Investigate_a_Dataset

February 3, 2019

1 Project: Investigating TMDb movie data

1.1 Table of Contents

Introduction
Data Wrangling
Exploratory Data Analysis
Conclusions
Limitations
Introduction

I have chosen to analyse the TMDb movie dataset as I really like movies. It contains data about 10,000 movies, including review and revenue data (the columns ending with "adj" show the 2010 values adjusted for inflation).

I am curious to know if runtime is correlated with budget. I might also see which Chris Pratt movies brought in the most revenue, because who doesn't love Chris Pratt?

```
In [35]: # Importing packages

         import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         pd.set_option('display.max_columns', None)
```

Data Wrangling

1.1.1 General Properties

```
In [36]: # Load data
         df = pd.read_csv("tmdb-movies.csv")
         df.head()
```

```
Out[36]:        id    imdb_id  popularity     budget     revenue  \
         0  135397  tt0369610   32.985763  150000000  1513528810
         1   76341  tt1392190   28.419936  150000000   378436354
         2  262500  tt2908446   13.112507  110000000   295238201
         3  140607  tt2488496   11.173104  200000000  2068178225
         4  168259  tt2820852    9.335014  190000000  1506249360
```

```
                original_title  \
0                Jurassic World
1            Mad Max: Fury Road
2                      Insurgent
3  Star Wars: The Force Awakens
4                      Furious 7

                                                 cast  \
0  Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...
1  Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...
2  Shailene Woodley|Theo James|Kate Winslet|Ansel...
3  Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...
4  Vin Diesel|Paul Walker|Jason Statham|Michelle ...

                                       homepage          director  \
0                    http://www.jurassicworld.com/   Colin Trevorrow
1                      http://www.madmaxmovie.com/     George Miller
2    http://www.thedivergentseries.movie/#insurgent  Robert Schwentke
3  http://www.starwars.com/films/star-wars-episod...       J.J. Abrams
4                          http://www.furious7.com/         James Wan

                        tagline  \
0              The park is open.
1             What a Lovely Day.
2     One Choice Can Destroy You
3  Every generation has a story.
4            Vengeance Hits Home

                                       keywords  \
0  monster|dna|tyrannosaurus rex|velociraptor|island
1   future|chase|post-apocalyptic|dystopia|australia
2  based on novel|revolution|dystopia|sequel|dyst...
3              android|spaceship|jedi|space opera|3d
4                car race|speed|revenge|suspense|car

                                       overview  runtime  \
0  Twenty-two years after the events of Jurassic ...      124
1  An apocalyptic story set in the furthest reach...      120
2  Beatrice Prior must confront her inner demons ...      119
3  Thirty years after defeating the Galactic Empi...      136
4  Deckard Shaw seeks revenge against Dominic Tor...      137

                                    genres  \
0  Action|Adventure|Science Fiction|Thriller
1  Action|Adventure|Science Fiction|Thriller
2         Adventure|Science Fiction|Thriller
3   Action|Adventure|Science Fiction|Fantasy
4                     Action|Crime|Thriller
```

```
                                    production_companies release_date  vote_count  \
0  Universal Studios|Amblin Entertainment|Legenda...       6/9/15        5562
1  Village Roadshow Pictures|Kennedy Miller Produ...      5/13/15        6185
2  Summit Entertainment|Mandeville Films|Red Wago...      3/18/15        2480
3          Lucasfilm|Truenorth Productions|Bad Robot     12/15/15        5292
4  Universal Pictures|Original Film|Media Rights ...       4/1/15        2947

   vote_average  release_year    budget_adj   revenue_adj
0           6.5          2015  1.379999e+08  1.392446e+09
1           7.1          2015  1.379999e+08  3.481613e+08
2           6.3          2015  1.012000e+08  2.716190e+08
3           7.5          2015  1.839999e+08  1.902723e+09
4           7.3          2015  1.747999e+08  1.385749e+09
```

In [37]:
```
# Let's see the type of each column and look for any weird data
df.describe()
```

Out[37]:

	id	popularity	budget	revenue	runtime
count	10866.000000	10866.000000	1.086600e+04	1.086600e+04	10866.000000
mean	66064.177434	0.646441	1.462570e+07	3.982332e+07	102.070863
std	92130.136561	1.000185	3.091321e+07	1.170035e+08	31.381405
min	5.000000	0.000065	0.000000e+00	0.000000e+00	0.000000
25%	10596.250000	0.207583	0.000000e+00	0.000000e+00	90.000000
50%	20669.000000	0.383856	0.000000e+00	0.000000e+00	99.000000
75%	75610.000000	0.713817	1.500000e+07	2.400000e+07	111.000000
max	417859.000000	32.985763	4.250000e+08	2.781506e+09	900.000000

	vote_count	vote_average	release_year	budget_adj	revenue_adj
count	10866.000000	10866.000000	10866.000000	1.086600e+04	1.086600e+04
mean	217.389748	5.974922	2001.322658	1.755104e+07	5.136436e+07
std	575.619058	0.935142	12.812941	3.430616e+07	1.446325e+08
min	10.000000	1.500000	1960.000000	0.000000e+00	0.000000e+00
25%	17.000000	5.400000	1995.000000	0.000000e+00	0.000000e+00
50%	38.000000	6.000000	2006.000000	0.000000e+00	0.000000e+00
75%	145.750000	6.600000	2011.000000	2.085325e+07	3.369710e+07
max	9767.000000	9.200000	2015.000000	4.250000e+08	2.827124e+09

Seems like there's some 0's in the budget and revenue that we have to be careful about. Year looks ok. Runtime 0 doesn't make much sense, I wonder what that means.

In [38]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 21 columns):
id                      10866 non-null int64
imdb_id                 10856 non-null object
popularity              10866 non-null float64
```

```
budget                  10866 non-null int64
revenue                 10866 non-null int64
original_title          10866 non-null object
cast                    10790 non-null object
homepage                2936 non-null object
director                10822 non-null object
tagline                 8042 non-null object
keywords                9373 non-null object
overview                10862 non-null object
runtime                 10866 non-null int64
genres                  10843 non-null object
production_companies    9836 non-null object
release_date            10866 non-null object
vote_count              10866 non-null int64
vote_average            10866 non-null float64
release_year            10866 non-null int64
budget_adj              10866 non-null float64
revenue_adj             10866 non-null float64
dtypes: float64(4), int64(6), object(11)
memory usage: 1.7+ MB
```

Most of the missing data seems to be in homepage, tagline, keywords and production companies, with some missing data in director, cast and production_companies. I will take a look at the rows with missing production companies since i want to look at that column.

```
In [39]: sum(df.duplicated())
```

```
Out[39]: 1
```

One duplicate so let's just drop it

```
In [40]: df.drop_duplicates(inplace = True)
```

1.1.2 Data Cleaning

Let's take a look at those null production companies

```
In [41]: df[df['production_companies'].isnull()].head(10)
```

```
Out[41]:          id    imdb_id  popularity   budget  revenue  \
         228  300792  tt1618448    0.584363        0        0
         259  360603  tt5133572    0.476341        0        0
         295  363483  tt5133810    0.417191        0        0
         298  354220  tt3826866    0.370258        0        0
         328  308457  tt3090670    0.367617        0        0
         370  318279  tt2545428    0.314199        0  2334228
         374  206197  tt1015471    0.302474        0        0
         382  306197  tt4145304    0.295946        0        0
```

```
388  323967  tt2016335    0.289526   700000        0
393  343284  tt3602128    0.283194  2000000        0

                original_title  \
228            Racing Extinction
259          Crown for Christmas
295        12 Gifts of Christmas
298  The Girl in the Photographs
328                Advantageous
370                       Meru
374      The Sisterhood of Night
382                  Unexpected
388                      Walter
393      Night Of The Living Deb

                                                    cast  \
228  Elon Musk|Jane Goodall|Louie Psihoyos|Leilani ...
259  Danica McKellar|Rupert Penry-Jones|Ellie Botte...
295  Katrina Law|Donna Mills|Aaron O'Connell|Melani...
298  Kal Penn|Claudia Lee|Kenny Wormald|Toby Heming...
328  Jacqueline Kim|James Urbaniak|Freya Adams|Ken ...
370    Conrad Anker|Grace Chin|Jimmy Chin|Amee Hinkley
374  Kara Hayward|Georgie Henley|Olivia DeJonge|Lau...
382  Cobie Smulders|Gail Bean|Anders Holm|Elizabeth...
388  Andrew J. West|Justin Kirk|Virginia Madsen|Wil...
393  Maria Thayer|Chris Marquette|Ray Wise|Michael ...

                                      homepage  \
228            http://www.racingextinction.com
259                                        NaN
295                                        NaN
298                                        NaN
328                                        NaN
370                   http://www.merufilm.com/
374  http://www.thesisterhoodofnight-movie.com/
382                                        NaN
388                                        NaN
393                                        NaN

                             director  \
228                     Louie Psihoyos
259                          Alex Zamm
295                     Peter Sullivan
298                         Nick Simon
328                     Jennifer Phang
370  Jimmy Chin|Elizabeth Chai Vasarhelyi
374                     Caryn Waechter
382                      Kris Swanberg
```

```
388                          Anna Mastro
393                          Kyle Rankin

                                                tagline  \
228  It's better to light one candle than curse the...
259                                                 NaN
295                                                 NaN
298                                                 NaN
328                                                 NaN
370                                                 NaN
374                    The Salem Witch Trials remixed.
382                    No one is ever really prepared.
388                 Heaven or hell. It's in his hands.
393                                                 NaN

                                              keywords  \
228     animal species|earth|scientist|globe|activist
259                                                 NaN
295                                           christmas
298    serial killer|tied feet|tied up while barefoot
328  identity|unemployment|dystopic future|woman di...
370  himalaya|climbing|india|mountaineering|woman d...
374                              witch|woman director
382                                    woman director
388                                god|woman director
393                                                 NaN

                                              overview  runtime  \
228  An unlikely team of activists and innovators h...       90
259  After getting fired from her job as a maid at ...       84
295  When Anna Parisi, an unemployed fine arts pain...       84
298  Images coming from the crimes committed by a d...       95
328  In a near-future city where soaring opulence o...       92
370  Meru is the electrifying story of three elite ...       89
374  When a teenage girl says she's the victim of a...      104
382  An inner-city high school teacher discovers sh...       90
388  A ticket-taker at the local cinema believes he...       87
393  After a one night stand Deb wakes up in the mi...       85

                          genres production_companies release_date  \
228         Adventure|Documentary                  NaN      1/24/15
259                      TV Movie                  NaN     11/27/15
295               Family|TV Movie                  NaN     11/26/15
298         Crime|Horror|Thriller                  NaN      9/14/15
328  Science Fiction|Drama|Family                  NaN      6/23/15
370         Adventure|Documentary                  NaN      1/25/15
374        Mystery|Drama|Thriller                  NaN      4/10/15
382                  Drama|Comedy                  NaN      7/24/15
```

```
388                  Drama|Comedy                    NaN      3/13/15
393                 Comedy|Horror                    NaN      8/29/15

     vote_count  vote_average  release_year    budget_adj    revenue_adj
228          36           7.8          2015  0.000000e+00  0.000000e+00
259          10           7.6          2015  0.000000e+00  0.000000e+00
295          12           6.3          2015  0.000000e+00  0.000000e+00
298          10           4.7          2015  0.000000e+00  0.000000e+00
328          29           6.4          2015  0.000000e+00  0.000000e+00
370          42           6.9          2015  0.000000e+00  2.147489e+06
374          25           6.6          2015  0.000000e+00  0.000000e+00
382          22           5.7          2015  0.000000e+00  0.000000e+00
388          12           5.2          2015  6.439997e+05  0.000000e+00
393          13           5.7          2015  1.839999e+06  0.000000e+00
```

Seems like a lot of the places with no production company have no budget or revenue. I'll first remove those and then let's see what we're left with (we'll need to remove them for the revenue investigation anyway)

```
In [42]: df = df.query("budget > 0 and revenue > 0")

In [43]: df[df['production_companies'].isnull()].head(20)

Out[43]:           id    imdb_id  popularity    budget   revenue  \
         1758   25183  tt1331064    0.118533   1000000   1296971
         1797   43937  tt0981042    0.182881   2500000    352810
         1800   30128  tt1220628    0.190162   7000000   1429299
         1871   42151  tt1489167    0.092519     31192     10000
         2303   56812  tt1572491    0.181532   7000000   3218666
         2782   19405  tt0265632    0.501163  10000000  44460850
         2805   50035  tt0258273    0.251798    250000   4186931
         2840   18734  tt0242587    0.185009    700000   1667192
         3047   13827  tt0976247    0.518011   6000000     69497
         3077   10188  tt1018785    0.451304  27000000  44352417
         3116   13191  tt0884224    0.374034   5000000   1296184
         3235   14070  tt1166100    0.215030   9100000  76000000
         3239   14301  tt1227926    0.352054    200000         3
         3752   65650  tt1582271    0.244803   6000000      5206
         3818   80379  tt2112999    0.331313    250000   1000000
         3853   62796  tt1831829    0.098896        20        15
         4062   37964  tt0297037    0.373221   8000000  27362712
         4352   14425  tt0110759    0.094568   9000000   4350774
         4889  126509  tt2247692    0.075043   2500000  33400000
         5636   80920  tt2012665    0.564100   5000000   1189612

                                    original_title  \
         1758                          Paper Heart
         1797                    Like Dandelion Dust
         1800          I Hope They Serve Beer in Hell
```

```
1871                            Down Terrace
2303                Balada triste de trompeta
2782                     Recess: School's Out
2805                         Lovely & Amazing
2840           L.I.E. Long Island Expressway
3047                            Surfer, Dude
3077  The Sisterhood of the Traveling Pants 2
3116                               War, Inc.
3235                                 Ghajini
3239           Dr. Horrible's Sing-Along Blog
3752                          The Good Doctor
3818   Louis C.K.: Live at the Beacon Theater
3853                           William & Kate
4062                              Brown Sugar
4352                                      PCU
4889                    2016: Obama's America
5636                               Repentance

                                                      cast  \
1758  Michael Cera|Charlyne Yi|Jake Johnson|Gill Sum...
1797  Mira Sorvino|Barry Pepper|Cole Hauser|Kate Lev...
1800  Matt Czuchry|Jesse Bradford|Marika Dominczyk|T...
1871  Robert Hill|Robin Hill|Julia Deakin|David Scha...
2303  Santiago Segura|Antonio de la Torre|RaÃžl ArÃ...
2782  Rickey D'Shon Collins|Jason Davis|Ashley Johns...
2805  Catherine Keener|Brenda Blethyn|Emily Mortimer...
2840  Paul Dano|Bruce Altman|Brian Cox|Billy Kay|Jam...
3047  Matthew McConaughey|Jeffrey Nordling|Willie Ne...
3077  Alexis Bledel|Amber Tamblyn|America Ferrera|Bl...
3116  John Cusack|Hilary Duff|Marisa Tomei|Joan Cusa...
3235  Aamir Khan|Asin Thottumkal|Jiah Khan|Pradeep R...
3239  Neil Patrick Harris|Nathan Fillion|Felicia Day...
3752  Orlando Bloom|Riley Keough|Taraji P. Henson|Ro...
3818                                        Louis C.K.
3853  Camilla Luddington|Nico Evers-Swindell|Samanth...
4062  Sanaa Lathan|Taye Diggs|Mos Def|Queen Latifah|...
4352  Jeremy Piven|Chris Young|Megan Ward|Jon Favrea...
4889                                                NaN
5636  Forest Whitaker|Anthony Mackie|Sanaa Lathan|Ni...

                                                  homepage  \
1758                                                   NaN
1797                                                   NaN
1800                                                   NaN
1871                    http://downterrace.blogspot.com/
2303         http://baladatristedetrompeta.blogspot.com/
2782                                                   NaN
2805                                                   NaN
```

```
2840  http://tartanvideo.com/film.asp?ProjectID={C66...
3047                 http://www.surferdudethemovie.com/
3077                                                NaN
3116      http://www.firstlookstudios.com/films/warinc/
3235                    http://www.rememberghajini.com/
3239                          http://www.drhorrible.com
3752                                                NaN
3818                           https://buy.louisck.net/
3853                                                NaN
4062                                                NaN
4352                                                NaN
4889                           http://2016themovie.com/
5636                                                NaN

                        director  \
1758            Nicholas Jasenovec
1797                      Jon Gunn
1800                     Bob Gosse
1871                   Ben Wheatley
2303            Ãlex de la Iglesia
2782                   Chuck Sheetz
2805             Nicole Holofcener
2840                Michael Cuesta
3047                  S.R. Bindler
3077                   Sanaa Hamri
3116                 Joshua Seftel
3235               A.R. Murugadoss
3239                   Joss Whedon
3752                    Lance Daly
3818                    Louis C.K.
3853                   Mark Rosman
4062                 Rick Famuyiwa
4352                  Hart Bochner
4889  Dinesh D'Souza|John Sullivan
5636               Philippe Caland

                                                 tagline  \
1758  A story about love that's taking on a life on ...
1797         Sometimes the greatest love is letting go.
1800                                                NaN
1871        You're only as good as the people you know.
2303                                                NaN
2782          Saving The World One Playground At A Time
2805                                                NaN
2840  On the Long Island Expressway there are lanes ...
3047                                                NaN
3077                    Some friends just fit together.
3116     When it comes to war... America means business
```

```
3235                                                 NaN
3239                      He has a Ph.D. in horribleness!
3752                                        Do no harm.
3818                                      Buy The Thing
3853  A Prince, an ordinary girl. And a very British...
4062      When did you first fall in love with hip-hop?
4352           Flunk âem if they canât take a joke.
4889             Love Him, Hate Him, You Don't Know Him
5636                Karma is Action, Vipaka is Reaction

                                                 keywords  \
1758          love|independent film|aftercreditsstinger
1797                      based on novel|independent film
1800  female nudity|based on novel|stripper|flop|bla...
1871                      murder|dark comedy|crime family
2303  militia|smashed head|mad|revenge motive|clown ...
2782  holiday|elementary school|friends|based on tv ...
2805                                    woman director
2840                                  independent film
3047                                     surfing|sport
3077  female friendship|best friend|summer vacation|...
3116                              hitman|political satire
3235                             tattoo|short-term memory
3239                                 musical|supervillain
3752                                                 NaN
3818                                               stand-up
3853      royal family|british|love|royalty|royal court
4062                                                 NaN
4352  mascot|beer keg|political correctness|ultimate...
4889                                                 NaN
5636                                                 NaN

                                                 overview  runtime  \
1758  Paper Heart follows Nick and Charlyne on a cro...        88
1797  A compelling drama that explores the different...       100
1800  Tucker decides to take an impromptu trip to ce...       106
1871  After serving jail time for a mysterious crime...        89
2303  The journey of Javier, the obese Sad Clown, st...       101
2782  Recess: School's Out is a 2001 animated film b...        83
2805  Self-esteem and insecurity are at the heart of...        91
2840  In this biting and disturbing coming-of-age ta...        97
3047  A wave twisting tale of a soul searching surfe...        85
3077  Four young women continue the journey toward a...       117
3116  War Inc. is set in the future, when the fictio...       106
3235  Sanjay a rich tycoon suffering from short term...       183
3239  Dr. Horrible, an aspiring supervillain with hi...        42
3752  Dr. Martin Blake, who has spent his life looki...        93
3818  Recorded November 10th, 2011 as part of the Ne...        62
```

```
3853  William &amp;amp; Kate is the first of two unrelat...       83
4062  Sidney is a writer who's just left her L.A. Ti...      109
4352  A high school senior visits college for the we...       79
4889  2016: Obama's America takes audiences on a gri...       87
5636  An earnest life-coach/author, Thomas Carter, i...       90

                                          genres production_companies  \
1758                      Comedy|Drama|Romance                    NaN
1797                             Drama|Family                    NaN
1800                             Comedy|Drama                    NaN
1871                      Drama|Action|Comedy                    NaN
2303            Drama|Action|Thriller|Foreign                    NaN
2782                  Animation|Comedy|Family                    NaN
2805                     Comedy|Drama|Romance                    NaN
2840                                    Drama                    NaN
3047                                   Comedy                    NaN
3077           Adventure|Comedy|Drama|Family                    NaN
3116        Action|Adventure|Comedy|Thriller                    NaN
3235    Action|Drama|Foreign|Mystery|Thriller                    NaN
3239  Adventure|Action|Comedy|Science Fiction|Music              NaN
3752                       Drama|Thriller                    NaN
3818                               Comedy                    NaN
3853                        Drama|Romance                    NaN
4062                       Comedy|Romance                    NaN
4352                               Comedy                    NaN
4889                          Documentary                    NaN
5636                       Horror|Thriller                    NaN

     release_date  vote_count  vote_average  release_year    budget_adj  \
1758      7/31/09          17           5.1          2009  1.016400e+06
1797       2/5/09          11           7.0          2009  2.541001e+06
1800      9/25/09          11           5.6          2009  7.114803e+06
1871       9/1/09          15           6.5          2009  3.170356e+04
2303     12/17/10          44           6.2          2010  7.000000e+06
2782      1/27/01          42           6.5          2001  1.231488e+07
2805      8/31/01          10           6.3          2001  3.078720e+05
2840      1/20/01          13           5.2          2001  8.620417e+05
3047       9/5/08          13           5.0          2008  6.076720e+06
3077       8/6/08         127           6.0          2008  2.734524e+07
3116      4/28/08          41           5.5          2008  5.063933e+06
3235     12/25/08          53           6.9          2008  9.216358e+06
3239      7/15/08         140           7.7          2008  2.025573e+05
3752       1/1/11          28           4.8          2011  5.816388e+06
3818     12/10/11          47           7.9          2011  2.423495e+05
3853      4/18/11          18           5.9          2011  1.938796e+01
4062      10/5/02          23           7.2          2002  9.698091e+06
4352      4/29/94          21           6.7          1994  1.324000e+07
4889      7/13/12          11           4.7          2012  2.374361e+06
```

```
5636      1/20/13          18           4.8          2013  4.680167e+06

       revenue_adj
1758  1.318242e+06
1797  3.585962e+05
1800  1.452740e+06
1871  1.016400e+04
2303  3.218666e+06
2782  5.475301e+07
2805  5.156156e+06
2840  2.053127e+06
3047  7.038563e+04
3077  4.491954e+07
3116  1.312758e+06
3235  7.697178e+07
3239  3.038360e+00
3752  5.046686e+03
3818  9.693980e+05
3853  1.454097e+01
4062  3.317076e+07
4352  6.400474e+06
4889  3.172146e+07
5636  1.113517e+06
```

Looks like some of the ones remaining are more like independent projects such as Joss Whedan's "Dr Horrible's Sing-Along Blog" or Louis CK stand up. It's weird that some movies don't have it, such as "Sisterhood of the travelling pants 2" and the Recess movie. I guess those are just missing data. I think for the purposes of looking at production companies over time I can remove them. Maybe for the question of revenue vs. runtime or chris pratt movies I can leave them. I'll fork the data cleaning here and make two dfs.

```
In [44]: df_production = df.dropna(subset=['production_companies'])

In [45]: df_rev = df[:]

In [46]: # need to remove NA's before looking for strings
         df_pratt = df.dropna(subset=['cast'])
         df_pratt = df_pratt[df_pratt['cast'].str.contains("Chris Pratt")]

In [47]: df_pratt

Out[47]:           id    imdb_id  popularity     budget     revenue  \
         0     135397  tt0369610   32.985763  150000000  1513528810
         630   118340  tt2015381   14.311205  170000000   773312399
         1454   10521  tt0901476    1.074072   30000000   114663461
         3448   63492  tt0770703    1.120851   20000000    30426096
         3457   60308  tt1210166    1.081676   50000000   110206216
         4414   97630  tt1790885    1.554441   40000000   132820716
         4446   72207  tt1195478    1.095479   30000000    53909751
```

```
5498  146239  tt2387559    1.304540   26000000    51164106

               original_title  \
0               Jurassic World
630    Guardians of the Galaxy
1454                Bride Wars
3448       What's Your Number?
3457                 Moneyball
4414           Zero Dark Thirty
4446  The Five-Year Engagement
5498              Delivery Man

                                                    cast  \
0     Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...
630   Chris Pratt|Zoe Saldana|Dave Bautista|Vin Dies...
1454  Anne Hathaway|Kate Hudson|Bryan Greenberg|Chri...
3448  Chris Evans|Anna Faris|Martin Freeman|Chris Pr...
3457  Brad Pitt|Jonah Hill|Robin Wright|Philip Seymo...
4414  Jessica Chastain|Jason Clarke|Mark Strong|Joel...
4446  Emily Blunt|Alison Brie|Jason Segel|Rhys Ifans...
5498  Vince Vaughn|Cobie Smulders|Chris Pratt|Britt ...

                                        homepage          director  \
0                 http://www.jurassicworld.com/   Colin Trevorrow
630                 http://marvel.com/guardians        James Gunn
1454                   http://www.bridewars.com       Gary Winick
3448       http://www.whatsyournumbermovie.com/        Mark Mylod
3457            http://www.moneyball-movie.com/    Bennett Miller
4414   http://www.zerodarkthirty-movie.com/site/   Kathryn Bigelow
4446  http://www.thefiveyearengagementmovie.com/  Nicholas Stoller
5498                                         NaN         Ken Scott

                                           tagline  \
0                                The park is open.
630                       All heroes start somewhere.
1454                           May the best bride win
3448       Ally's looking for the best ex of her life.
3457                        What are you really worth?
4414                    The Greatest Manhunt in History
4446  A comedy about the journey between popping the...
5498   You're never quite ready for what life delivers.

                                             keywords  \
0     monster|dna|tyrannosaurus rex|velociraptor|island
630   marvel comic|spaceship|space|scene during end ...
1454       bride|friendship|engagement|rivalry|wedding
3448   based on novel|loser|magazine|womanizer|mission
3457  underdog|based on novel|baseball|teamwork|partner
```

```
4414         assassination|cia|hotel|terrorist|prisoner
4446     sex|san francisco|restaurant|frustration|chase
5498                                 remake|sperm donor

                                                 overview  runtime  \
0     Twenty-two years after the events of Jurassic ...      124
630   Light years from Earth, 26 years after being a...      121
1454  Two best friends become rivals when their resp...       89
3448  Ally Darling (Anna Faris) is realizing she's a...      106
3457  The story of Oakland Athletics general manager...      133
4414  A chronicle of the decade-long hunt for al-Qae...      157
4446  Exactly one year after Tom meets Violet, he su...      124
5498  An affable underachiever finds out he's father...      105

                                          genres  \
0     Action|Adventure|Science Fiction|Thriller
630            Action|Science Fiction|Adventure
1454                                     Comedy
3448                             Comedy|Romance
3457                                      Drama
4414                     Thriller|Drama|History
4446                                     Comedy
5498                                     Comedy

                                    production_companies release_date  \
0     Universal Studios|Amblin Entertainment|Legenda...       6/9/15
630   Marvel Studios|Moving Picture Company (MPC)|Bu...      7/30/14
1454  Dune Entertainment|Regency Enterprises|Fox 200...       1/9/09
3448  Regency Enterprises|Contrafilm|New Regency Pic...      9/30/11
3457          Columbia Pictures|Scott Rudin Productions      9/22/11
4414  Columbia Pictures|Annapurna Pictures|First Lig...     12/19/12
4446  Universal Pictures|Dentsu|Relativity Media|Apa...      4/27/12
5498                DreamWorks SKG|Touchstone Pictures     10/10/13

      vote_count  vote_average  release_year    budget_adj   revenue_adj
0           5562           6.5          2015  1.379999e+08  1.392446e+09
630         5612           7.9          2014  1.565855e+08  7.122911e+08
1454         501           5.8          2009  3.049201e+07  1.165440e+08
3448         390           6.2          2011  1.938796e+07  2.949500e+07
3457         899           6.9          2011  4.846990e+07  1.068337e+08
4414        1240           6.5          2012  3.798977e+07  1.261457e+08
4446         319           5.6          2012  2.849233e+07  5.120048e+07
5498         377           6.1          2013  2.433687e+07  4.789131e+07
```

In [48]:
```
df_rev.head()
```

Out[48]:
```
        id    imdb_id  popularity     budget     revenue  \
0  135397  tt0369610   32.985763  150000000  1513528810
```

```
1   76341  tt1392190   28.419936  150000000   378436354
2  262500  tt2908446   13.112507  110000000   295238201
3  140607  tt2488496   11.173104  200000000  2068178225
4  168259  tt2820852    9.335014  190000000  1506249360

                original_title  \
0                Jurassic World
1            Mad Max: Fury Road
2                      Insurgent
3  Star Wars: The Force Awakens
4                      Furious 7

                                                 cast  \
0  Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...
1  Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...
2  Shailene Woodley|Theo James|Kate Winslet|Ansel...
3  Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...
4  Vin Diesel|Paul Walker|Jason Statham|Michelle ...

                                          homepage          director  \
0                    http://www.jurassicworld.com/   Colin Trevorrow
1                      http://www.madmaxmovie.com/     George Miller
2    http://www.thedivergentseries.movie/#insurgent  Robert Schwentke
3  http://www.starwars.com/films/star-wars-episod...       J.J. Abrams
4                          http://www.furious7.com/         James Wan

                        tagline  \
0              The park is open.
1             What a Lovely Day.
2     One Choice Can Destroy You
3  Every generation has a story.
4            Vengeance Hits Home

                                          keywords  \
0  monster|dna|tyrannosaurus rex|velociraptor|island
1   future|chase|post-apocalyptic|dystopia|australia
2  based on novel|revolution|dystopia|sequel|dyst...
3              android|spaceship|jedi|space opera|3d
4                car race|speed|revenge|suspense|car

                                          overview  runtime  \
0  Twenty-two years after the events of Jurassic ...      124
1  An apocalyptic story set in the furthest reach...      120
2  Beatrice Prior must confront her inner demons ...      119
3  Thirty years after defeating the Galactic Empi...      136
4  Deckard Shaw seeks revenge against Dominic Tor...      137

                                  genres  \
```

```
0  Action|Adventure|Science Fiction|Thriller
1  Action|Adventure|Science Fiction|Thriller
2         Adventure|Science Fiction|Thriller
3   Action|Adventure|Science Fiction|Fantasy
4                       Action|Crime|Thriller

                                production_companies release_date  vote_count  \
0  Universal Studios|Amblin Entertainment|Legenda...       6/9/15        5562
1  Village Roadshow Pictures|Kennedy Miller Produ...      5/13/15        6185
2  Summit Entertainment|Mandeville Films|Red Wago...      3/18/15        2480
3          Lucasfilm|Truenorth Productions|Bad Robot     12/15/15        5292
4  Universal Pictures|Original Film|Media Rights ...       4/1/15        2947

   vote_average  release_year    budget_adj   revenue_adj
0           6.5          2015  1.379999e+08  1.392446e+09
1           7.1          2015  1.379999e+08  3.481613e+08
2           6.3          2015  1.012000e+08  2.716190e+08
3           7.5          2015  1.839999e+08  1.902723e+09
4           7.3          2015  1.747999e+08  1.385749e+09
```

In [49]:
```
# Just curious if those 0 runtime's disappeared
df.describe()
```

Out[49]:

	id	popularity	budget	revenue	runtime	\
count	3854.000000	3854.000000	3.854000e+03	3.854000e+03	3854.000000	
mean	39888.185262	1.191554	3.720370e+07	1.076866e+08	109.220291	
std	67222.527399	1.475162	4.220822e+07	1.765393e+08	19.922820	
min	5.000000	0.001117	1.000000e+00	2.000000e+00	15.000000	
25%	6073.500000	0.462368	1.000000e+07	1.360003e+07	95.000000	
50%	11321.500000	0.797511	2.400000e+07	4.480000e+07	106.000000	
75%	38573.250000	1.368324	5.000000e+07	1.242125e+08	119.000000	
max	417859.000000	32.985763	4.250000e+08	2.781506e+09	338.000000	

	vote_count	vote_average	release_year	budget_adj	revenue_adj
count	3854.000000	3854.000000	3854.000000	3.854000e+03	3.854000e+03
mean	527.720291	6.168163	2001.261028	4.423999e+07	1.370647e+08
std	879.956821	0.794920	11.282575	4.480925e+07	2.161114e+08
min	10.000000	2.200000	1960.000000	9.693980e-01	2.370705e+00
25%	71.000000	5.700000	1995.000000	1.309053e+07	1.835735e+07
50%	204.000000	6.200000	2004.000000	3.001611e+07	6.173068e+07
75%	580.000000	6.700000	2010.000000	6.061307e+07	1.632577e+08
max	9767.000000	8.400000	2015.000000	4.250000e+08	2.827124e+09

Seems like it, although there are still some super short movies
Exploratory Data Analysis

Tip: Now that you've trimmed and cleaned your data, you're ready to move on to exploration. Compute statistics and create visualizations with the goal of addressing the research questions that you posed in the Introduction section. It is recommended

that you be systematic with your approach. Look at one variable at a time, and then follow it up by looking at relationships between variables.

1.1.3 Q1: What was Chris Pratt's most successful movie (in terms of revenue)?

In [50]:
```
# Use this, and more code cells, to explore your data. Don't forget to add
#   Markdown cells to document your observations and findings.
df_pratt.sort_values(by='revenue_adj', ascending=False).loc[0,'original_title']
```

Out[50]:
```
'Jurassic World'
```

In [51]:
```
df_pratt['% change'] = df_pratt.sort_values(by='revenue_adj', ascending=True)['revenue_
```

In [52]:
```
df_pratt.sort_values(by='revenue_adj', ascending=False)
```

Out[52]:

	id	imdb_id	popularity	budget	revenue	original_title	cast
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...
630	118340	tt2015381	14.311205	170000000	773312399	Guardians of the Galaxy	Chris Pratt|Zoe Saldana|Dave Bautista|Vin Dies...
4414	97630	tt1790885	1.554441	40000000	132820716	Zero Dark Thirty	Jessica Chastain|Jason Clarke|Mark Strong|Joel...
1454	10521	tt0901476	1.074072	30000000	114663461	Bride Wars	Anne Hathaway|Kate Hudson|Bryan Greenberg|Chri...
3457	60308	tt1210166	1.081676	50000000	110206216	Moneyball	Brad Pitt|Jonah Hill|Robin Wright|Philip Seymo...
4446	72207	tt1195478	1.095479	30000000	53909751	The Five-Year Engagement	Emily Blunt|Alison Brie|Jason Segel|Rhys Ifans...
5498	146239	tt2387559	1.304540	26000000	51164106	Delivery Man	Vince Vaughn|Cobie Smulders|Chris Pratt|Britt ...
3448	63492	tt0770703	1.120851	20000000	30426096	What's Your Number?	Chris Evans|Anna Faris|Martin Freeman|Chris Pr...

	homepage	director
0	http://www.jurassicworld.com/	Colin Trevorrow
630	http://marvel.com/guardians	James Gunn
4414	http://www.zerodarkthirty-movie.com/site/	Kathryn Bigelow

```
1454                    http://www.bridewars.com       Gary Winick
3457             http://www.moneyball-movie.com/    Bennett Miller
4446  http://www.thefiveyearengagementmovie.com/  Nicholas Stoller
5498                                          NaN         Ken Scott
3448        http://www.whatsyournumbermovie.com/        Mark Mylod

                                                tagline  \
0                                    The park is open.
630                         All heroes start somewhere.
4414                    The Greatest Manhunt in History
1454                              May the best bride win
3457                           What are you really worth?
4446  A comedy about the journey between popping the...
5498   You're never quite ready for what life delivers.
3448        Ally's looking for the best ex of her life.

                                               keywords  \
0     monster|dna|tyrannosaurus rex|velociraptor|island
630   marvel comic|spaceship|space|scene during end ...
4414        assassination|cia|hotel|terrorist|prisoner
1454       bride|friendship|engagement|rivalry|wedding
3457  underdog|based on novel|baseball|teamwork|partner
4446     sex|san francisco|restaurant|frustration|chase
5498                                remake|sperm donor
3448    based on novel|loser|magazine|womanizer|mission

                                               overview  runtime  \
0     Twenty-two years after the events of Jurassic ...      124
630   Light years from Earth, 26 years after being a...      121
4414  A chronicle of the decade-long hunt for al-Qae...      157
1454  Two best friends become rivals when their resp...       89
3457  The story of Oakland Athletics general manager...      133
4446  Exactly one year after Tom meets Violet, he su...      124
5498  An affable underachiever finds out he's father...      105
3448  Ally Darling (Anna Faris) is realizing she's a...      106

                                        genres  \
0     Action|Adventure|Science Fiction|Thriller
630           Action|Science Fiction|Adventure
4414                    Thriller|Drama|History
1454                                    Comedy
3457                                     Drama
4446                                    Comedy
5498                                    Comedy
3448                             Comedy|Romance

                                  production_companies release_date  \
0     Universal Studios|Amblin Entertainment|Legenda...       6/9/15
```

```
630   Marvel Studios|Moving Picture Company (MPC)|Bu...      7/30/14
4414  Columbia Pictures|Annapurna Pictures|First Lig...     12/19/12
1454  Dune Entertainment|Regency Enterprises|Fox 200...      1/9/09
3457          Columbia Pictures|Scott Rudin Productions     9/22/11
4446  Universal Pictures|Dentsu|Relativity Media|Apa...      4/27/12
5498                DreamWorks SKG|Touchstone Pictures     10/10/13
3448  Regency Enterprises|Contrafilm|New Regency Pic...      9/30/11

      vote_count  vote_average  release_year    budget_adj   revenue_adj  \
0           5562           6.5          2015  1.379999e+08  1.392446e+09
630         5612           7.9          2014  1.565855e+08  7.122911e+08
4414        1240           6.5          2012  3.798977e+07  1.261457e+08
1454         501           5.8          2009  3.049201e+07  1.165440e+08
3457         899           6.9          2011  4.846990e+07  1.068337e+08
4446         319           5.6          2012  2.849233e+07  5.120048e+07
5498         377           6.1          2013  2.433687e+07  4.789131e+07
3448         390           6.2          2011  1.938796e+07  2.949500e+07

      % change
0     0.954883
630   4.646574
4414  0.082387
1454  0.090892
3457  1.086576
4446  0.069097
5498  0.623710
3448       NaN
```

Jurassic park was clearly the most successful in terms of revenue. It made nearly 95% more than the next highest (Guardians of the Galaxy) after adjusting for inflation.

1.1.4 Q2: Is runtime correlated with budget?

```
In [53]: df_rev.head()
```

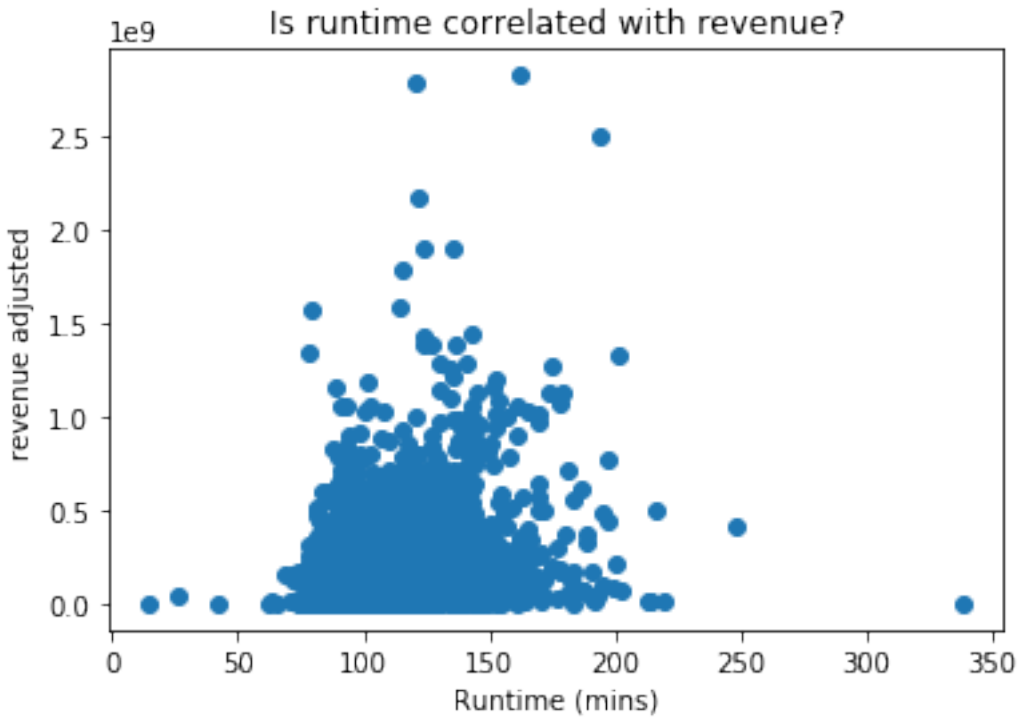
```
Out[53]:        id    imdb_id  popularity     budget     revenue  \
         0  135397  tt0369610   32.985763  150000000  1513528810
         1   76341  tt1392190   28.419936  150000000   378436354
         2  262500  tt2908446   13.112507  110000000   295238201
         3  140607  tt2488496   11.173104  200000000  2068178225
         4  168259  tt2820852    9.335014  190000000  1506249360

                         original_title  \
         0                Jurassic World
         1             Mad Max: Fury Road
         2                      Insurgent
         3  Star Wars: The Force Awakens
         4                      Furious 7
```

```
                                                cast  \
0  Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...
1  Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...
2  Shailene Woodley|Theo James|Kate Winslet|Ansel...
3  Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...
4  Vin Diesel|Paul Walker|Jason Statham|Michelle ...

                                        homepage          director  \
0                    http://www.jurassicworld.com/   Colin Trevorrow
1                      http://www.madmaxmovie.com/     George Miller
2    http://www.thedivergentseries.movie/#insurgent  Robert Schwentke
3  http://www.starwars.com/films/star-wars-episod...       J.J. Abrams
4                          http://www.furious7.com/         James Wan

                       tagline  \
0             The park is open.
1            What a Lovely Day.
2    One Choice Can Destroy You
3  Every generation has a story.
4           Vengeance Hits Home

                                        keywords  \
0  monster|dna|tyrannosaurus rex|velociraptor|island
1   future|chase|post-apocalyptic|dystopia|australia
2  based on novel|revolution|dystopia|sequel|dyst...
3             android|spaceship|jedi|space opera|3d
4               car race|speed|revenge|suspense|car

                                        overview  runtime  \
0  Twenty-two years after the events of Jurassic ...      124
1  An apocalyptic story set in the furthest reach...      120
2  Beatrice Prior must confront her inner demons ...      119
3  Thirty years after defeating the Galactic Empi...      136
4  Deckard Shaw seeks revenge against Dominic Tor...      137

                                    genres  \
0  Action|Adventure|Science Fiction|Thriller
1  Action|Adventure|Science Fiction|Thriller
2         Adventure|Science Fiction|Thriller
3   Action|Adventure|Science Fiction|Fantasy
4                     Action|Crime|Thriller

                                production_companies release_date  vote_count  \
0  Universal Studios|Amblin Entertainment|Legenda...       6/9/15        5562
1  Village Roadshow Pictures|Kennedy Miller Produ...      5/13/15        6185
2  Summit Entertainment|Mandeville Films|Red Wago...      3/18/15        2480
3          Lucasfilm|Truenorth Productions|Bad Robot     12/15/15        5292
```

```
4  Universal Pictures|Original Film|Media Rights ...       4/1/15        2947

   vote_average  release_year    budget_adj   revenue_adj
0           6.5          2015  1.379999e+08  1.392446e+09
1           7.1          2015  1.379999e+08  3.481613e+08
2           6.3          2015  1.012000e+08  2.716190e+08
3           7.5          2015  1.839999e+08  1.902723e+09
4           7.3          2015  1.747999e+08  1.385749e+09
```

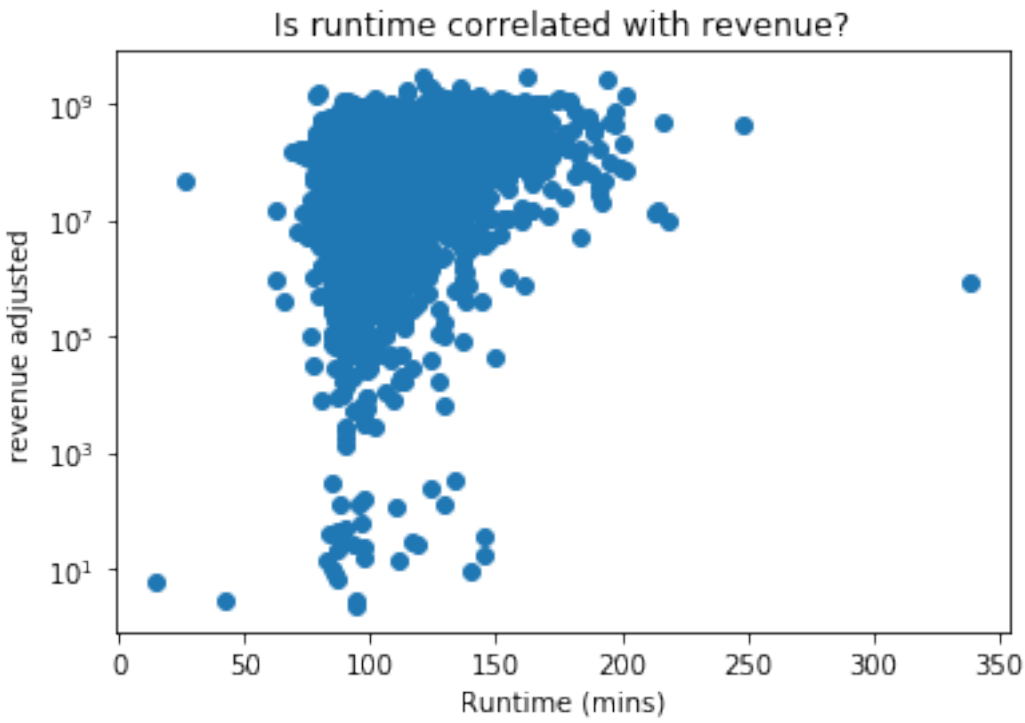
```
In [54]: def plot_scatter_adj(df, comp_var = 'revenue'):
             plt.scatter(df['runtime'], df[str(comp_var) + '_adj'])
             plt.xlabel("Runtime (mins)")
             plt.ylabel(str(comp_var) + " adjusted")
             plt.title("Is runtime correlated with " + str(comp_var) + "?")
```

```
In [55]: plot_scatter_adj(df_rev)
         plt.show()
```



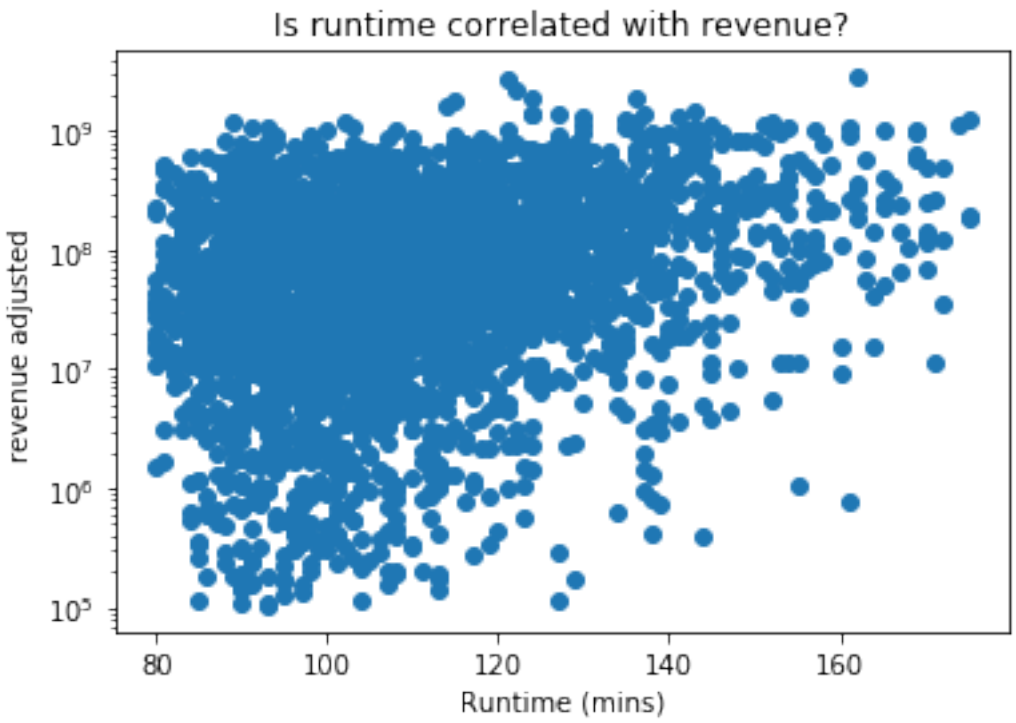
Hmm, not so clear. Could be because there's a big range in revenue in this dataset. Let's try with a log scale on the y axis.

```
In [56]: plot_scatter_adj(df_rev)
         plt.yscale("log")
         plt.show()
```



Again, not clear. Seems that a lot of movies are around the 80-175 minute mark. Let's take one last look by reducing the range of interest. Also, while we're doing it, let's try only looking at movies that made at least $100,000

```
In [57]: df_rev_adj = df_rev.query('runtime >= 80 and runtime <= 175 and revenue_adj >= 100000')

In [58]: plot_scatter_adj(df_rev_adj)
         plt.yscale("log")
         plt.show()
```

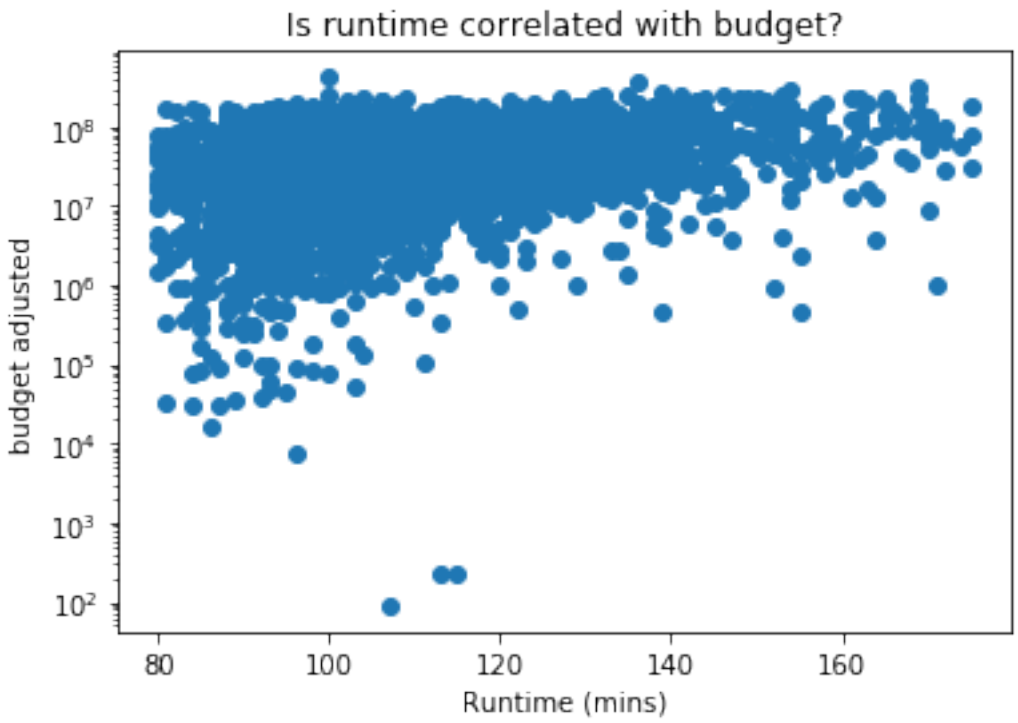


Nah, I don't see it. There may be a very weak positive correlation but it's not clear. I'll just run the statistical test out of curiosity but it seems that runtime and revenue aren't strongly correlated. Although I can imagine that longer movies cost more so that might naturally lead to more revenue since bigger budgets come from larger studios which tend to make big blockbusters. Let's plot runtime vs. budget out of curiosity too.

```
In [59]: from scipy.stats.stats import pearsonr
         # Returns a tuple with (Pearson's correlation coefficient, 2-tailed p-value)
         # Correlation coefficient between -1 and +1 where 0 is no correlation, +1 is perfect po
         # and -1 is perfect negative correlation
         pearsonr(df_rev_adj['runtime'], df_rev_adj['revenue_adj'])

Out[59]: (0.28623398889001084, 2.7900527626608778e-70)
```

Yup, seems weekly positively correlated.

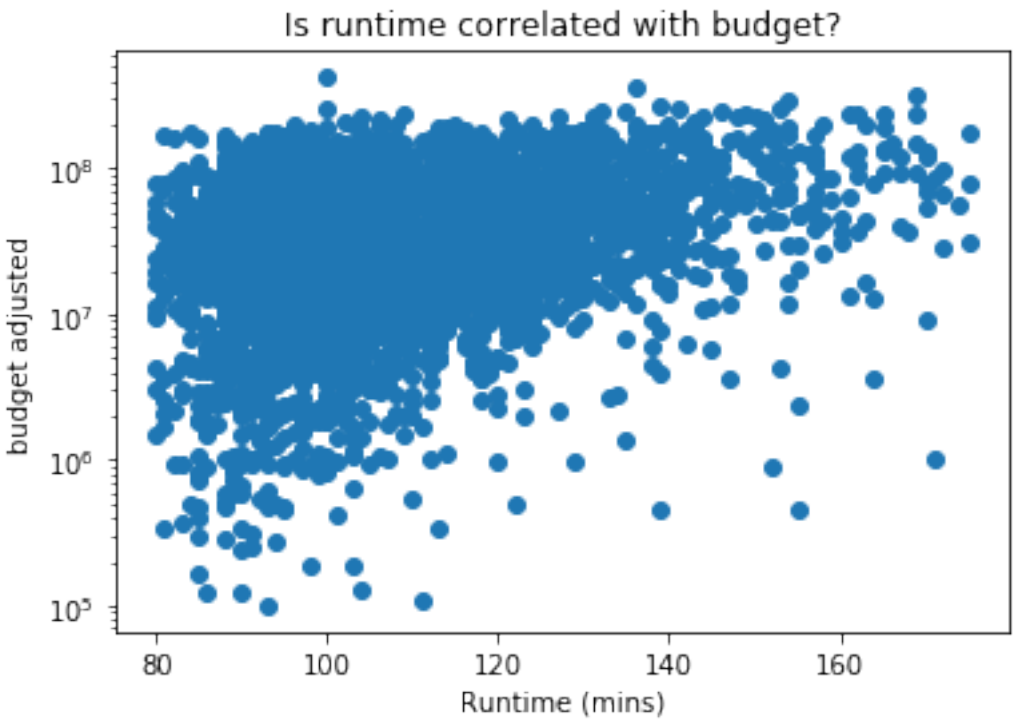
```
In [60]: plot_scatter_adj(df_rev_adj, 'budget')
         plt.yscale("log")
         plt.show()
```



Let's remove those outliers

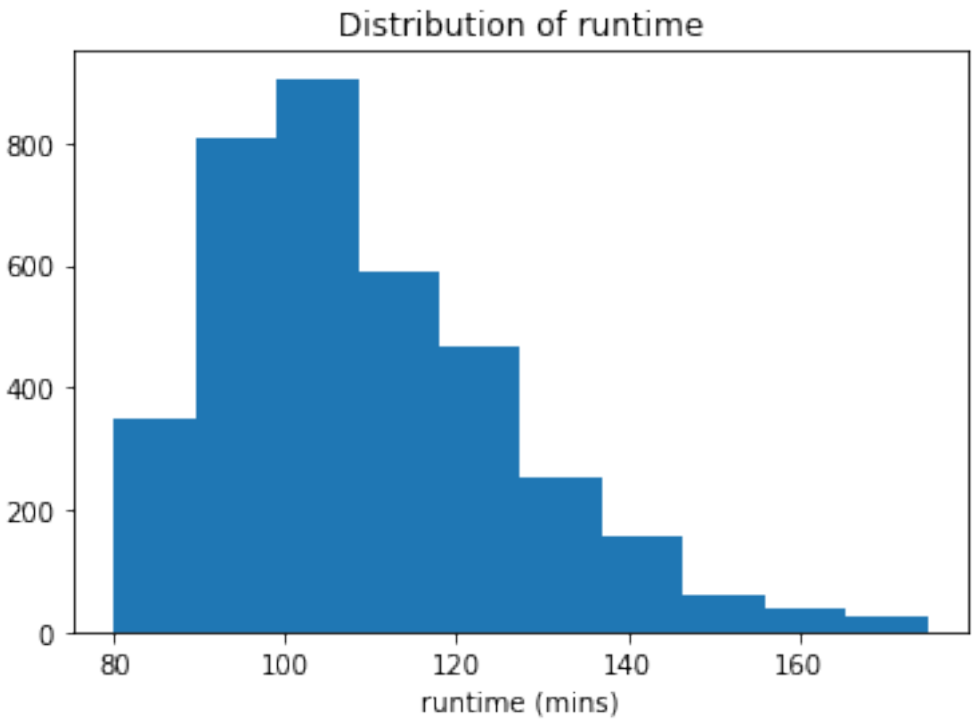
```
In [61]: df_budg_adj = df_rev_adj.query('budget_adj >= 100000')
```

```
In [62]: plot_scatter_adj(df_budg_adj, 'budget')
         plt.yscale("log")
         plt.show()
```



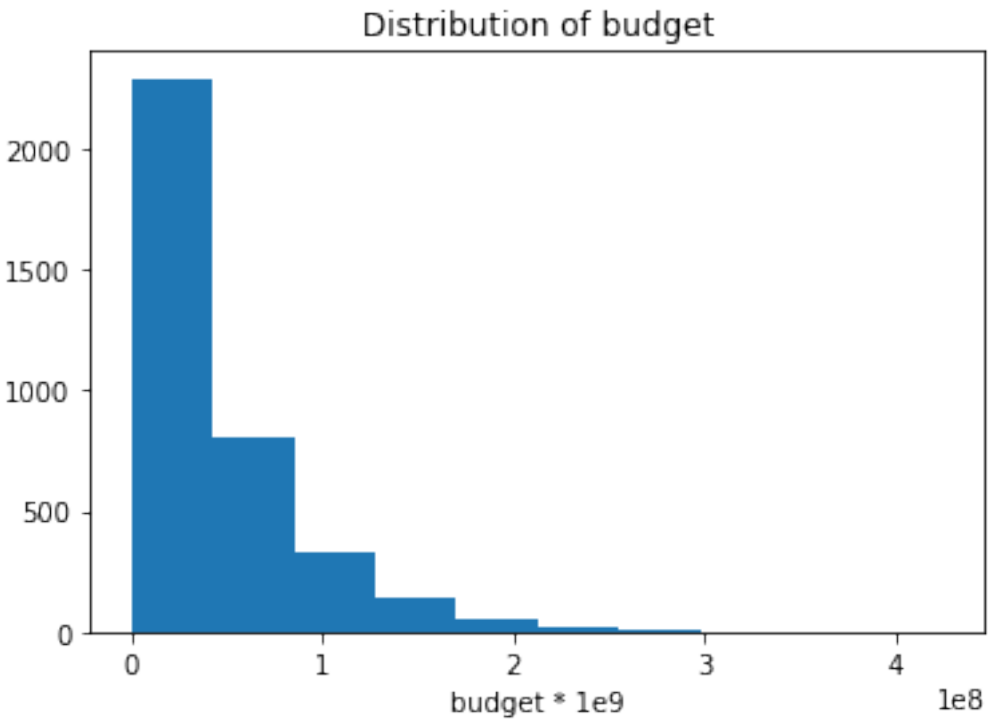
Still no obvious correlation. Nothing to see here. I think the problem probably comes from the fact that the runtimes, revenues and budgets are not normally distributed. Let's check the histograms

```
In [63]: plt.hist(df_budg_adj['runtime'])
         plt.xlabel("runtime (mins)")
         plt.title("Distribution of runtime");
```

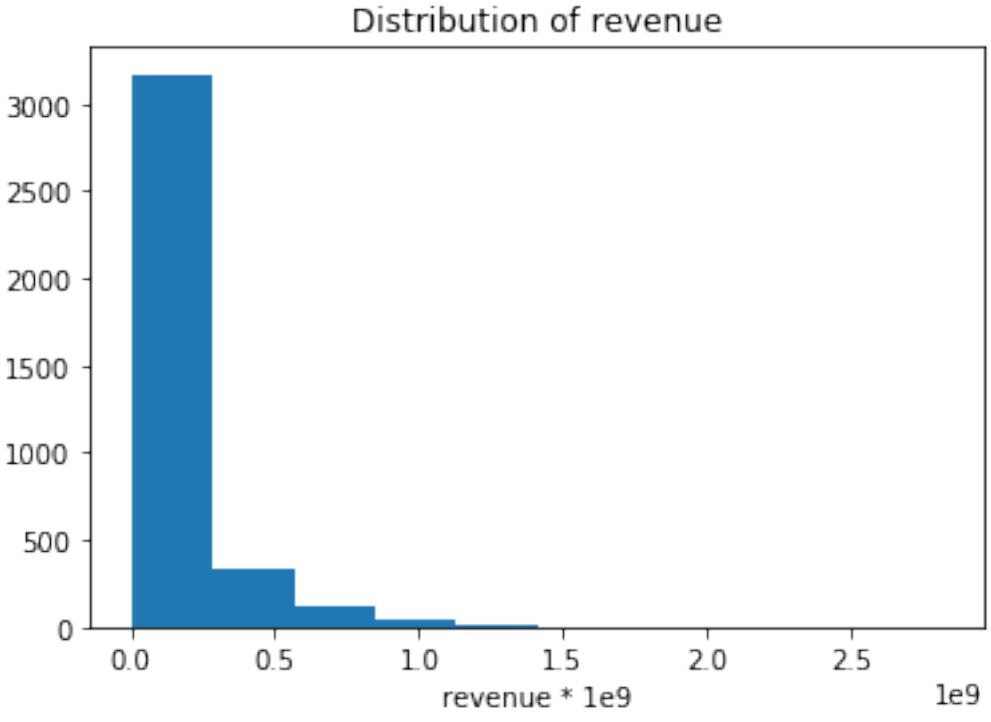


Yeah this is a bit left skewed. Most movies are around 100-120 minutes with a few really long movies over 150mins

```
In [64]: plt.hist(df_budg_adj['budget_adj'])
         plt.xlabel("budget * 1e9")
         plt.title("Distribution of budget");
```




```
In [65]: plt.hist(df_rev_adj['revenue_adj'])
         plt.xlabel("revenue * 1e9")
         plt.title("Distribution of revenue");
```



Wow. Revenue is really right skewed, as is budget. Must be why it's so hard to see

Conclusions

So I initially wanted to know which Chris Pratt movie brought in the most revenue and it turns out to be Jurassic World. It made around 95% more than the next highest which was Guardians of the Galaxy.

I also wanted to see if runtime was correlated with revenue. Turns out that there is only a very weak positive correlation between revenue and runtime. I dug a bit deeper into the relationship between runtime and budget, with the hypothesis that longer movies require more money to make, but again found a weak correlation. I think the main issue is that the budget/revenue distributions are very right-skewed so it's difficult to find discrepancies between them. Similarly, most movies are between 80-120 mins so it's not likely that bigger budget movies would just keep getting longer and longer. There's a cap and a minimum for big budget movies (they're not going to make a 60min nor a 300min movie).

If I had to do it again, I would probably look at revenue vs. reviews but I think that's what most people would have looked at so I wanted to try something different

Limitations

I don't split my analysis of runtime vs. revenue/budget into groups which is potentially why there is no obvious correlation. I group together both small studios and large, with varying budgets/revenues. I only look at movies with a revenue/budget over $100K, and maybe I would have seen a correlation in the low-budget/revenue movies.

I remove all movies with a revenue/budget of $0 and with a production/cast which is missing, without looking more closely at why they have these extreme values.

```
In [66]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

```
Out[66]: 0
```

```
In [ ]:
```