Day_37_021223

January 23, 2024

[3]:
```
import numpy as np
import pandas as pd
!gdown 1s2TkjSpzNc4SyxqRrQleZyDIHlc7bxnd
!gdown 1Ws-_s1fHZ9nHfGLVUQurbHDvStePlEJm
movies = pd.read_csv("movies.csv",index_col=0)
directors = pd.read_csv("directors.csv",index_col=0)
data = pd.merge(movies,directors,left_on="director_id",right_on='id',how='left')
data.drop('id_y',axis=1,inplace=True)
data.rename({"id_x":"movies_id"},axis=1,inplace=True)
data
```

```
Downloading…
From: https://drive.google.com/uc?id=1s2TkjSpzNc4SyxqRrQleZyDIHlc7bxnd
To: C:\Data\Data_science\Data Science RIA\3 Python\Pandas\Codes\movies.csv

  0%|          | 0.00/112k [00:00<?, ?B/s]
100%|##########| 112k/112k [00:00<00:00, 1.60MB/s]
Downloading…
From: https://drive.google.com/uc?id=1Ws-_s1fHZ9nHfGLVUQurbHDvStePlEJm
To: C:\Data\Data_science\Data Science RIA\3 Python\Pandas\Codes\directors.csv

  0%|          | 0.00/65.4k [00:00<?, ?B/s]
100%|##########| 65.4k/65.4k [00:00<00:00, 41.2MB/s]
```

[3]:

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

[1465 rows x 13 columns]
```

1 Grouping in Pandas

[4]:
```
data.groupby('director_name').nunique()
```

[4]:
```
                           movies_id  budget  popularity  revenue  title  \
director_name
Adam McKay                         6       6           6        6      6
Adam Shankman                      8       8           7        8      8
Alejandro González Iñárritu        6       6           6        6      6
Alex Proyas                        5       5           5        5      5
Alexander Payne                    5       5           5        5      5
…                                …       …           …        …      …
Wes Craven                        10       7           9       10     10
Wolfgang Petersen                  7       7           7        7      7
Woody Allen                       18       9          13       10     18
Zack Snyder                        7       7           7        7      7
Zhang Yimou                        6       4           6        4      6

                           vote_average  vote_count  director_id  year  \
```

```
director_name
Adam McKay                              6           6           1     6
Adam Shankman                           8           8           1     7
Alejandro González Iñárritu             6           6           1     6
Alex Proyas                             5           5           1     5
Alexander Payne                         3           5           1     5
…                                     …           …           …     …
Wes Craven                              9          10           1     9
Wolfgang Petersen                       6           7           1     7
Woody Allen                            12          18           1    18
Zack Snyder                             5           7           1     7
Zhang Yimou                             5           6           1     6

                              month  day  gender
director_name
Adam McKay                        3    2       1
Adam Shankman                     5    2       1
Alejandro González Iñárritu       5    3       1
Alex Proyas                       4    3       1
Alexander Payne                   4    2       0
…                               …  …    …
Wes Craven                        6    5       1
Wolfgang Petersen                 5    3       1
Woody Allen                       9    6       1
Zack Snyder                       4    4       1
Zhang Yimou                       2    3       1

[199 rows x 12 columns]
```

2 Get how many groups are groupped

[5]:
```
data.groupby('director_name').ngroups
```

[5]:
```
199
```

3 Displaying the groups

[6]:
```
data.groupby('director_name').groups
```

[6]:
```
{'Adam McKay': [176, 323, 366, 505, 839, 916], 'Adam Shankman': [265, 300, 350,
404, 458, 843, 999, 1231], 'Alejandro González Iñárritu': [106, 749, 1015, 1034,
1077, 1405], 'Alex Proyas': [95, 159, 514, 671, 873], 'Alexander Payne': [793,
1006, 1101, 1211, 1281], 'Andrew Adamson': [11, 43, 328, 501, 947], 'Andrew
Niccol': [533, 603, 701, 722, 1439], 'Andrzej Bartkowiak': [349, 549, 754, 911,
924], 'Andy Fickman': [517, 681, 909, 926, 973, 1023], 'Andy Tennant': [314,
320, 464, 593, 676, 885], 'Ang Lee': [99, 134, 748, 840, 1089, 1110, 1132,
```

1184], 'Anne Fletcher': [610, 650, 736, 789, 1206], 'Antoine Fuqua': [310, 338, 424, 467, 576, 808, 818, 1105], 'Atom Egoyan': [946, 1128, 1164, 1194, 1347, 1416], 'Barry Levinson': [313, 319, 471, 594, 878, 898, 1013, 1037, 1082, 1143, 1185, 1345, 1378], 'Barry Sonnenfeld': [13, 48, 90, 205, 591, 778, 783], 'Ben Stiller': [209, 212, 547, 562, 850], 'Bill Condon': [102, 307, 902, 1233, 1381], 'Bobby Farrelly': [352, 356, 481, 498, 624, 630, 654, 806, 928, 972, 1111], 'Brad Anderson': [1163, 1197, 1350, 1419, 1430], 'Brett Ratner': [24, 39, 188, 207, 238, 292, 405, 456, 920], 'Brian De Palma': [228, 255, 318, 439, 747, 905, 919, 1088, 1232, 1261, 1317, 1354], 'Brian Helgeland': [512, 607, 623, 742, 933], 'Brian Levant': [418, 449, 568, 761, 860, 1003], 'Brian Robbins': [416, 441, 669, 962, 988, 1115], 'Bryan Singer': [6, 32, 33, 44, 122, 216, 297, 1326], 'Cameron Crowe': [335, 434, 488, 503, 513, 698], 'Catherine Hardwicke': [602, 695, 724, 937, 1406, 1412], 'Chris Columbus': [117, 167, 204, 218, 229, 509, 656, 897, 996, 1086, 1129], 'Chris Weitz': [17, 500, 794, 869, 1202, 1267], 'Christopher Nolan': [3, 45, 58, 59, 74, 565, 641, 1341], 'Chuck Russell': [177, 410, 657, 1069, 1097, 1339], 'Clint Eastwood': [369, 426, 447, 482, 490, 520, 530, 535, 645, 727, 731, 786, 787, 899, 974, 986, 1167, 1190, 1313], 'Curtis Hanson': [494, 579, 606, 711, 733, 1057, 1310], 'Danny Boyle': [527, 668, 1083, 1085, 1126, 1168, 1287, 1385], 'Darren Aronofsky': [113, 751, 1187, 1328, 1363, 1458], 'Darren Lynn Bousman': [1241, 1243, 1283, 1338, 1440], 'David Ayer': [50, 273, 741, 1024, 1146, 1407], 'David Cronenberg': [541, 767, 994, 1055, 1254, 1268, 1334], 'David Fincher': [62, 213, 253, 383, 398, 478, 522, 555, 618, 785], 'David Gordon Green': [543, 862, 884, 927, 1376, 1418, 1432, 1459], 'David Koepp': [443, 644, 735, 1041, 1209], 'David Lynch': [583, 1161, 1264, 1340, 1456], 'David O. Russell': [422, 556, 609, 896, 982, 989, 1229, 1304], 'David R. Ellis': [582, 634, 756, 888, 934], 'David Zucker': [569, 619, 965, 1052, 1175], 'Dennis Dugan': [217, 260, 267, 293, 303, 718, 780, 977, 1247], 'Donald Petrie': [427, 507, 570, 649, 858, 894, 1106, 1331], 'Doug Liman': [52, 148, 251, 399, 544, 1318, 1451], 'Edward Zwick': [92, 182, 346, 566, 791, 819, 825], 'F. Gary Gray': [308, 402, 491, 523, 697, 833, 1272, 1380], 'Francis Ford Coppola': [487, 559, 622, 646, 772, 1076, 1155, 1253, 1312], 'Francis Lawrence': [63, 72, 109, 120, 679], 'Frank Coraci': [157, 249, 275, 451, 577, 599, 963], 'Frank Oz': [193, 355, 473, 580, 712, 813, 987], 'Garry Marshall': [329, 496, 528, 571, 784, 893, 1029, 1169], 'Gary Fleder': [518, 667, 689, 867, 981, 1165], 'Gary Winick': [258, 797, 798, 804, 1454], 'Gavin O'Connor': [820, 841, 939, 953, 1444], 'George A. Romero': [250, 1066, 1096, 1278, 1367, 1396], 'George Clooney': [343, 450, 831, 966, 1302], 'George Miller': [78, 103, 233, 287, 1250, 1403, 1450], 'Gore Verbinski': [1, 8, 9, 107, 119, 633, 1040], 'Guillermo del Toro': [35, 252, 419, 486, 1118], 'Gus Van Sant': [595, 1018, 1027, 1159, 1240, 1311, 1398], 'Guy Ritchie': [124, 215, 312, 1093, 1225, 1269, 1420], 'Harold Ramis': [425, 431, 558, 586, 788, 1137, 1166, 1325], 'Ivan Reitman': [274, 643, 816, 883, 910, 935, 1134, 1242], 'James Cameron': [0, 19, 170, 173, 344, 1100, 1320], 'James Ivory': [1125, 1152, 1180, 1291, 1293, 1390, 1397], 'James Mangold': [140, 141, 557, 560, 829, 845, 958, 1145], 'James Wan': [30, 617, 1002, 1047, 1337, 1417, 1424], 'Jan de Bont': [155, 224, 231, 270, 781], 'Jason Friedberg': [812, 1010, 1012, 1014, 1036], 'Jason Reitman': [792, 1092, 1213, 1295, 1299], 'Jaume Collet-Serra': [516, 540, 640, 725, 1011, 1189], 'Jay Roach': [195, 359, 389,

```
397, 461, 703, 859, 1072], 'Jean-Pierre Jeunet': [423, 485, 605, 664, 765], 'Joe
Dante': [284, 525, 638, 1226, 1298, 1428], 'Joe Wright': [85, 432, 553, 803,
814, 855], 'Joel Coen': [428, 670, 691, 707, 721, 889, 906, 980, 1157, 1238,
1305], 'Joel Schumacher': [128, 184, 348, 484, 572, 614, 652, 764, 876, 886,
1108, 1230, 1280], 'John Carpenter': [537, 663, 686, 861, 938, 1028, 1080, 1102,
1329, 1371], 'John Glen': [601, 642, 801, 847, 864], 'John Landis': [524, 868,
1276, 1384, 1435], 'John Madden': [457, 882, 1020, 1249, 1257], 'John
McTiernan': [127, 214, 244, 351, 534, 563, 648, 782, 838, 1074], 'John
Singleton': [294, 489, 732, 796, 1120, 1173, 1316], 'John Whitesell': [499, 632,
763, 1119, 1148], 'John Woo': [131, 142, 264, 371, 420, 675, 1182], 'Jon
Favreau': [46, 54, 55, 382, 759, 1346], 'Jon M. Chu': [100, 225, 810, 1099,
1186], 'Jon Turteltaub': [64, 180, 372, 480, 760, 846, 1171], 'Jonathan Demme':
[277, 493, 1000, 1123, 1215], 'Jonathan Liebesman': [81, 143, 339, 1117, 1301],
'Judd Apatow': [321, 710, 717, 865, 881], 'Justin Lin': [38, 123, 246, 1437,
1447], 'Kenneth Branagh': [80, 197, 421, 879, 1094, 1277, 1288], 'Kenny Ortega':
[412, 852, 1228, 1315, 1365], 'Kevin Reynolds': [53, 502, 639, 1019, 1059], …}
```

4 Accessing the grouped elements

[7]:
```
data.groupby('director_name').get_group('James Cameron')
```

[7]:

	movies_id	budget	popularity	revenue	title	vote_average	vote_count	director_id	year	month	day	director_name	gender
0	43597	237000000	150	2787965087	Avatar	7.2	11800	4762	2009	Dec	Thursday	James Cameron	Male
19	43622	200000000	100	1845034188	Titanic	7.5	7562	4762	1997	Nov	Tuesday	James Cameron	Male
170	43876	100000000	101	520000000	Terminator 2: Judgment Day	7.7	4185	4762	1991	Jul	Monday	James Cameron	Male
173	43879	115000000	38	378882411	True Lies	6.8	1116	4762	1994	Jul	Thursday	James Cameron	Male
344	44184	70000000	24	90000098	The Abyss	7.1	808	4762	1989	Aug	Wednesday	James Cameron	Male
1100	46000	18500000	67	183316455	Aliens	7.7	3220	4762	1986	Jul	Friday	James Cameron	Male
1320	47036	6400000	74	78371200	The Terminator	7.3	4128	4762	1984				

```
1320   Oct     Friday  James Cameron   Male
```

[8]:
```
data.groupby('director_name')['title'].count().sort_values(ascending=False)
```

[8]:
```
director_name
Steven Spielberg    26
Clint Eastwood      19
Martin Scorsese     19
Woody Allen         18
Robert Rodriguez    16
                    ..
Paul Weitz           5
John Madden          5
Paul Verhoeven       5
John Whitesell       5
Kevin Reynolds       5
Name: title, Length: 199, dtype: int64
```

[9]:
```
data.groupby('director_name')['title'].value_counts()
```

[9]:
```
director_name  title
Adam McKay     Anchorman 2: The Legend Continues                1
               Anchorman: The Legend of Ron Burgundy            1
               The Other Guys                                   1
               The Big Short                                    1
               Talladega Nights: The Ballad of Ricky Bobby      1
                                                               ..
Zhang Yimou    Hero                                             1
               Curse of the Golden Flower                       1
               Coming Home                                      1
               A Woman, a Gun and a Noodle Shop                 1
               The Flowers of War                               1
Name: count, Length: 1465, dtype: int64
```

[10]:
```
data.groupby('director_name')['year'].aggregate(['min','max'])
```

[10]:
```
                              min   max
director_name
Adam McKay                   2004  2015
Adam Shankman                2001  2012
Alejandro González Iñárritu  2000  2015
Alex Proyas                  1994  2016
Alexander Payne              1999  2013
…                             …     …
Wes Craven                   1984  2011
Wolfgang Petersen            1981  2006
Woody Allen                  1977  2013
```

```
Zack Snyder                    2004  2016
Zhang Yimou                    2002  2014

[199 rows x 2 columns]
```

5 Get me the list of High budget directors

- Atleast 1 movie with 1 Million Budget

Getting max budget of directors

[11]:
```
data_dir_budget = data.groupby('director_name')['budget'].max().reset_index()
```

Names of directors who have more than 1 million budget movies

[12]:
```
names = data_dir_budget.
  ↪loc[data_dir_budget['budget']>=100000000]['director_name']
```

Checking whether names are present in data

[13]:
```
data.loc[data['director_name'].isin(names)]
```

[13]:
```
      movies_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
…           …          …           …           …
1450      48267     400000          33   100000000
1451      48268     200000          13     4505922
1452      48274          0           5     2611555
1458      48335      60000          27     3221152
1460      48363          0           3      321952

                                      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
…                                        …             …           …
1450                               Mad Max           6.6        1213
1451                              Swingers           6.8         253
1452                                 Three           6.3          31
1458                                    Pi           7.1         586
1460                         The Last Waltz           7.9          64

      director_id  year month       day      director_name gender
```

```
0            4762  2009   Dec  Thursday      James Cameron   Male
1            4763  2007   May  Saturday     Gore Verbinski   Male
2            4764  2015   Oct    Monday         Sam Mendes   Male
3            4765  2012   Jul    Monday  Christopher Nolan   Male
4            4767  2007   May   Tuesday          Sam Raimi   Male
…             …     …     …       …                …     …
1450         4845  1979   Apr  Thursday      George Miller   Male
1451         4813  1996   Oct    Friday         Doug Liman   Male
1452         4936  2010   Dec  Thursday         Tom Tykwer   Male
1458         4881  1998   Jul    Friday   Darren Aronofsky   Male
1460         4809  1978   May    Monday    Martin Scorsese   Male

[679 rows x 13 columns]
```

[21]:
```
def high_budget(data):
    return data['budget'].max() >= 100000000

data.groupby('director_name').filter(high_budget)
```

[21]:
```
      movies_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
…           …          …           …           …
1450      48267     400000          33   100000000
1451      48268     200000          13     4505922
1452      48274          0           5     2611555
1458      48335      60000          27     3221152
1460      48363          0           3      321952

                                      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
…                                         …             …           …
1450                                Mad Max           6.6        1213
1451                               Swingers           6.8         253
1452                                  Three           6.3          31
1458                                     Pi           7.1         586
1460                          The Last Waltz           7.9          64

      director_id  year month       day      director_name gender
0            4762  2009   Dec  Thursday      James Cameron   Male
```

```
1            4763  2007   May  Saturday     Gore Verbinski   Male
2            4764  2015   Oct    Monday         Sam Mendes   Male
3            4765  2012   Jul    Monday  Christopher Nolan   Male
4            4767  2007   May   Tuesday          Sam Raimi   Male
…             …     …     …       …                 …     …
1450         4845  1979   Apr  Thursday      George Miller   Male
1451         4813  1996   Oct    Friday         Doug Liman   Male
1452         4936  2010   Dec  Thursday         Tom Tykwer   Male
1458         4881  1998   Jul    Friday   Darren Aronofsky   Male
1460         4809  1978   May    Monday    Martin Scorsese   Male

[679 rows x 13 columns]
```

6 Find out the Risky Movies

- Average Revenue of the Director - 10, 20, 15, 20, 18 - 21M
- Risky - 21M : 25M,30M, 18M, 10M, 50M

[30]:
```
def is_risky(x):
    x['is_risky'] = (x['budget']-x['revenue'].mean())>= 0
    return x

data_risky = data.groupby('director_name').apply(is_risky)
```

[31]:
```
data_risky.head()
```

[31]:
```
                    movies_id     budget  popularity    revenue  \
director_name
Adam McKay    176      43882  100000000          24  170432927
              323      44151   72500000          12  162966177
              366      44236   65000000          22  128107642
              505      44503   50000000          38  173649015
              839      45301   28000000          57  133346506

                                                      title  vote_average  \
director_name
Adam McKay    176                                The Other Guys           6.1
              323  Talladega Nights: The Ballad of Ricky Bobby           6.2
              366                                  Step Brothers           6.5
              505            Anchorman 2: The Legend Continues           6.0
              839                                  The Big Short           7.3

                    vote_count  director_id  year month        day  \
director_name
Adam McKay    176        1383         4925  2010   Aug     Friday
              323         491         4925  2006   Aug     Friday
              366        1062         4925  2008   Jul     Friday
```

```
              505         923         4925  2013   Dec  Wednesday
              839        2607         4925  2015   Dec     Friday

                  director_name gender  is_risky
director_name
Adam McKay    176    Adam McKay   Male     False
              323    Adam McKay   Male     False
              366    Adam McKay   Male     False
              505    Adam McKay   Male     False
              839    Adam McKay   Male     False
```

[32]:
```
data_risky.loc[data_risky['is_risky']==True]
```

[32]:
```
                        movies_id     budget  popularity    revenue  \
director_name
Andrzej Bartkowiak 349       44192   60000000          29   55987321
Atom Egoyan        946       45538   25000000           4          0
                   1194      46370   15000000          26    8459458
                   1347      47224    5000000           7    3263585
Brett Ratner       24        43630  210000000           3  459359555
…                            …         …          …         …
Uwe Boll           944       45536   25000000           7    2405420
                   1058      45834   20000000           9   10442808
                   1383      47453    3500000           4          0
Wayne Wang         468       44419   55000000          19  154906693
Zhang Yimou        192       43914   94000000          12   95311434

                                              title  vote_average  \
director_name
Andrzej Bartkowiak 349                         Doom           5.0
Atom Egoyan        946          Where the Truth Lies           5.9
                   1194                        Chloe           5.9
                   1347          The Sweet Hereafter           6.8
Brett Ratner       24          X-Men: The Last Stand           6.3
…                                                …           …
Uwe Boll           944                   BloodRayne           3.5
                   1058            Alone in the Dark           3.1
                   1383  In the Name of the King III           3.3
Wayne Wang         468             Maid in Manhattan           5.6
Zhang Yimou        192            The Flowers of War           7.1

                        vote_count  director_id  year month        day  \
director_name
Andrzej Bartkowiak 349         609         5061  2005   Oct   Thursday
Atom Egoyan        946          66         5599  2005   Oct     Friday
                   1194        498         5599  2009   Mar  Wednesday
                   1347        103         5599  1997   May  Wednesday
```

```
Brett Ratner       24          3525         4786  2006   May  Wednesday
…                             …          …   …   …        …
Uwe Boll           944          118         5148  2005   Oct   Saturday
                   1058         173         5148  2005   Jan     Friday
                   1383          19         5148  2013   Dec     Friday
Wayne Wang         468          485         5162  2002   Dec     Friday
Zhang Yimou        192          187         4945  2011   Dec   Thursday

                                director_name gender  is_risky
director_name
Andrzej Bartkowiak 349   Andrzej Bartkowiak   Male      True
Atom Egoyan        946          Atom Egoyan   Male      True
                   1194         Atom Egoyan   Male      True
                   1347         Atom Egoyan   Male      True
Brett Ratner       24          Brett Ratner   Male      True
…                                     …    …       …
Uwe Boll           944             Uwe Boll   Male      True
                   1058            Uwe Boll   Male      True
                   1383            Uwe Boll   Male      True
Wayne Wang         468           Wayne Wang    NaN      True
Zhang Yimou        192          Zhang Yimou   Male      True

[131 rows x 14 columns]
```

[ ]: