August 3, 2021

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

[2]:
```
pd.read_csv('C:/Users/  /OneDrive/   /movies.csv')
```

[2]:
```
          title   distributor genre release_time  time screening_rat director  \
0                                    2012-11-22    96
1                    ( )          2015-11-19   130
2                   ( )          2013-06-05   123       15
3                   ( )NEW        2012-07-12   101
4                ( )            2010-11-04   108       15
..          …             …   …           …   …           …       …
595                  ( )NEW        2014-08-13   111
596                 ( )          2013-03-14   127       15
597               ( )            2010-09-30    99
598              CJ              2015-05-14   102       15
599               CJ             2013-01-30   120       15

     dir_prev_bfnum  dir_prev_num  num_staff  num_actor  box_off_num
0               NaN             0         91          2        23398
1        1161602.50             2        387          3      7072501
2         220775.25             4        343          4      6959083
3          23894.00             2         20          6       217866
4              1.00             1        251          2       483387
..               …           …         …         …           …
595         3833.00             1        510          7      1475091
596       496061.00             1        286          6      1716438
597             NaN             0        123          4         2475
598             NaN             0        431          4      2192525
599             NaN             0        363          5      7166532

[600 rows x 12 columns]
```

[3]:
```
all_movie = pd.read_csv('C:/Users/  /OneDrive/   /movies.csv')
```

[4]:
```
print(all_movie.shape)
```

```
(600, 12)
```

[5]:
```
all_movie.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 600 entries, 0 to 599
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   title           600 non-null    object
 1   distributor     600 non-null    object
 2   genre           600 non-null    object
 3   release_time    600 non-null    object
 4   time            600 non-null    int64
 5   screening_rat   600 non-null    object
 6   director        600 non-null    object
 7   dir_prev_bfnum  270 non-null    float64
 8   dir_prev_num    600 non-null    int64
 9   num_staff       600 non-null    int64
 10  num_actor       600 non-null    int64
 11  box_off_num     600 non-null    int64
dtypes: float64(1), int64(5), object(6)
memory usage: 56.4+ KB
```

[6]:
```
all_movie.isna()
```

[6]:

	title	distributor	genre	release_time	time	screening_rat	director	\
0	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	
..	…	…	…	…	…	…	…	
595	False	False	False	False	False	False	False	
596	False	False	False	False	False	False	False	
597	False	False	False	False	False	False	False	
598	False	False	False	False	False	False	False	
599	False	False	False	False	False	False	False	

	dir_prev_bfnum	dir_prev_num	num_staff	num_actor	box_off_num
0	True	False	False	False	False
1	False	False	False	False	False
2	False	False	False	False	False
3	False	False	False	False	False
4	False	False	False	False	False
..	…	…	…	…	…
595	False	False	False	False	False
596	False	False	False	False	False
597	True	False	False	False	False
598	True	False	False	False	False

```
599            True         False      False      False        False

[600 rows x 12 columns]
```

[7]:
```
all_movie.fillna(0)
```

[7]:
```
         title   distributor genre release_time  time screening_rat director  \
0                                  2012-11-22    96
1                   ( )          2015-11-19   130
2                  ( )          2013-06-05   123       15
3                  ( )NEW        2012-07-12   101
4               ( )           2010-11-04   108       15
..          …            …   …           …   …            …       …
595                 ( )NEW        2014-08-13   111
596                ( )          2013-03-14   127       15
597              ( )            2010-09-30    99
598              CJ             2015-05-14   102       15
599               CJ             2013-01-30   120       15

     dir_prev_bfnum  dir_prev_num  num_staff  num_actor  box_off_num
0              0.00             0         91          2        23398
1        1161602.50             2        387          3      7072501
2         220775.25             4        343          4      6959083
3          23894.00             2         20          6       217866
4              1.00             1        251          2       483387
..               …           …        …         …           …
595          3833.00             1        510          7      1475091
596        496061.00             1        286          6      1716438
597            0.00             0        123          4         2475
598            0.00             0        431          4      2192525
599            0.00             0        363          5      7166532

[600 rows x 12 columns]
```

[8]:
```
all_Movie = all_movie.fillna(0)
```

[9]:
```
all_Movie.describe()
```

[9]:
```
             time  dir_prev_bfnum  dir_prev_num   num_staff   num_actor  \
count  600.000000    6.000000e+02    600.000000  600.000000  600.000000
mean   100.863333    4.726993e+05      0.876667  151.118333    3.706667
std     18.097528    1.309474e+06      1.183409  165.654671    2.446889
min     45.000000    0.000000e+00      0.000000    0.000000    0.000000
25%     89.000000    0.000000e+00      0.000000   17.000000    2.000000
50%    100.000000    0.000000e+00      0.000000   82.500000    3.000000
75%    114.000000    3.761416e+05      2.000000  264.000000    4.000000
max    180.000000    1.761531e+07      5.000000  869.000000   25.000000
```

```
        box_off_num
count  6.000000e+02
mean   7.081818e+05
std    1.828006e+06
min    1.000000e+00
25%    1.297250e+03
50%    1.259100e+04
75%    4.798868e+05
max    1.426277e+07
```

[10]:
```
all_Movie[['genre', 'box_off_num']].groupby('genre').mean().
↪sort_values('box_off_num')
```

[10]:
```
          box_off_num
genre
       6.627000e+03
      6.717226e+04
       8.261100e+04
      1.819267e+05
 /    4.259680e+05
       5.275482e+05
        5.908325e+05
       6.256898e+05
       1.193914e+06
SF      1.788346e+06
       2.203974e+06
       2.263695e+06
```

[11]:
```
all_Movie[['screening_rat', 'box_off_num']].groupby('screening_rat').mean().
↪sort_values('box_off_num')
```

[11]:
```
                box_off_num
screening_rat
             1.351005e+05
            3.641813e+05
12           8.449809e+05
15           1.247519e+06
```

[12]:
```
all_Movie[['director', 'box_off_num']].groupby('director').mean().
↪sort_values('box_off_num')
```

[12]:
```
           box_off_num
director
                 1.0
                 2.0
                 8.0
```

```
                10.0
                36.0
…                …
           9135806.0
           9350351.0
          11374879.0
          12845252.0
          14262766.0

[472 rows x 1 columns]
```

[13]:
```
all_Movie[['distributor', 'box_off_num']].groupby('distributor').mean().
↪sort_values('box_off_num')
```

[13]:
```
                  box_off_num
distributor
             2.000000e+00
               8.000000e+00
               4.200000e+01
                4.600000e+01
                5.400000e+01
…                        …
               2.541603e+06
  ( )          2.634823e+06
( )            3.117859e+06
( )               3.386656e+06
CJ E&M Pictures  4.122337e+06

[169 rows x 1 columns]
```

[14]:
```
all_Movie['genre_num'] = all_Movie.genre.map({'  ':1, '   ':2, '   ':3, '   ':4,␣
↪' /  ':5, '   ':6,
                                               ' ':7, '  ':8, '  ':9, 'SF':10,␣
↪' ':11, '  ':12})
```

[15]:
```
dis_rank = all_Movie.groupby('distributor').box_off_num.median().
↪reset_index(name = 'dis_rank_num').sort_values(by = 'dis_rank_num')
dis_rank
```

[15]:
```
         distributor  dis_rank_num
141                        2.0
65                          8.0
92                         42.0
131                         46.0
68                          54.0
..                 …           …
50       CJ E&M         2242510.0
```

```
121                    2541603.0
96          ( )        2634823.0
27        ( )          3117859.0
49    CJ E&M Pictures     4122337.0

[169 rows x 2 columns]
```

[16]:
```
dis_rank.shape[0]
```

[16]:
```
169
```

[17]:
```
dis_rank['dis_rank_num'] = [i+1 for i in range(dis_rank.shape[0])]
dis_rank
```

[17]:
```
          distributor  dis_rank_num
141                            1
65                              2
92                              3
131                             4
68                              5
..                 …            …
50        CJ E&M                165
121                            166
96          ( )               167
27        ( )                 168
49    CJ E&M Pictures             169

[169 rows x 2 columns]
```

[18]:
```
all_Movie = pd.merge(all_Movie, dis_rank, how='left')
```

[19]:
```
all_Movie
```

[19]:
```
          title   distributor genre release_time  time screening_rat director  \
0                                  2012-11-22    96
1                     ( )          2015-11-19   130
2                   ( )            2013-06-05   123       15
3                   ( )NEW         2012-07-12   101
4                ( )             2010-11-04   108       15
..          …            …   …           …   …            …       …
595                   ( )NEW         2014-08-13   111
596                 ( )            2013-03-14   127       15
597               ( )              2010-09-30    99
598               CJ               2015-05-14   102       15
599                CJ              2013-01-30   120       15

     dir_prev_bfnum  dir_prev_num  num_staff  num_actor  box_off_num  \
```

```
0              0.00             0         91          2        23398
1        1161602.50             2        387          3      7072501
2         220775.25             4        343          4      6959083
3          23894.00             2         20          6       217866
4              1.00             1        251          2       483387
..               …           …        …        …          …
595          3833.00             1        510          7      1475091
596        496061.00             1        286          6      1716438
597             0.00             0        123          4         2475
598             0.00             0        431          4      2192525
599             0.00             0        363          5      7166532

     genre_num  dis_rank_num
0         11.0           151
1         12.0           164
2         11.0           164
3          9.0           158
4          9.0           167
..         …           …
595        8.0           158
596        8.0           164
597        7.0            49
598       12.0           159
599       11.0           159

[600 rows x 14 columns]
```

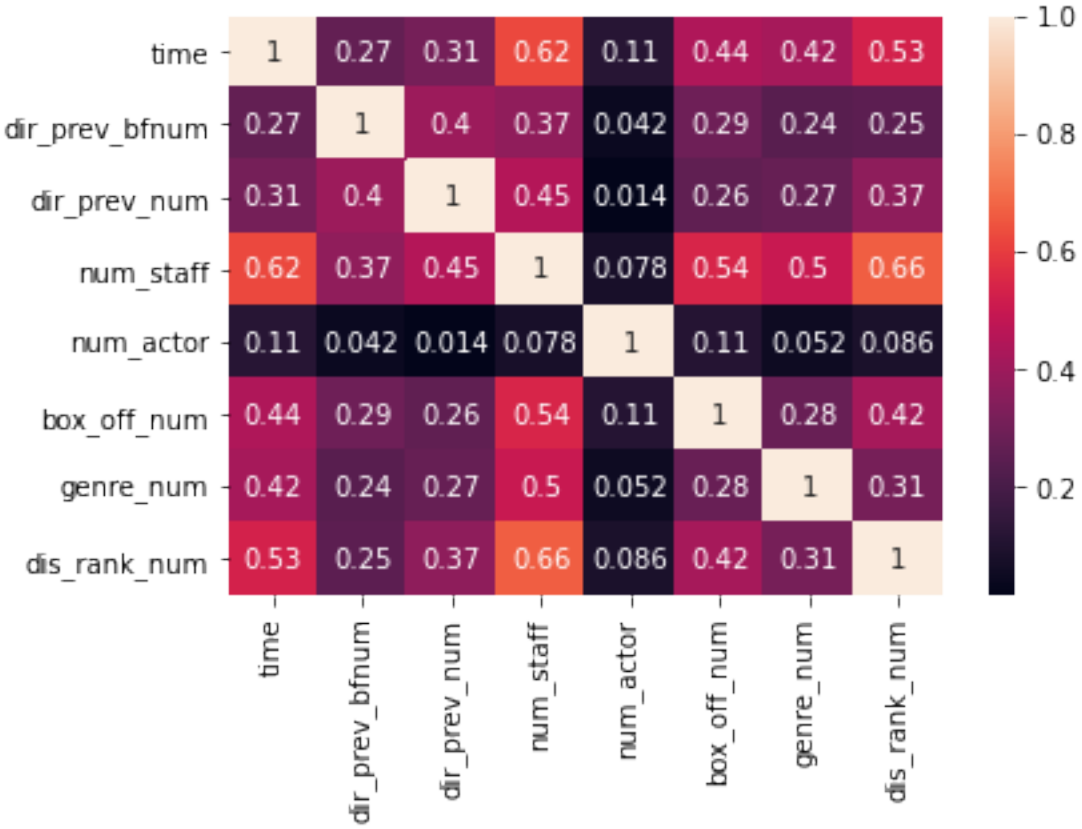
[20]:
```
all_Movie.corr()
```

[20]:
```
                    time  dir_prev_bfnum  dir_prev_num  num_staff  num_actor  \
time            1.000000        0.266065      0.306727   0.623205   0.114153
dir_prev_bfnum  0.266065        1.000000      0.396616   0.369657   0.042491
dir_prev_num    0.306727        0.396616      1.000000   0.450706   0.014006
num_staff       0.623205        0.369657      0.450706   1.000000   0.077871
num_actor       0.114153        0.042491      0.014006   0.077871   1.000000
box_off_num     0.441452        0.293791      0.259674   0.544265   0.111179
genre_num       0.420855        0.239070      0.274047   0.501566   0.051658
dis_rank_num    0.533877        0.250900      0.367591   0.664916   0.086059

                box_off_num  genre_num  dis_rank_num
time               0.441452   0.420855      0.533877
dir_prev_bfnum     0.293791   0.239070      0.250900
dir_prev_num       0.259674   0.274047      0.367591
num_staff          0.544265   0.501566      0.664916
num_actor          0.111179   0.051658      0.086059
box_off_num        1.000000   0.277633      0.419216
genre_num          0.277633   1.000000      0.311629
```

```
dis_rank_num        0.419216    0.311629        1.000000
```

[31]:
```
import seaborn as sns

sns.heatmap(all_Movie.corr(), annot=True)
```

[31]:
```
<AxesSubplot:>
```



from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X_movie, y_movie, random_state=0, test_size=0.2)

train, test = train_test_split(all_Movie, randdom_state=0, test_size=0.2)

[ ]: