi
1       1  infini  how are you?
2       2   kiran           bye
```

```
3       4     NaN          nice
```

```
[11]: users.rename(columns={'userid': 'id'}, inplace=True)
```

```
[12]: users
```

```
[12]:    id    name
      0   1  infini
      1   2   kiran
      2   3   sayed
```

```
[13]: users.merge(msgs, left_on='id', right_on='userid')
```

```
[13]:    id    name  userid          msg
      0   1  infini       1           hi
      1   1  infini       1  how are you?
      2   2   kiran       2          bye
```

0.1.1 IMDB Movie Usecase

```
[14]: !gdown 1s2TkjSpzNc4SyxqRrQleZyDIHlc7bxnd
```

```
Downloading…
From: https://drive.google.com/uc?id=1s2TkjSpzNc4SyxqRrQleZyDIHlc7bxnd
To: /Users/satish/Desktop/scaler/Dec Tue Batch - DAV-1/movies.csv
100%|                     | 112k/112k [00:00<00:00, 1.59MB/s]
```

```
[15]: !gdown 1Ws-_s1fHZ9nHfGLVUQurbHDvStePlEJm
```

```
Downloading…
From: https://drive.google.com/uc?id=1Ws-_s1fHZ9nHfGLVUQurbHDvStePlEJm
To: /Users/satish/Desktop/scaler/Dec Tue Batch - DAV-1/directors.csv
100%|                    | 65.4k/65.4k [00:00<00:00, 1.21MB/s]
```

```
[16]: movies = pd.read_csv('movies.csv')
      movies
```

```
[16]:       Unnamed: 0     id     budget  popularity     revenue  \
      0              0  43597  237000000         150  2787965087
      1              1  43598  300000000         139   961000000
      2              2  43599  245000000         107   880674609
      3              3  43600  250000000         112  1084939099
      4              5  43602  258000000         115   890871626
      …            …      …          …           …           …
      1460        4736  48363          0           3      321952
      1461        4743  48370      27000          19     3151130
      1462        4748  48375          0           7           0
```

```
1463        4749  48376          0           3           0
1464        4768  48395     220000          14     2040920

                                         title  vote_average  vote_count  \
0                                       Avatar           7.2       11800
1     Pirates of the Caribbean: At World's End           6.9        4500
2                                      Spectre           6.3        4466
3                        The Dark Knight Rises           7.6        9106
4                                 Spider-Man 3           5.9        3576
…                                        …           …         …
1460                            The Last Waltz           7.9          64
1461                                    Clerks           7.4         755
1462                                   Rampage           6.0         131
1463                                   Slacker           6.4          77
1464                               El Mariachi           6.6         238

      director_id  year month       day
0            4762  2009   Dec  Thursday
1            4763  2007   May  Saturday
2            4764  2015   Oct    Monday
3            4765  2012   Jul    Monday
4            4767  2007   May   Tuesday
…           …   …   …       …
1460         4809  1978   May    Monday
1461         5369  1994   Sep   Tuesday
1462         5148  2009   Aug    Friday
1463         5535  1990   Jul    Friday
1464         5097  1992   Sep    Friday

[1465 rows x 12 columns]
```

```
[17]: movies.drop('Unnamed: 0', axis=1, inplace=True)
```

```
[18]: movies
```

```
[18]:          id     budget  popularity     revenue  \
0       43597  237000000         150  2787965087
1       43598  300000000         139   961000000
2       43599  245000000         107   880674609
3       43600  250000000         112  1084939099
4       43602  258000000         115   890871626
…     …        …         …         …
1460    48363          0           3      321952
1461    48370      27000          19     3151130
1462    48375          0           7           0
1463    48376          0           3           0
1464    48395     220000          14     2040920
```

```
                                         title  vote_average  vote_count  \
0                                       Avatar           7.2       11800
1     Pirates of the Caribbean: At World's End           6.9        4500
2                                      Spectre           6.3        4466
3                        The Dark Knight Rises           7.6        9106
4                                 Spider-Man 3           5.9        3576
…                                        …           …         …
1460                            The Last Waltz           7.9          64
1461                                    Clerks           7.4         755
1462                                   Rampage           6.0         131
1463                                   Slacker           6.4          77
1464                               El Mariachi           6.6         238

      director_id  year month       day
0            4762  2009   Dec  Thursday
1            4763  2007   May  Saturday
2            4764  2015   Oct    Monday
3            4765  2012   Jul    Monday
4            4767  2007   May   Tuesday
…           …   …   …       …
1460         4809  1978   May    Monday
1461         5369  1994   Sep   Tuesday
1462         5148  2009   Aug    Friday
1463         5535  1990   Jul    Friday
1464         5097  1992   Sep    Friday

[1465 rows x 11 columns]
```

[19]:

```python
movies.shape
```

[19]:

```
(1465, 11)
```

[23]:

```python
directors = pd.read_csv('directors.csv', index_col=0)
directors
```

[23]:

```
          director_name    id gender
0         James Cameron  4762   Male
1        Gore Verbinski  4763   Male
2            Sam Mendes  4764   Male
3     Christopher Nolan  4765   Male
4        Andrew Stanton  4766   Male
…                 …   …    …
2344      Shane Carruth  7106   Male
2345   Neill Dela Llana  7107    NaN
2346        Scott Smith  7108    NaN
2347        Daniel Hsia  7109   Male
```

```
2348   Brian Herzlinger  7110   Male

[2349 rows x 3 columns]
```

```
[ ]:
```

```
[24]: directors.shape
```

```
[24]: (2349, 3)
```

```
[25]: movies.head()
```

```
[25]:       id     budget  popularity     revenue  \
      0  43597  237000000         150  2787965087
      1  43598  300000000         139   961000000
      2  43599  245000000         107   880674609
      3  43600  250000000         112  1084939099
      4  43602  258000000         115   890871626

                                           title  vote_average  vote_count  \
      0                                   Avatar           7.2       11800
      1  Pirates of the Caribbean: At World's End           6.9        4500
      2                                  Spectre           6.3        4466
      3                    The Dark Knight Rises           7.6        9106
      4                             Spider-Man 3           5.9        3576

         director_id  year month       day
      0         4762  2009   Dec  Thursday
      1         4763  2007   May  Saturday
      2         4764  2015   Oct    Monday
      3         4765  2012   Jul    Monday
      4         4767  2007   May   Tuesday
```

```
[26]: directors.head()
```

```
[26]:       director_name    id gender
      0      James Cameron  4762   Male
      1     Gore Verbinski  4763   Male
      2         Sam Mendes  4764   Male
      3  Christopher Nolan  4765   Male
      4     Andrew Stanton  4766   Male
```

```
[28]: movies['director_id'].nunique()
```

```
[28]: 199
```

```
[29]: directors['id'].nunique()
```

[29]:
```
2349
```

[30]:
```
movies['director_id'].isin(directors['id'])
```

[30]:
```
0       True
1       True
2       True
3       True
4       True
        …
1460    True
1461    True
1462    True
1463    True
1464    True
Name: director_id, Length: 1465, dtype: bool
```

[31]:
```
np.all(movies['director_id'].isin(directors['id']))
```

[31]:
```
True
```

[32]:
```
### Do we need to keep all rows from movies data frame [merge the dataframes]
```

[33]:
```
data = movies.merge(directors, left_on='director_id', right_on='id', how='left')
data
```

[33]:
```
       id_x     budget  popularity     revenue  \
0     43597  237000000         150  2787965087
1     43598  300000000         139   961000000
2     43599  245000000         107   880674609
3     43600  250000000         112  1084939099
4     43602  258000000         115   890871626
…       …          …          …           …
1460  48363          0           3      321952
1461  48370      27000          19     3151130
1462  48375          0           7           0
1463  48376          0           3           0
1464  48395     220000          14     2040920

                                         title  vote_average  vote_count  \
0                                       Avatar           7.2       11800
1     Pirates of the Caribbean: At World's End           6.9        4500
2                                      Spectre           6.3        4466
3                        The Dark Knight Rises           7.6        9106
4                                 Spider-Man 3           5.9        3576
…                                          …             …           …
1460                           The Last Waltz           7.9          64
```

```
1461                                    Clerks           7.4          755
1462                                   Rampage           6.0          131
1463                                   Slacker           6.4           77
1464                                El Mariachi           6.6          238

      director_id  year month       day      director_name  id_y gender
0            4762  2009   Dec  Thursday      James Cameron  4762    Male
1            4763  2007   May  Saturday     Gore Verbinski  4763    Male
2            4764  2015   Oct    Monday         Sam Mendes  4764    Male
3            4765  2012   Jul    Monday  Christopher Nolan  4765    Male
4            4767  2007   May   Tuesday          Sam Raimi  4767    Male
…               …     …   …         …                  …     …     …
1460         4809  1978   May    Monday    Martin Scorsese  4809    Male
1461         5369  1994   Sep   Tuesday        Kevin Smith  5369    Male
1462         5148  2009   Aug    Friday           Uwe Boll  5148    Male
1463         5535  1990   Jul    Friday  Richard Linklater  5535    Male
1464         5097  1992   Sep    Friday   Robert Rodriguez  5097     NaN

[1465 rows x 14 columns]
```

```
[34]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1465 entries, 0 to 1464
Data columns (total 14 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id_x           1465 non-null   int64
 1   budget         1465 non-null   int64
 2   popularity     1465 non-null   int64
 3   revenue        1465 non-null   int64
 4   title          1465 non-null   object
 5   vote_average   1465 non-null   float64
 6   vote_count     1465 non-null   int64
 7   director_id    1465 non-null   int64
 8   year           1465 non-null   int64
 9   month          1465 non-null   object
 10  day            1465 non-null   object
 11  director_name  1465 non-null   object
 12  id_y           1465 non-null   int64
 13  gender         1341 non-null   object
dtypes: float64(1), int64(8), object(5)
memory usage: 171.7+ KB
```

```
[35]: data.drop(['director_id', 'id_y'], axis=1, inplace=True)
      data
```

```
[35]:        id_x     budget  popularity     revenue  \
      0     43597  237000000         150  2787965087
      1     43598  300000000         139   961000000
      2     43599  245000000         107   880674609
      3     43600  250000000         112  1084939099
      4     43602  258000000         115   890871626
      …       …        …           …          …
      1460  48363          0           3      321952
      1461  48370      27000          19     3151130
      1462  48375          0           7           0
      1463  48376          0           3           0
      1464  48395     220000          14     2040920

                                             title  vote_average  vote_count  \
      0                                     Avatar           7.2       11800
      1     Pirates of the Caribbean: At World's End           6.9        4500
      2                                    Spectre           6.3        4466
      3                      The Dark Knight Rises           7.6        9106
      4                               Spider-Man 3           5.9        3576
      …                                          …           …          …
      1460                          The Last Waltz           7.9          64
      1461                                  Clerks           7.4         755
      1462                                 Rampage           6.0         131
      1463                                 Slacker           6.4          77
      1464                              El Mariachi           6.6         238

            year month       day       director_name gender
      0     2009   Dec  Thursday       James Cameron   Male
      1     2007   May  Saturday      Gore Verbinski   Male
      2     2015   Oct    Monday          Sam Mendes   Male
      3     2012   Jul    Monday   Christopher Nolan   Male
      4     2007   May   Tuesday           Sam Raimi   Male
      …      …    …        …               …    …
      1460  1978   May    Monday     Martin Scorsese   Male
      1461  1994   Sep   Tuesday         Kevin Smith   Male
      1462  2009   Aug    Friday            Uwe Boll   Male
      1463  1990   Jul    Friday   Richard Linklater   Male
      1464  1992   Sep    Friday    Robert Rodriguez    NaN

      [1465 rows x 12 columns]
```

```
[36]: data.describe()
```

```
[36]:                id_x        budget   popularity       revenue  vote_average  \
      count   1465.000000  1.465000e+03  1465.000000  1.465000e+03   1465.000000
      mean   45225.191126  4.802295e+07    30.855973  1.432539e+08      6.368191
      std     1189.096396  4.935541e+07    34.845214  2.064918e+08      0.818033
```

```
min    43597.000000  0.000000e+00     0.000000  0.000000e+00      3.000000
25%    44236.000000  1.400000e+07    11.000000  1.738013e+07      5.900000
50%    45022.000000  3.300000e+07    23.000000  7.578164e+07      6.400000
75%    45990.000000  6.600000e+07    41.000000  1.792469e+08      6.900000
max    48395.000000  3.800000e+08   724.000000  2.787965e+09      8.300000

         vote_count         year
count   1465.000000  1465.000000
mean    1146.396587  2002.615017
std     1578.077438     8.680141
min        1.000000  1976.000000
25%      216.000000  1998.000000
50%      571.000000  2004.000000
75%     1387.000000  2009.000000
max    13752.000000  2016.000000
```

```
[37]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1465 entries, 0 to 1464
Data columns (total 12 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id_x           1465 non-null   int64
 1   budget         1465 non-null   int64
 2   popularity     1465 non-null   int64
 3   revenue        1465 non-null   int64
 4   title          1465 non-null   object
 5   vote_average   1465 non-null   float64
 6   vote_count     1465 non-null   int64
 7   year           1465 non-null   int64
 8   month          1465 non-null   object
 9   day            1465 non-null   object
 10  director_name  1465 non-null   object
 11  gender         1341 non-null   object
dtypes: float64(1), int64(6), object(5)
memory usage: 148.8+ KB
```

```
[38]: data.describe(include=object)
```

```
[38]:          title month     day     director_name gender
       count      1465  1465    1465              1465   1341
       unique     1465    12       7               199      2
       top      Avatar   Dec  Friday  Steven Spielberg   Male
       freq          1   193     654                26   1309
```

[39]:
```
data['budget'] = data['budget']/1000000
data
```

[39]:
```
        id_x   budget  popularity     revenue  \
0      43597  237.000         150  2787965087
1      43598  300.000         139   961000000
2      43599  245.000         107   880674609
3      43600  250.000         112  1084939099
4      43602  258.000         115   890871626
…        …      …          …           …
1460   48363    0.000           3      321952
1461   48370    0.027          19     3151130
1462   48375    0.000           7           0
1463   48376    0.000           3           0
1464   48395    0.220          14     2040920

                                         title  vote_average  vote_count  \
0                                       Avatar           7.2       11800
1     Pirates of the Caribbean: At World's End           6.9        4500
2                                      Spectre           6.3        4466
3                        The Dark Knight Rises           7.6        9106
4                                 Spider-Man 3           5.9        3576
…                                          …             …           …
1460                            The Last Waltz           7.9          64
1461                                    Clerks           7.4         755
1462                                   Rampage           6.0         131
1463                                   Slacker           6.4          77
1464                                El Mariachi           6.6         238

      year month       day      director_name gender
0     2009   Dec  Thursday      James Cameron   Male
1     2007   May  Saturday     Gore Verbinski   Male
2     2015   Oct    Monday         Sam Mendes   Male
3     2012   Jul    Monday  Christopher Nolan   Male
4     2007   May   Tuesday          Sam Raimi   Male
…       …    …         …                 …    …
1460  1978   May    Monday    Martin Scorsese   Male
1461  1994   Sep   Tuesday        Kevin Smith   Male
1462  2009   Aug    Friday           Uwe Boll   Male
1463  1990   Jul    Friday  Richard Linklater   Male
1464  1992   Sep    Friday    Robert Rodriguez    NaN

[1465 rows x 12 columns]
```

[40]:
```
## 1. All highly rated movies  [ratings>7]
```

[41]:
```
data['vote_average']>7
```

[41]:

```
0        True
1       False
2       False
3        True
4       False
        …
1460     True
1461     True
1462    False
1463    False
1464    False
Name: vote_average, Length: 1465, dtype: bool
```

[42]:

```
data.loc[data['vote_average']>7]
```

[42]:

```
       id_x   budget  popularity     revenue  \
0     43597  237.000         150  2787965087
3     43600  250.000         112  1084939099
14    43616  250.000         120   956019788
16    43619  250.000          94   958400000
19    43622  200.000         100  1845034188
…     …      …         …         …
1456  48321    0.010          20     7000000
1457  48323    0.000           5           0
1458  48335    0.060          27     3221152
1460  48363    0.000           3      321952
1461  48370    0.027          19     3151130

                                          title  vote_average  vote_count  \
0                                        Avatar           7.2       11800
3                         The Dark Knight Rises           7.6        9106
14    The Hobbit: The Battle of the Five Armies           7.1        4760
16          The Hobbit: The Desolation of Smaug           7.6        4524
19                                      Titanic           7.5        7562
…                                           …            …          …
1456                                 Eraserhead           7.5         485
1457                                 The Mighty           7.1          51
1458                                         Pi           7.1         586
1460                             The Last Waltz           7.9          64
1461                                     Clerks           7.4         755

      year month        day      director_name gender
0     2009   Dec   Thursday      James Cameron   Male
3     2012   Jul     Monday  Christopher Nolan   Male
14    2014   Dec  Wednesday      Peter Jackson   Male
16    2013   Dec  Wednesday      Peter Jackson   Male
19    1997   Nov    Tuesday      James Cameron   Male
```

```
…    …   …        …                …    …
1456  1977   Mar   Saturday        David Lynch   Male
1457  1998   Oct     Friday      Peter Chelsom   Male
1458  1998   Jul     Friday   Darren Aronofsky   Male
1460  1978   May     Monday    Martin Scorsese   Male
1461  1994   Sep    Tuesday        Kevin Smith   Male

[301 rows x 12 columns]
```

[43]:

```python
data[data['vote_average']>7]
```

[43]:

```
       id_x   budget  popularity     revenue   \
0     43597  237.000         150  2787965087
3     43600  250.000         112  1084939099
14    43616  250.000         120   956019788
16    43619  250.000          94   958400000
19    43622  200.000         100  1845034188
…     …      …         …         …
1456  48321    0.010          20     7000000
1457  48323    0.000           5           0
1458  48335    0.060          27     3221152
1460  48363    0.000           3      321952
1461  48370    0.027          19     3151130

                                          title  vote_average  vote_count  \
0                                        Avatar           7.2       11800
3                         The Dark Knight Rises           7.6        9106
14    The Hobbit: The Battle of the Five Armies           7.1        4760
16          The Hobbit: The Desolation of Smaug           7.6        4524
19                                      Titanic           7.5        7562
…                                          …           …         …
1456                                 Eraserhead           7.5         485
1457                                 The Mighty           7.1          51
1458                                         Pi           7.1         586
1460                             The Last Waltz           7.9          64
1461                                     Clerks           7.4         755

      year month        day      director_name gender
0     2009   Dec   Thursday      James Cameron   Male
3     2012   Jul     Monday  Christopher Nolan   Male
14    2014   Dec  Wednesday      Peter Jackson   Male
16    2013   Dec  Wednesday      Peter Jackson   Male
19    1997   Nov    Tuesday      James Cameron   Male
…    …   …        …                …    …
1456  1977   Mar   Saturday        David Lynch   Male
1457  1998   Oct     Friday      Peter Chelsom   Male
1458  1998   Jul     Friday   Darren Aronofsky   Male
```

```
1460  1978   May     Monday    Martin Scorsese   Male
1461  1994   Sep    Tuesday        Kevin Smith   Male

[301 rows x 12 columns]
```

[44]:
```
data.loc[data['vote_average']>7, ['title', 'director_name']]
```

[44]:
```
                                        title       director_name
0                                      Avatar       James Cameron
3                       The Dark Knight Rises   Christopher Nolan
14    The Hobbit: The Battle of the Five Armies       Peter Jackson
16          The Hobbit: The Desolation of Smaug       Peter Jackson
19                                    Titanic       James Cameron
…                                         …                 …
1456                               Eraserhead         David Lynch
1457                               The Mighty       Peter Chelsom
1458                                       Pi    Darren Aronofsky
1460                           The Last Waltz     Martin Scorsese
1461                                   Clerks         Kevin Smith

[301 rows x 2 columns]
```

[45]:
```
###Highly rated movies released after 2014
```

[46]:
```
data.loc[(data['vote_average']>7) & (data['year']>2014)]
```

[46]:
```
       id_x  budget  popularity     revenue                    title  \
30    43641   190.0         102  1506249360                Furious 7
78    43724   150.0         434   378858340       Mad Max: Fury Road
106   43773   135.0         100   532950503             The Revenant
162   43867   108.0         167   630161890              The Martian
312   44128    75.0          48   108145109  The Man from U.N.C.L.E.
394   44281    44.0          68   155760117        The Hateful Eight
625   44770    35.0          53   194564672               The Intern
635   44784    40.0          48   165478348          Bridge of Spies
808   45194    30.0          65    91709827                Southpaw
833   45293    28.0          61   201634991   Straight Outta Compton
839   45301    28.0          57   133346506            The Big Short
1344  47181     5.0          22    24804129                     Race

      vote_average  vote_count  year month        day  \
30             7.3        4176  2015   Apr  Wednesday
78             7.2        9427  2015   May  Wednesday
106            7.3        6396  2015   Dec     Friday
162            7.6        7268  2015   Sep  Wednesday
312            7.1        2265  2015   Aug   Thursday
394            7.6        4274  2015   Dec     Friday
```

```
625            7.1        1881  2015   Sep   Thursday
635            7.2        2583  2015   Oct   Thursday
808            7.3        2067  2015   Jun     Monday
833            7.7        1355  2015   Aug   Thursday
839            7.3        2607  2015   Dec     Friday
1344           7.1         478  2016   Feb     Friday

                    director_name  gender
30                      James Wan    Male
78                  George Miller    Male
106   Alejandro González Iñárritu    Male
162                  Ridley Scott    Male
312                   Guy Ritchie    Male
394             Quentin Tarantino    Male
625                  Nancy Meyers  Female
635              Steven Spielberg    Male
808                 Antoine Fuqua    Male
833                  F. Gary Gray    Male
839                   Adam McKay    Male
1344              Stephen Hopkins    Male
```

[47]:

```
#find movies release on either Friday or Sunday
```

[48]:

```
data.loc[(data['day'] =='Friday') | (data['day']=='Sunday')]
```

[48]:

```
       id_x  budget  popularity    revenue  \
22    43627  200.00          35  783766341
25    43632  150.00          21  836297228
53    43672  175.00          44  264218220
61    43696   38.00           6  207283925
65    43701  160.00          21  181674817
…       …      …           …         …
1458  48335    0.06          27    3221152
1459  48359    0.00           2          0
1462  48375    0.00           7          0
1463  48376    0.00           3          0
1464  48395    0.22          14    2040920

                                   title  vote_average  vote_count  year  \
22                          Spider-Man 2           6.7        4321  2004
25    Transformers: Revenge of the Fallen           6.0        3138  2009
53                            Waterworld           5.9         992  1995
61                 The Fast and the Furious           6.6        3428  2001
65                              Poseidon           5.5         583  2006
…                                      …             …           …     …
1458                                   Pi           7.1         586  1998
1459                    George Washington           6.4          36  2000
```

```
1462                              Rampage           6.0         131  2009
1463                              Slacker           6.4          77  1990
1464                          El Mariachi           6.6         238  1992

     month     day       director_name gender
22     Jun  Friday           Sam Raimi   Male
25     Jun  Friday         Michael Bay   Male
53     Jul  Friday      Kevin Reynolds    NaN
61     Jun  Friday           Rob Cohen   Male
65     May  Friday   Wolfgang Petersen   Male
…    …     …                 …    …
1458   Jul  Friday    Darren Aronofsky   Male
1459   Oct  Sunday  David Gordon Green   Male
1462   Aug  Friday            Uwe Boll   Male
1463   Jul  Friday   Richard Linklater   Male
1464   Sep  Friday    Robert Rodriguez    NaN

[700 rows x 12 columns]
```

[49]:
```
#Top 5 popular movies
```

[50]:
```
data.sort_values(['popularity'], ascending=False).head()
```

[50]:
```
      id_x  budget  popularity     revenue  \
58   43692   165.0         724   675120017
78   43724   150.0         434   378858340
119  43796   140.0         271   655011224
120  43797   125.0         206   752100229
45   43662   185.0         187  1004558444

                                                 title  vote_average  \
58                                        Interstellar           8.1
78                                  Mad Max: Fury Road           7.2
119  Pirates of the Caribbean: The Curse of the Bla…           7.5
120              The Hunger Games: Mockingjay - Part 1           6.6
45                                     The Dark Knight           8.2

     vote_count  year month        day      director_name gender
58        10867  2014   Nov  Wednesday  Christopher Nolan   Male
78         9427  2015   May  Wednesday      George Miller   Male
119        6985  2003   Jul  Wednesday     Gore Verbinski   Male
120        5584  2014   Nov    Tuesday   Francis Lawrence   Male
45        12002  2008   Jul  Wednesday  Christopher Nolan   Male
```

[51]:
```
def encode(data):
    if data == "Male":
        return 0
```

```python
    else:
        return 1

data['gender'] = data['gender'].apply(encode)
data
```

[51]:

```
       id_x   budget  popularity     revenue  \
0     43597  237.000         150  2787965087
1     43598  300.000         139   961000000
2     43599  245.000         107   880674609
3     43600  250.000         112  1084939099
4     43602  258.000         115   890871626
…     …      …          …         …
1460  48363    0.000           3      321952
1461  48370    0.027          19     3151130
1462  48375    0.000           7           0
1463  48376    0.000           3           0
1464  48395    0.220          14     2040920

                                         title  vote_average  vote_count  \
0                                       Avatar           7.2       11800
1     Pirates of the Caribbean: At World's End           6.9        4500
2                                      Spectre           6.3        4466
3                        The Dark Knight Rises           7.6        9106
4                                 Spider-Man 3           5.9        3576
…                                          …           …          …
1460                            The Last Waltz           7.9          64
1461                                    Clerks           7.4         755
1462                                   Rampage           6.0         131
1463                                   Slacker           6.4          77
1464                               El Mariachi           6.6         238

      year month        day      director_name  gender
0     2009   Dec   Thursday      James Cameron       0
1     2007   May   Saturday     Gore Verbinski       0
2     2015   Oct     Monday         Sam Mendes       0
3     2012   Jul     Monday  Christopher Nolan       0
4     2007   May    Tuesday          Sam Raimi       0
…    …    …       …                …      …
1460  1978   May     Monday    Martin Scorsese       0
1461  1994   Sep    Tuesday        Kevin Smith       0
1462  2009   Aug     Friday           Uwe Boll       0
1463  1990   Jul     Friday  Richard Linklater       0
1464  1992   Sep     Friday   Robert Rodriguez       1

[1465 rows x 12 columns]
```

[ ]:

```
### Find sum of revenue and budget per movie
```

[55]:

```
data[['revenue', 'budget']].sum(axis=1)
```

[55]:

```
0       2.787965e+09
1       9.610003e+08
2       8.806749e+08
3       1.084939e+09
4       8.908719e+08
            …
1460    3.219520e+05
1461    3.151130e+06
1462    0.000000e+00
1463    0.000000e+00
1464    2.040920e+06
Length: 1465, dtype: float64
```

[ ]:

```
#Profit per movie [apply funciton]
```

[59]:

```
!gdown 15zIxR-IvXI8s9EoHMUZXvP40HXo5bIkP
```

```
Downloading…
From: https://drive.google.com/uc?id=15zIxR-IvXI8s9EoHMUZXvP40HXo5bIkP
To: /Users/satish/Desktop/scaler/Dec Tue Batch - DAV-1/Sample - Superstore.xlsx
100%|                    | 1.21M/1.21M [00:00<00:00, 7.51MB/s]
```

[60]:

```
exl = pd.read_excel('Sample - Superstore.xlsx', sheet_name='Orders')
exl
```

```
/usr/local/lib/python3.9/site-packages/openpyxl/worksheet/_reader.py:329:
UserWarning: Unknown extension is not supported and will be removed
  warn(msg)
```

[60]:

```
      Row ID        Order ID Order Date  Ship Date       Ship Mode  \
0        1.0  CA-2021-152156 2021-11-08 2021-11-11    Second Class
1        2.0  CA-2021-152156 2021-11-08 2021-11-11    Second Class
2        3.0  CA-2021-138688 2021-06-12 2021-06-16    Second Class
3        4.0  US-2020-108966 2020-10-11 2020-10-18  Standard Class
4        5.0  US-2020-108966 2020-10-11 2020-10-18  Standard Class
…        …             …          …          …               …
9989  9990.0  CA-2019-110422 2019-01-21 2019-01-23    Second Class
9990  9991.0  CA-2022-121258 2022-02-26 2022-03-03  Standard Class
9991  9992.0  CA-2022-121258 2022-02-26 2022-03-03  Standard Class
9992  9993.0  CA-2022-121258 2022-02-26 2022-03-03  Standard Class
9993  9994.0  CA-2022-119914 2022-05-04 2022-05-09    Second Class

     Customer ID     Customer Name    Segment Country/Region             City  \
0       CG-12520       Claire Gute   Consumer  United States        Henderson
1       CG-12520       Claire Gute   Consumer  United States        Henderson
2       DV-13045   Darrin Van Huff  Corporate  United States      Los Angeles
3       SO-20335    Sean O'Donnell   Consumer  United States  Fort Lauderdale
4       SO-20335    Sean O'Donnell   Consumer  United States  Fort Lauderdale
…          …               …        …            …              …
9989    TB-21400  Tom Boeckenhauer   Consumer  United States            Miami
9990    DB-13060       Dave Brooks   Consumer  United States       Costa Mesa
9991    DB-13060       Dave Brooks   Consumer  United States       Costa Mesa
9992    DB-13060       Dave Brooks   Consumer  United States       Costa Mesa
9993    CC-12220      Chris Cortes   Consumer  United States      Westminster

      … Postal Code  Region       Product ID         Category Sub-Category  \
0     …     42420.0   South  FUR-BO-10001798        Furniture    Bookcases
1     …     42420.0   South  FUR-CH-10000454        Furniture       Chairs
2     …     90036.0    West  OFF-LA-10000240  Office Supplies       Labels
3     …     33311.0   South  FUR-TA-10000577        Furniture       Tables
4     …     33311.0   South  OFF-ST-10000760  Office Supplies      Storage
…   …         …     …              …              …          …
9989  …     33180.0   South  FUR-FU-10001889        Furniture  Furnishings
9990  …     92627.0    West  FUR-FU-10000747        Furniture  Furnishings
9991  …     92627.0    West  TEC-PH-10003645       Technology       Phones
9992  …     92627.0    West  OFF-PA-10004041  Office Supplies        Paper
9993  …     92683.0    West  OFF-AP-10002684  Office Supplies   Appliances

                                           Product Name     Sales  Quantity  \
0                     Bush Somerset Collection Bookcase  261.9600       2.0
1     Hon Deluxe Fabric Upholstered Stacking Chairs,…  731.9400       3.0
2     Self-Adhesive Address Labels for Typewriters b…   14.6200       2.0
3         Bretford CR4500 Series Slim Rectangular Table  957.5775       5.0
4                        Eldon Fold 'N Roll Cart System   22.3680       2.0
…                                                 …       …       …
9989                             Ultra Door Pull Handle   25.2480       3.0
9990  Tenex B1-RE Series Chair Mats for Low Pile Car…   91.9600       2.0
9991                              Aastra 57i VoIP phone  258.5760       2.0
9992  It's Hot Message Books with Stickers, 2 3/4" x 5"   29.6000       4.0
9993  Acco 7-Outlet Masterpiece Power Center, Wihtou…  243.1600       2.0

      Discount    Profit
0         0.00   41.9136
1         0.00  219.5820
2         0.00    6.8714
3         0.45 -383.0310
4         0.20    2.5164
…          …        …
9989      0.20    4.1028
9990      0.00   15.6332
```

```
9991      0.20   19.3932
9992      0.00   13.3200
9993      0.00   72.9480

[9994 rows x 21 columns]
```

[ ]: