# IMDB Case Study using Pandas

In [ ]:

```
import numpy as np
import pandas as pd
movies=pd.read_csv('movies.csv')
directors=pd.read_csv('directors.csv')
```

In [ ]:

```
movies.head()
```

Out[ ]:

| | Unnamed: 0 | id | budget | popularity | revenue | title | vote_average | vote_count | director_id | year | month | day |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 43597 | 237000000 | 150 | 2787965087 | Avatar | 7.2 | 11800 | 4762 | 2009 | Dec | Thursday |
| **1** | 1 | 43598 | 300000000 | 139 | 961000000 | Pirates of the Caribbean: At World's End | 6.9 | 4500 | 4763 | 2007 | May | Saturday |
| **2** | 2 | 43599 | 245000000 | 107 | 880674609 | Spectre | 6.3 | 4466 | 4764 | 2015 | Oct | Monday |
| **3** | 3 | 43600 | 250000000 | 112 | 1084939099 | The Dark Knight Rises | 7.6 | 9106 | 4765 | 2012 | Jul | Monday |
| **4** | 5 | 43602 | 258000000 | 115 | 890871626 | Spider-Man 3 | 5.9 | 3576 | 4767 | 2007 | May | Tuesday |

In [ ]:

```
directors.head()
```

Out[ ]:

| | Unnamed: 0 | director_name | id | gender |
|---|---|---|---|---|
| **0** | 0 | James Cameron | 4762 | Male |
| **1** | 1 | Gore Verbinski | 4763 | Male |
| **2** | 2 | Sam Mendes | 4764 | Male |
| **3** | 3 | Christopher Nolan | 4765 | Male |
| **4** | 4 | Andrew Stanton | 4766 | Male |

In [ ]:

```
df=movies.merge(directors,left_on='director_id',right_on='id',how='inner')
```

In [ ]:

```
df.head()
```

Out[ ]:

| | Unnamed: 0_x | id_x | budget | popularity | revenue | title | vote_average | vote_count | director_id | year | month | day |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 43597 | 237000000 | 150 | 2787965087 | Avatar | 7.2 | 11800 | 4762 | 2009 | Dec | Thursday |
| **1** | 1 | 43598 | 300000000 | 139 | 961000000 | Pirates of the Caribbean: At World's End | 6.9 | 4500 | 4763 | 2007 | May | Saturday |
| **2** | 2 | 43599 | 245000000 | 107 | 880674609 | Spectre | 6.3 | 4466 | 4764 | 2015 | Oct | Monday |
| **3** | 3 | 43600 | 250000000 | 112 | 1084939099 | The Dark Knight Rises | 7.6 | 9106 | 4765 | 2012 | Jul | Monday |
| **4** | 5 | 43602 | 258000000 | 115 | 890871626 | Spider-Man 3 | 5.9 | 3576 | 4767 | 2007 | May | Tuesday |

In [ ]:

```
df.drop(['director_id','id_y','id_x'],axis=1,inplace=True)
```

In [ ]:

```
df.head()
```

Out[ ]:

|   | Unnamed: 0_x | budget | popularity | revenue | title | vote_average | vote_count | year | month | day | Unnamed: 0_y | directo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 237000000 | 150 | 2787965087 | Avatar | 7.2 | 11800 | 2009 | Dec | Thursday | 0 | C |
| **1** | 1 | 300000000 | 139 | 961000000 | Pirates of the Caribbean: At World's End | 6.9 | 4500 | 2007 | May | Saturday | 1 | Gore V |
| **2** | 2 | 245000000 | 107 | 880674609 | Spectre | 6.3 | 4466 | 2015 | Oct | Monday | 2 | Sam |
| **3** | 3 | 250000000 | 112 | 1084939099 | The Dark Knight Rises | 7.6 | 9106 | 2012 | Jul | Monday | 3 | Chr |
| **4** | 5 | 258000000 | 115 | 890871626 | Spider-Man 3 | 5.9 | 3576 | 2007 | May | Tuesday | 5 | Sa |

In [ ]:

```
df['budget']=df['budget']/10000000
```

In [ ]:

```
df.head()
```

Out[ ]:

|   | Unnamed: 0_x | budget | popularity | revenue | title | vote_average | vote_count | year | month | day | Unnamed: 0_y | director_n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 23.7 | 150 | 2787965087 | Avatar | 7.2 | 11800 | 2009 | Dec | Thursday | 0 | Ja Cam |
| **1** | 1 | 30.0 | 139 | 961000000 | Pirates of the Caribbean: At World's End | 6.9 | 4500 | 2007 | May | Saturday | 1 | Gore Verb |
| **2** | 2 | 24.5 | 107 | 880674609 | Spectre | 6.3 | 4466 | 2015 | Oct | Monday | 2 | Sam Me |
| **3** | 3 | 25.0 | 112 | 1084939099 | The Dark Knight Rises | 7.6 | 9106 | 2012 | Jul | Monday | 3 | Christo N |
| **4** | 5 | 25.8 | 115 | 890871626 | Spider-Man 3 | 5.9 | 3576 | 2007 | May | Tuesday | 5 | Sam F |

In [ ]:

```
df['revenue']=np.round(df['revenue']/1000000,2)
```

In [ ]:

```
df.head()
```

Out[ ]:

|   | Unnamed: 0_x | budget | popularity | revenue | title | vote_average | vote_count | year | month | day | Unnamed: 0_y | director_nam |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 23.7 | 150 | 2787.97 | Avatar | 7.2 | 11800 | 2009 | Dec | Thursday | 0 | James Cameron |
| **1** | 1 | 30.0 | 139 | 961.00 | Pirates of the Caribbean: At World's End | 6.9 | 4500 | 2007 | May | Saturday | 1 | Gore Verbinsk |
| **2** | 2 | 24.5 | 107 | 880.67 | Spectre | 6.3 | 4466 | 2015 | Oct | Monday | 2 | Sam Mende |
| **3** | 3 | 25.0 | 112 | 1084.94 | The Dark Knight Rises | 7.6 | 9106 | 2012 | Jul | Monday | 3 | Christophe Nolar |
| **4** | 5 | 25.8 | 115 | 890.87 | Spider-Man 3 | 5.9 | 3576 | 2007 | May | Tuesday | 5 | Sam Raim |

In [ ]:

```
df.shape
```

Out[ ]:

```
(1465, 13)
```

In [ ]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1465 entries, 0 to 1464
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Unnamed: 0_x   1465 non-null   int64  
 1   budget         1465 non-null   float64
 2   popularity     1465 non-null   int64  
 3   revenue        1465 non-null   float64
 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  Unnamed: 0_y   1465 non-null   int64  
 11  director_name  1465 non-null   object 
 12  gender         1341 non-null   object 
dtypes: float64(3), int64(5), object(5)
memory usage: 148.9+ KB
```

In [ ]:

```
df.describe(include='number')
```

Out[ ]:

| | Unnamed: 0_x | budget | popularity | revenue | vote_average | vote_count | year | Unnamed: 0_y |
|---|---|---|---|---|---|---|---|---|
| **count** | 1465.000000 | 1465.000000 | 1465.000000 | 1465.000000 | 1465.000000 | 1465.000000 | 1465.000000 | 1465.000000 |
| **mean** | 1627.391809 | 4.802295 | 30.855973 | 143.253952 | 6.368191 | 1146.396587 | 2002.615017 | 278.192491 |
| **std** | 1187.182894 | 4.935541 | 34.845214 | 206.491831 | 0.818033 | 1578.077438 | 8.680141 | 258.059631 |
| **min** | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 3.000000 | 1.000000 | 1976.000000 | 0.000000 |
| **25%** | 639.000000 | 1.400000 | 11.000000 | 17.380000 | 5.900000 | 216.000000 | 1998.000000 | 83.000000 |
| **50%** | 1425.000000 | 3.300000 | 23.000000 | 75.780000 | 6.400000 | 571.000000 | 2004.000000 | 202.000000 |
| **75%** | 2393.000000 | 6.600000 | 41.000000 | 179.250000 | 6.900000 | 1387.000000 | 2009.000000 | 417.000000 |
| **max** | 4768.000000 | 38.000000 | 724.000000 | 2787.970000 | 8.300000 | 13752.000000 | 2016.000000 | 1442.000000 |

In [ ]:

```
df.describe(include='object')
```

Out[ ]:

| | title | month | day | director_name | gender |
|---|---|---|---|---|---|
| **count** | 1465 | 1465 | 1465 | 1465 | 1341 |
| **unique** | 1465 | 12 | 7 | 199 | 2 |
| **top** | El Mariachi | Dec | Friday | Steven Spielberg | Male |
| **freq** | 1 | 193 | 654 | 26 | 1309 |

In [ ]:

```
#find the gender wise vote_Average
df.groupby('gender')['vote_average'].mean()
```

Out[ ]:

| | vote_average |
|---|---|
| **gender** | |
| **Female** | 6.262500 |
| **Male** | 6.381742 |

**dtype:** float64

In [ ]:

```
df.groupby('director_name')['title'].count()
```

Out[ ]:

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

199 rows × 1 columns


dtype: int64
```

In [ ]:

```
df.groupby('director_name')['title'].count().sort_values(ascending=False)
```

Out[ ]:

```
                        title
         director_name
     Steven Spielberg      26
      Martin Scorsese      19
       Clint Eastwood      19
          Woody Allen      18
     Robert Rodriguez      16
                  ...     ...
       Stephen Daldry       5
           Tom Tykwer       5
             Tim Hill       5
             Uwe Boll       5
           Wayne Wang       5

199 rows × 1 columns


dtype: int64
```

In [ ]:

```
df.groupby('director_name').get_group('Steven Spielberg')
```

Out[ ]:

```
      Unnamed:  budget  popularity  revenue         title  vote_average  vote_count  year  month        day  Unnamed:  directo
          0_x                                                                                                       0_y
                                                  Indiana
                                                Jones and
                                                      the
 37         53   18.50          75   786.64   Kingdom of           5.7        2495  2008    May  Wednesday        37
                                                the Crystal                                                                 Sp
                                                      S...
105        175   14.00          44   183.35      The BFG           6.0        1000  2016    Jun  Wednesday        37
                                                                                                                            Sp
                                                War of the
110        185   13.20          48   591.74       Worlds           6.2        2322  2005    Jun    Tuesday        37
                                                                                                                            Sp
                                                      The
114        190   13.00          89   371.94   Adventures           6.7        2061  2011    Oct    Tuesday        37
                                                of Tintin                                                                   Sp
                                                 Minority
166        275   10.20          65   358.37       Report           7.1        2608  2002    Jun   Thursday        37
                                                                                                                            Sp
                                                     A.I.
219        363   10.00          34   235.93     Artificial           6.8        1974  2001    Jun     Friday        37
                                             Intelligence                                                                   Sp

                                                 The Lost
```

|  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **296** | 508 | 7.30 | 2 | 229.07 | World: Jurassic Park | 6.2 | 2487 | 1997 | May | Friday | 37 | Sp |
| **304** | 521 | 6.00 | 57 | 219.42 | The Terminal | 7.0 | 1910 | 2004 | Jun | Thursday | 37 | Sp |
| **309** | 528 | 7.00 | 29 | 130.36 | Munich | 6.9 | 696 | 2005 | Dec | Thursday | 37 | Sp |
| **333** | 572 | 7.00 | 33 | 300.85 | Hook | 6.6 | 1532 | 1991 | Dec | Wednesday | 37 | Sp |
| **342** | 585 | 6.60 | 29 | 177.58 | War Horse | 7.0 | 992 | 2011 | Dec | Sunday | 37 | Sp |
| **360** | 628 | 7.00 | 76 | 481.84 | Saving Private Ryan | 7.9 | 5048 | 1998 | Jul | Friday | 37 | Sp |
| **363** | 633 | 6.50 | 36 | 275.29 | Lincoln | 6.7 | 1429 | 2012 | Nov | Friday | 37 | Sp |
| **390** | 675 | 6.30 | 40 | 920.10 | Jurassic Park | 7.6 | 4856 | 1993 | Jun | Friday | 37 | Sp |
| **495** | 883 | 5.20 | 73 | 352.11 | Catch Me If You Can | 7.7 | 3795 | 2002 | Dec | Wednesday | 37 | Sp |
| **552** | 1006 | 4.80 | 80 | 474.17 | Indiana Jones and the Last Crusade | 7.6 | 3152 | 1989 | May | Wednesday | 37 | Sp |
| **635** | 1187 | 4.00 | 48 | 165.48 | Bridge of Spies | 7.2 | 2583 | 2015 | Oct | Thursday | 37 | Sp |
| **647** | 1211 | 3.60 | 3 | 74.00 | Amistad | 6.8 | 316 | 1997 | Dec | Wednesday | 37 | Sp |
| **766** | 1510 | 3.50 | 10 | 31.76 | 1941 | 5.6 | 143 | 1979 | Dec | Thursday | 37 | Sp |
| **834** | 1697 | 2.80 | 66 | 333.00 | Indiana Jones and the Temple of Doom | 7.1 | 2781 | 1984 | May | Wednesday | 37 | Sp |
| **901** | 1818 | 2.20 | 104 | 321.37 | Schindler's List | 8.3 | 4329 | 1993 | Nov | Monday | 37 | Sp |
| **995** | 2085 | 1.80 | 68 | 389.93 | Raiders of the Lost Ark | 7.7 | 3854 | 1981 | Jun | Friday | 37 | Sp |
| **997** | 2087 | 2.00 | 52 | 303.79 | Close Encounters of the Third Kind | 7.2 | 1098 | 1977 | Nov | Wednesday | 37 | Sp |
| **1135** | 2520 | 1.50 | 17 | 146.29 | The Color Purple | 7.7 | 338 | 1985 | Dec | Wednesday | 37 | Sp |
| **1239** | 2967 | 1.05 | 56 | 792.91 | E.T. the Extra-Terrestrial | 7.3 | 3269 | 1982 | Apr | Saturday | 37 | Sp |
| **1252** | 3006 | 1.00 | 12 | 29.45 | Twilight Zone: The Movie | 6.2 | 161 | 1983 | Jun | Friday | 37 | Sp |

In [ ]:

```
df.groupby('director_name').groups
```

Out[ ]:

```
{'Adam McKay': [176, 323, 366, 505, 839, 916], 'Adam Shankman': [265, 300, 350, 404, 458, 843, 999, 1231], 'A
lejandro González Iñárritu': [106, 749, 1015, 1034, 1077, 1405], 'Alex Proyas': [95, 159, 514, 671, 873], 'Al
exander Payne': [793, 1006, 1101, 1211, 1281], 'Andrew Adamson': [11, 43, 328, 501, 947], 'Andrew Niccol': [5
33, 603, 701, 722, 1439], 'Andrzej Bartkowiak': [349, 549, 754, 911, 924], 'Andy Fickman': [517, 681, 909, 92
6, 973, 1023], 'Andy Tennant': [314, 320, 464, 593, 676, 885], 'Ang Lee': [99, 134, 748, 840, 1089, 1110, 113
2, 1184], 'Anne Fletcher': [610, 650, 736, 789, 1206], 'Antoine Fuqua': [310, 338, 424, 467, 576, 808, 818, 1
105], 'Atom Egoyan': [946, 1128, 1164, 1194, 1347, 1416], 'Barry Levinson': [313, 319, 471, 594, 878, 898, 10
13, 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], 'Camer
on Crowe': [335, 434, 488, 503, 513, 698], 'Catherine Hardwicke': [602, 695, 724, 937, 1406, 1412], 'Chris Co
lumbus': [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, 116
7, 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 Lync
h': [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, 26
7, 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': [30
8, 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], 'Fra
nk Oz': [193, 355, 473, 580, 712, 813, 987], 'Garry Marshall': [329, 496, 528, 571, 784, 893, 1029, 1169], 'G
ary Fleder': [518, 667, 689, 867, 981, 1165], 'Gary Winick': [258, 797, 798, 804, 1454], 'Gavin O'Connor': [8
20, 841, 939, 953, 1444], 'George A. Romero': [250, 1066, 1096, 1278, 1367, 1396], 'George Clooney': [343, 45
0, 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, 124
0, 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': [15
5, 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, 7
03, 859, 1072], 'Jean-Pierre Jeunet': [423, 485, 605, 664, 765], 'Joe Dante': [284, 525, 638, 1226, 1298, 142
8], '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 C
arpenter': [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': [1
27, 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, 11
17, 1301], 'Judd Apatow': [321, 710, 717, 865, 881], 'Justin Lin': [38, 123, 246, 1437, 1447], 'Kenneth Brana
gh': [80, 197, 421, 879, 1094, 1277, 1288], 'Kenny Ortega': [412, 852, 1228, 1315, 1365], 'Kevin Reynolds': [
53, 502, 639, 1019, 1059], ...}
```

In [ ]:

```
df.groupby('director_name').ngroups
```

Out[ ]:

```
199
```

In [ ]:

```
#which director is more productive
final_data=df.groupby('director_name').aggregate({'year':['min','max'],'title':['count']})
```

In [ ]:

```
final_data
```

Out[ ]:

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

199 rows × 3 columns

In [ ]:

```
final_data.sort_values(by=('title','count'),ascending=False,inplace=True)
```

In [ ]:

```
final_data
```

Out[ ]:

| | year | | title |
|---|---|---|---|
| | min | max | count |
| director_name | | | |
| Steven Spielberg | 1977 | 2016 | 26 |
| Martin Scorsese | 1976 | 2013 | 19 |
| Clint Eastwood | 1982 | 2014 | 19 |
| Woody Allen | 1977 | 2013 | 18 |
| Robert Rodriguez | 1992 | 2014 | 16 |
| ... | ... | ... | ... |
| Stephen Daldry | 2000 | 2014 | 5 |
| Tom Tykwer | 1998 | 2012 | 5 |
| Tim Hill | 1999 | 2011 | 5 |
| Uwe Boll | 2005 | 2013 | 5 |
| Wayne Wang | 1999 | 2011 | 5 |

199 rows × 3 columns

In [ ]:

```
final_data['year_active']=final_data[('year','max')]-final_data[('year','min')]
final_data
```

Out[ ]:

|  | year |  | title | year_active |
|---|---|---|---|---|
|  | min | max | count |  |
| director_name |  |  |  |  |
| Steven Spielberg | 1977 | 2016 | 26 | 39 |
| Martin Scorsese | 1976 | 2013 | 19 | 37 |
| Clint Eastwood | 1982 | 2014 | 19 | 32 |
| Woody Allen | 1977 | 2013 | 18 | 36 |
| Robert Rodriguez | 1992 | 2014 | 16 | 22 |
| ... | ... | ... | ... | ... |
| Stephen Daldry | 2000 | 2014 | 5 | 14 |
| Tom Tykwer | 1998 | 2012 | 5 | 14 |
| Tim Hill | 1999 | 2011 | 5 | 12 |
| Uwe Boll | 2005 | 2013 | 5 | 8 |
| Wayne Wang | 1999 | 2011 | 5 | 12 |

199 rows × 4 columns

In [ ]:

```
final_data['productivity']=final_data['title','count']/final_data['year_active']
final_data['productivity']=final_data['productivity']*100
final_data
```

Out[ ]:

|  | year |  | title | year_active | productivity |
|---|---|---|---|---|---|
|  | min | max | count |  |  |
| director_name |  |  |  |  |  |
| Steven Spielberg | 1977 | 2016 | 26 | 39 | 66.666667 |
| Martin Scorsese | 1976 | 2013 | 19 | 37 | 51.351351 |
| Clint Eastwood | 1982 | 2014 | 19 | 32 | 59.375000 |
| Woody Allen | 1977 | 2013 | 18 | 36 | 50.000000 |
| Robert Rodriguez | 1992 | 2014 | 16 | 22 | 72.727273 |
| ... | ... | ... | ... | ... | ... |
| Stephen Daldry | 2000 | 2014 | 5 | 14 | 35.714286 |
| Tom Tykwer | 1998 | 2012 | 5 | 14 | 35.714286 |
| Tim Hill | 1999 | 2011 | 5 | 12 | 41.666667 |
| Uwe Boll | 2005 | 2013 | 5 | 8 | 62.500000 |
| Wayne Wang | 1999 | 2011 | 5 | 12 | 41.666667 |

199 rows × 5 columns

In [ ]:

```
final_data.sort_values('productivity',ascending=False,inplace=True)
```

In [ ]:

```
final_data
```

Out[ ]:

| | year | | title | year_active | productivity |
|---|---|---|---|---|---|
| | min | max | count | | |
| director_name | | | | | |
| Tyler Perry | 2006 | 2013 | 9 | 7 | 128.571429 |
| Jason Friedberg | 2006 | 2010 | 5 | 4 | 125.000000 |
| Shawn Levy | 2002 | 2014 | 11 | 12 | 91.666667 |
| Adam Shankman | 2001 | 2012 | 8 | 11 | 72.727273 |
| Robert Rodriguez | 1992 | 2014 | 16 | 22 | 72.727273 |
| ... | ... | ... | ... | ... | ... |
| Lawrence Kasdan | 1985 | 2012 | 5 | 27 | 18.518519 |
| Luc Besson | 1985 | 2014 | 5 | 29 | 17.241379 |
| Michael Apted | 1980 | 2010 | 5 | 30 | 16.666667 |
| Robert Redford | 1980 | 2010 | 5 | 30 | 16.666667 |
| Sidney Lumet | 1976 | 2006 | 5 | 30 | 16.666667 |

199 rows × 5 columns

In [ ]:
```
df
```

Out[ ]:

| | Unnamed: 0_x | budget | popularity | revenue | title | vote_average | vote_count | year | month | day | Unnamed: 0_y | director_r |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 23.7000 | 150 | 2787.97 | Avatar | 7.2 | 11800 | 2009 | Dec | Thursday | 0 | Ja Can |
| 1 | 1 | 30.0000 | 139 | 961.00 | Pirates of the Caribbean: At World's End | 6.9 | 4500 | 2007 | May | Saturday | 1 | Gore Verb |
| 2 | 2 | 24.5000 | 107 | 880.67 | Spectre | 6.3 | 4466 | 2015 | Oct | Monday | 2 | Sam Me |
| 3 | 3 | 25.0000 | 112 | 1084.94 | The Dark Knight Rises | 7.6 | 9106 | 2012 | Jul | Monday | 3 | Christo N |
| 4 | 5 | 25.8000 | 115 | 890.87 | Spider-Man 3 | 5.9 | 3576 | 2007 | May | Tuesday | 5 | Sam R |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 1460 | 4736 | 0.0000 | 3 | 0.32 | The Last Waltz | 7.9 | 64 | 1978 | May | Monday | 47 | M Scor |
| 1461 | 4743 | 0.0027 | 19 | 3.15 | Clerks | 7.4 | 755 | 1994 | Sep | Tuesday | 607 | Kevin S |
| 1462 | 4748 | 0.0000 | 7 | 0.00 | Rampage | 6.0 | 131 | 2009 | Aug | Friday | 386 | Uwe |
| 1463 | 4749 | 0.0000 | 3 | 0.00 | Slacker | 6.4 | 77 | 1990 | Jul | Friday | 773 | Ric Link |
| 1464 | 4768 | 0.0220 | 14 | 2.04 | El Mariachi | 6.6 | 238 | 1992 | Sep | Friday | 335 | R Rodr |

1465 rows × 13 columns

In [ ]:
```
df.groupby('director_name').get_group('Tyler Perry')
```

Out[ ]:

| | Unnamed: 0_x | budget | popularity | revenue | title | vote_average | vote_count | year | month | day | Unnamed: 0_y | director_na |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **915** | 1843 | 0.0 | 5 | 0.00 | A Madea Christmas | 7.0 | 35 | 2013 | Dec | Friday | 792 | Tyler P |
| **991** | 2073 | 0.0 | 1 | 37.00 | For Colored Girls | 7.0 | 22 | 2010 | Nov | Friday | 792 | Tyler P |
| **1007** | 2107 | 2.0 | 2 | 60.07 | Why Did I Get Married Too? | 6.1 | 29 | 2010 | Apr | Friday | 792 | Tyler P |
| **1009** | 2110 | 0.0 | 7 | 0.00 | Madea's Witness Protection | 5.9 | 52 | 2012 | Jun | Friday | 792 | Tyler P |
| **1062** | 2287 | 0.0 | 2 | 0.00 | I Can Do Bad All By Myself | 6.0 | 40 | 2009 | Sep | Friday | 792 | Tyler P |
| **1098** | 2393 | 0.0 | 3 | 90.51 | Madea Goes to Jail | 6.4 | 52 | 2009 | Feb | Monday | 792 | Tyler P |
| **1140** | 2531 | 1.5 | 2 | 55.18 | Why Did I Get Married? | 6.1 | 33 | 2007 | Oct | Friday | 792 | Tyler P |
| **1172** | 2677 | 0.0 | 4 | 0.00 | Good Deeds | 6.2 | 45 | 2012 | Feb | Thursday | 792 | Tyler P |
| **1244** | 2984 | 0.6 | 5 | 57.23 | Madea's Family Reunion | 6.0 | 77 | 2006 | Feb | Friday | 792 | Tyler P |

# Insights

- Tyler Perry is the most productive director
- Steven spielberg have director most number of movies

In [ ]:

```
#  1.Which month number of movies have been directed
df.groupby('month')['title'].count()
```

Out[ ]:

| | title |
|---|---|
| **month** | |
| **Apr** | 90 |
| **Aug** | 111 |
| **Dec** | 193 |
| **Feb** | 104 |
| **Jan** | 60 |
| **Jul** | 127 |
| **Jun** | 133 |
| **Mar** | 99 |
| **May** | 116 |
| **Nov** | 117 |
| **Oct** | 149 |
| **Sep** | 166 |

**dtype:** int64

# Insight 1:

December has the highest number of movies release(193) ,and least number of movies released in January,suggesting it is a peak month for the film industry,likely due to holiday seasons and increased audience engagement.

In [ ]:

```
# 2.Is there any luck day for movie release
df.groupby('day')['title'].count()
```

Out[ ]:

|  | title |
|---|---|
| **day** |  |
| **Friday** | 654 |
| **Monday** | 67 |
| **Saturday** | 47 |
| **Sunday** | 46 |
| **Thursday** | 277 |
| **Tuesday** | 98 |
| **Wednesday** | 276 |

**dtype:** int64

# Insight 2:

Friday is the most popular day for movie release (654), indicating a strategic choice to maximize weekened audience engagement

In [ ]:

```
# 3. Find out top 5 profitable movies
df[['title','revenue']].sort_values(by='revenue',ascending=False).head()
```

Out[ ]:

|  | title | revenue |
|---|---|---|
| **0** | Avatar | 2787.97 |
| **19** | Titanic | 1845.03 |
| **30** | Furious 7 | 1506.25 |
| **36** | Transformers: Dark of the Moon | 1123.75 |
| **199** | The Lord of the Rings: The Return of the King | 1118.89 |

# Insights 3:

The top 5 Movies made a highest revenue top 1 is **'Avatar**' Generated **27billion** Revenue and top 5 is **'The Lord of the rings:the return of the king'** Generated **11billion** revenue

In [ ]:

```
df
```

Out[ ]:

| | Unnamed: 0_x | budget | popularity | revenue | title | vote_average | vote_count | year | month | day | Unnamed: 0_y | director_r |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 23.7000 | 150 | 2787.97 | Avatar | 7.2 | 11800 | 2009 | Dec | Thursday | 0 | Ja Can |
| **1** | 1 | 30.0000 | 139 | 961.00 | Pirates of the Caribbean: At World's End | 6.9 | 4500 | 2007 | May | Saturday | 1 | Gore Verb |
| **2** | 2 | 24.5000 | 107 | 880.67 | Spectre | 6.3 | 4466 | 2015 | Oct | Monday | 2 | Sam Me |
| **3** | 3 | 25.0000 | 112 | 1084.94 | The Dark Knight Rises | 7.6 | 9106 | 2012 | Jul | Monday | 3 | Christo N |
| **4** | 5 | 25.8000 | 115 | 890.87 | Spider-Man 3 | 5.9 | 3576 | 2007 | May | Tuesday | 5 | Sam R |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **1460** | 4736 | 0.0000 | 3 | 0.32 | The Last Waltz | 7.9 | 64 | 1978 | May | Monday | 47 | M Scor |
| **1461** | 4743 | 0.0027 | 19 | 3.15 | Clerks | 7.4 | 755 | 1994 | Sep | Tuesday | 607 | Kevin S |
| **1462** | 4748 | 0.0000 | 7 | 0.00 | Rampage | 6.0 | 131 | 2009 | Aug | Friday | 386 | Uwe |
| **1463** | 4749 | 0.0000 | 3 | 0.00 | Slacker | 6.4 | 77 | 1990 | Jul | Friday | 773 | Ric Link |
| **1464** | 4768 | 0.0220 | 14 | 2.04 | El Mariachi | 6.6 | 238 | 1992 | Sep | Friday | 335 | R Rodr |

1465 rows × 13 columns

In [ ]:

```
#4.Director with highest vote_avg and popularity
df.groupby('director_name')[['vote_average','popularity']].max().head()
```

Out[ ]:

| | vote_average | popularity |
|---|---|---|
| **director_name** | | |
| **Adam McKay** | 7.3 | 57 |
| **Adam Shankman** | 7.5 | 35 |
| **Alejandro González Iñárritu** | 7.6 | 100 |
| **Alex Proyas** | 7.3 | 95 |
| **Alexander Payne** | 7.4 | 40 |

# Insight 4:

The top5 director with the highest vote average and popularity .Top1 director Adam Shankman vote average is 7.5 and popularity is 100 .

In [ ]:

```
#5.Top3 budget movies
df.sort_values('budget',ascending=False).head(3)
```

Out[ ]:

| | Unnamed: 0_x | budget | popularity | revenue | title | vote_average | vote_count | year | month | day | Unnamed: 0_y | director_n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **12** | 17 | 38.0 | 135 | 1045.71 | Pirates of the Caribbean: On Stranger Tides | 6.4 | 4948 | 2011 | May | Saturday | 13 | Rob Mar |
| **1** | 1 | 30.0 | 139 | 961.00 | Pirates of the Caribbean: At World's End | 6.9 | 4500 | 2007 | May | Saturday | 1 | Gore Verb |
| **6** | 10 | 27.0 | 57 | 391.08 | Superman Returns | 5.4 | 1400 | 2006 | Jun | Wednesday | 10 | Bryan Si |

# Insight 5:

The top 3 budget movies were directed by **Rob Marshall**,**Gore Verbinski** and **Bryan Singer**

In [ ]:

```
#6.director who have directed top budget movies
df.sort_values('budget',ascending=False)[['director_name','budget']].head()
```

Out[ ]:

| | director_name | budget |
|---|---|---|
| **12** | Rob Marshall | 38.0 |
| **1** | Gore Verbinski | 30.0 |
| **6** | Bryan Singer | 27.0 |
| **4** | Sam Raimi | 25.8 |
| **9** | Gore Verbinski | 25.5 |

# Insight 6:

**Rob Marshall** Directed a movie with the highest budget of **380 million**, **Gore Verbinski** Directed two movies with budgets over **$250 million**