pandas-lecture-4-dec-batch

June 10, 2023

0.1 Pandas-Lecture - 4

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

[3]:
```
movies = pd.read_csv('movies.csv', index_col=0)
directors = pd.read_csv('directors.csv', index_col=0)
data = movies.merge(directors, left_on='director_id', right_on='id', how='left')
data.drop(['director_id', 'id_y'], axis=1, inplace=True)
data
```

[3]:

	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]
```

[30]:
```
data['revenue'] = data['revenue']/1000000
data['budget'] = data['budget']/1000000
```

[11]:
```
data.loc[data['director_name'] == 'Christopher Nolan', ['title']].count()
```

[11]:
```
title    8
dtype: int64
```

[12]:
```
data['director_name'].value_counts()
```

[12]:
```
Steven Spielberg      26
Martin Scorsese       19
Clint Eastwood        19
Woody Allen           18
Ridley Scott          16
                      ..
Tim Hill               5
Jonathan Liebesman     5
Roman Polanski         5
Larry Charles          5
Nicole Holofcener      5
Name: director_name, Length: 199, dtype: int64
```

[13]:
```
data.groupby('director_name')
```

[13]:
```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x121486b20>
```

[14]:
```
data['director_name'].nunique()
```

[14]:
```
199
```

[15]:
```
data.groupby('director_name').ngroups
```

[15]: 199

[16]:
```
data.groupby('director_name').groups
```

[16]:
```
{'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], …}
```

[17]:
```
data.groupby('director_name').get_group('Alexander Payne')
```

[17]:

	id_x	budget	popularity	revenue	title	vote_average	vote_count	year	month	day	director_name	gender
793	45163	30000000	19	105834556	About Schmidt	6.7	362	2002	Dec	Friday	Alexander Payne	NaN
1006	45699	20000000	40	177243185	The Descendants	6.7	934	2011	Sep	Friday	Alexander Payne	NaN
1101	46004	16000000	23	109502303	Sideways	6.9	478	2004	Oct	Friday	Alexander Payne	NaN
1211	46446	12000000	29	17654912	Nebraska	7.4	636	2013	Sep	Saturday	Alexander Payne	NaN
1281	46813	0	13	0	Election	6.7	270	1999	Apr	Friday	Alexander Payne	NaN

[21]:
```
data.groupby('director_name')['title'].count().sort_values(ascending=False)
```

[21]:
```
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
```

[28]:
```
data.groupby('director_name')['year'].aggregate(['min', 'max'])
```

[28]:

director_name	min	max
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]

[32]:
```
# Highest movie budget of a director
data_dir_budget = data.groupby('director_name')['budget'].max().reset_index()
data_dir_budget
```

[32]:

	director_name	budget
0	Adam McKay	100.0
1	Adam Shankman	80.0
2	Alejandro González Iñárritu	135.0
3	Alex Proyas	140.0
4	Alexander Payne	30.0
..	…	…
194	Wes Craven	40.0
195	Wolfgang Petersen	175.0
196	Woody Allen	30.0
197	Zack Snyder	250.0

```
198                  Zhang Yimou    94.0

[199 rows x 2 columns]
```

[35]:
```
names = data_dir_budget.loc[data_dir_budget['budget'] >= 100, 'director_name']
names
```

[35]:
```
0                       Adam McKay
2      Alejandro González Iñárritu
3                      Alex Proyas
5                   Andrew Adamson
10                         Ang Lee
                  …
187                    Tom Shadyac
188                     Tom Tykwer
189                     Tony Scott
195               Wolfgang Petersen
197                    Zack Snyder
Name: director_name, Length: 85, dtype: object
```

[36]:
```
data.loc[data['director_name'].isin(names)]
```

[36]:

	id_x	budget	popularity	revenue	title	vote_average	vote_count
0	43597	237.00	150	2787.965087	Avatar	7.2	11800
1	43598	300.00	139	961.000000	Pirates of the Caribbean: At World's End	6.9	4500
2	43599	245.00	107	880.674609	Spectre	6.3	4466
3	43600	250.00	112	1084.939099	The Dark Knight Rises	7.6	9106
4	43602	258.00	115	890.871626	Spider-Man 3	5.9	3576
…	…	…	…	…	…	…	…
1450	48267	0.40	33	100.000000	Mad Max	6.6	1213
1451	48268	0.20	13	4.505922	Swingers	6.8	253
1452	48274	0.00	5	2.611555	Three	6.3	31
1458	48335	0.06	27	3.221152	Pi	7.1	586
1460	48363	0.00	3	0.321952	The Last Waltz	7.9	64

```
      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
…     …    …       …                  …    …
1450  1979   Apr  Thursday      George Miller   Male
1451  1996   Oct    Friday         Doug Liman   Male
1452  2010   Dec  Thursday         Tom Tykwer   Male
1458  1998   Jul    Friday   Darren Aronofsky   Male
1460  1978   May    Monday    Martin Scorsese   Male

[679 rows x 12 columns]
```

[37]:
```
def high_budget(x):
    return x['budget'].max()>=100

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

[37]:
```
       id_x  budget  popularity      revenue  \
0     43597  237.00         150  2787.965087
1     43598  300.00         139   961.000000
2     43599  245.00         107   880.674609
3     43600  250.00         112  1084.939099
4     43602  258.00         115   890.871626
…     …      …          …           …
1450  48267    0.40          33   100.000000
1451  48268    0.20          13     4.505922
1452  48274    0.00           5     2.611555
1458  48335    0.06          27     3.221152
1460  48363    0.00           3     0.321952

                                       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
```

```
      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
…     …    …       …                …    …
1450  1979   Apr  Thursday      George Miller   Male
1451  1996   Oct    Friday         Doug Liman   Male
1452  2010   Dec  Thursday         Tom Tykwer   Male
1458  1998   Jul    Friday   Darren Aronofsky   Male
1460  1978   May    Monday    Martin Scorsese   Male

[679 rows x 12 columns]
```

[40]:
```
def is_risky(x):
    x['is_risky']= (x['budget'] - x['revenue'].mean()) >=0
    return x

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

[40]:
```
       id_x   budget  popularity      revenue  \
0     43597  237.000         150  2787.965087
1     43598  300.000         139   961.000000
2     43599  245.000         107   880.674609
3     43600  250.000         112  1084.939099
4     43602  258.000         115   890.871626
…     …      …         …           …
1460  48363    0.000           3     0.321952
1461  48370    0.027          19     3.151130
1462  48375    0.000           7     0.000000
1463  48376    0.000           3     0.000000
1464  48395    0.220          14     2.040920

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

[1465 rows x 13 columns]
```

[41]:
```
data_risky.loc[data_risky['is_risky']]
```

[41]:
```
       id_x  budget  popularity      revenue  \
7     43608   200.0         107   586.090727
12    43614   380.0         135  1045.713802
15    43618   200.0          37   310.669540
20    43624   209.0          64   303.025485
24    43630   210.0           3   459.359555
…     …     …          …           …
1347  47224     5.0           7     3.263585
1349  47229     5.0           3     4.842699
1351  47233     5.0           6     0.000000
1356  47263    15.0          10     0.000000
1383  47453     3.5           4     0.000000

                                          title  vote_average  vote_count  \
7                             Quantum of Solace           6.1        2965
12    Pirates of the Caribbean: On Stranger Tides           6.4        4948
15                                    Robin Hood           6.2        1398
20                                    Battleship           5.5        2114
24                          X-Men: The Last Stand           6.3        3525
…                                             …           …          …
1347                        The Sweet Hereafter           6.8         103
1349                      90 Minutes in Heaven            5.4          40
1351                             Light Sleeper           5.7          15
1356                       Dying of the Light            4.5         118
1383             In the Name of the King III           3.3          19

      year month       day   director_name gender  is_risky
7     2008   Oct  Thursday    Marc Forster   Male      True
12    2011   May  Saturday    Rob Marshall   Male      True
```

```
15    2010   May  Wednesday    Ridley Scott   Male      True
20    2012   Apr  Wednesday      Peter Berg   Male      True
24    2006   May  Wednesday    Brett Ratner   Male      True
…     …    …        …              …     …        …
1347  1997   May  Wednesday     Atom Egoyan   Male      True
1349  2015   Sep     Friday  Michael Polish   Male      True
1351  1992   Aug     Friday   Paul Schrader    NaN      True
1356  2014   Dec   Thursday   Paul Schrader    NaN      True
1383  2013   Dec     Friday        Uwe Boll   Male      True

[131 rows x 13 columns]
```

[42]:
```
data['director_name'].value_counts()
```

[42]:
```
Steven Spielberg      26
Martin Scorsese       19
Clint Eastwood        19
Woody Allen           18
Ridley Scott          16
                      ..
Tim Hill               5
Jonathan Liebesman     5
Roman Polanski         5
Larry Charles          5
Nicole Holofcener      5
Name: director_name, Length: 199, dtype: int64
```

[44]:
```
data_agg = data.groupby('director_name')[['title', 'year']].aggregate({'title':␣
↪'count', 'year':['min', 'max']})
data_agg
```

[44]:
```
                           title  year
                           count   min   max
director_name
Adam McKay                     6  2004  2015
Adam Shankman                  8  2001  2012
Alejandro González Iñárritu    6  2000  2015
Alex Proyas                    5  1994  2016
Alexander Payne                5  1999  2013
…                            …    …     …
Wes Craven                    10  1984  2011
Wolfgang Petersen              7  1981  2006
Woody Allen                   18  1977  2013
Zack Snyder                    7  2004  2016
Zhang Yimou                    6  2002  2014

[199 rows x 3 columns]
```

[45]:
```
data.columns
```

[45]:
```
Index(['id_x', 'budget', 'popularity', 'revenue', 'title', 'vote_average',
       'vote_count', 'year', 'month', 'day', 'director_name', 'gender'],
      dtype='object')
```

[46]:
```
data_agg.columns
```

[46]:
```
MultiIndex([('title', 'count'),
            ( 'year',   'min'),
            ( 'year',   'max')],
           )
```

[47]:
```
data_agg['year']
```

[47]:

director_name	min	max
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]