Investigate_a_Dataset

January 8, 2018

1 Project: Investigate a movie database to determine characteristics of movies sucessfull at the box office from 2010 to 2015 (inclusive)

1.1 Table of Contents

1.2 Introduction

This project analysyses the movie data present in the TMDB database which provides revenue, genre, cast, title, and other movie information for movies from 1960 to 2015. We will use this database to clean-up and extract information we need for the analysis of movies released from 2010 to 2015.

In [1]:
```
# Use this cell to set up import statements for all of the packages that you
#   plan to use.

# Remember to include a 'magic word' so that your visualizations are plotted
#   inline with the notebook. See this page for more:
#   http://ipython.readthedocs.io/en/stable/interactive/magics.html

# Importing Packages

import numpy as np
import pandas as pd
```

Data Wrangling
List the first few rows to understand the fields and data format. Examine the data for null/empty values and query data stats to understand max, min etc. to make the appropriate cleaning decisions before analyzing the data.

1.2.1 General Properties

```
In [2]: # Load your data and print out a few lines. Perform operations to inspect data
        #   types and look for instances of missing or possibly errant data.

        #Reading CSV and print the top 5 rows

        df = pd.read_csv('tmdb_movies.csv')
        df.head()
```

```
Out[2]:        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      ...

                                                     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

[5 rows x 21 columns]
```

In [3]:
```
#Check if thare are any null value fields

pd.isnull(df).sum()
```

Out[3]:
```
id                         0
imdb_id                   10
popularity                 0
budget                     0
revenue                    0
original_title             0
cast                      76
homepage                7930
director                  44
tagline                 2824
keywords                1493
overview                   4
runtime                    0
genres                    23
production_companies    1030
release_date               0
vote_count                 0
vote_average               0
```

```
release_year             0
budget_adj               0
revenue_adj              0
dtype: int64
```

In [4]:
```
#Check if there are any Zeroes, we can see from Min value that budget, revenue, runtime,

df.describe()
```

Out[4]:

	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

In [5]:
```
#Check the rows with zero revenue
df[df["revenue"] == 0]
```

Out[5]:

	id	imdb_id	popularity	budget	revenue
48	265208	tt2231253	2.932340	30000000	0
67	334074	tt3247714	2.331636	20000000	0
74	347096	tt3478232	2.165433	0	0
75	308369	tt2582496	2.141506	0	0
92	370687	tt3608646	1.876037	0	0
93	307663	tt3480796	1.872696	10000000	0
100	326359	tt4007502	1.724712	0	0
101	254302	tt0462335	1.661789	0	0
103	292040	tt3321300	1.646664	0	0
116	297291	tt3086386	1.380320	0	0
122	277355	tt1945084	1.342839	0	0
133	157827	tt2217859	1.251681	11000000	0
140	300803	tt3829170	1.144808	0	0
143	378373	tt3532278	1.128081	0	0
145	294963	tt2494362	1.073349	1800000	0
147	245698	tt1596345	1.063055	0	0

149	346808	tt3181776	1.041922	20000000	0
151	290637	tt3733778	1.036825	0	0
152	244458	tt1567437	1.027620	0	0
154	314405	tt3278330	1.008474	12000000	0
156	157843	tt1837636	0.973316	15000000	0
158	290762	tt2245003	0.953647	0	0
159	251516	tt3472226	0.953046	630019	0
164	228968	tt2917388	0.917040	0	0
165	347969	tt2479478	0.913085	60000000	0
166	237756	tt2393845	0.906860	0	0
169	311291	tt3544082	0.894477	0	0
174	342474	tt3289712	0.861179	0	0
175	277217	tt3440298	0.848748	0	0
176	207936	tt2338424	0.843174	0	0
...	...	...	...	...	...
10834	12639	tt0060897	0.310688	0	0
10836	38720	tt0061170	0.239435	0	0
10837	19728	tt0060177	0.291704	0	0
10838	22383	tt0060862	0.151845	0	0
10839	13353	tt0060550	0.276133	0	0
10840	34388	tt0060437	0.102530	0	0
10841	42701	tt0062262	0.264925	75000	0
10842	36540	tt0061199	0.253437	0	0
10843	29710	tt0060588	0.252399	0	0
10844	23728	tt0059557	0.236098	0	0
10845	5065	tt0059014	0.230873	0	0
10846	17102	tt0059127	0.212716	0	0
10847	28763	tt0060548	0.034555	0	0
10849	28270	tt0060445	0.206537	0	0
10850	26268	tt0060490	0.202473	0	0
10851	15347	tt0060182	0.342791	0	0
10852	37301	tt0060165	0.227220	0	0
10853	15598	tt0060086	0.163592	0	0
10854	31602	tt0060232	0.146402	0	0
10855	13343	tt0059221	0.141026	700000	0
10856	20277	tt0061135	0.140934	0	0
10857	5921	tt0060748	0.131378	0	0
10858	31918	tt0060921	0.317824	0	0
10859	20620	tt0060955	0.089072	0	0
10860	5060	tt0060214	0.087034	0	0
10861	21	tt0060371	0.080598	0	0
10862	20379	tt0060472	0.065543	0	0
10863	39768	tt0060161	0.065141	0	0
10864	21449	tt0061177	0.064317	0	0
10865	22293	tt0060666	0.035919	19000	0

	original_title \
48	Wild Card

67	Survivor
74	Mythica: The Darkspore
75	Me and Earl and the Dying Girl
92	Mythica: The Necromancer
93	Vice
100	Frozen Fever
101	High-Rise
103	Spooks: The Greater Good
116	The Scorpion King: The Lost Throne
122	Everly
133	Louder Than Bombs
140	Dragonheart 3: The Sorcerer's Curse
143	Brothers of the Wind
145	Bone Tomahawk
147	Pawn Sacrifice
149	Momentum
151	Pay the Ghost
152	The Voices
154	Il racconto dei racconti
156	Queen of the Desert
158	Miss You Already
159	Kung Fury
164	Kidnapping Mr. Heineken
165	The Ridiculous 6
166	Kill Me Three Times
169	45 Years
174	Jenny's Wedding
175	Descendants
176	Tumbledown
...	...
10834	Return of the Seven
10836	Walk Don't Run
10837	The Blue Max
10838	The Professionals
10839	It's the Great Pumpkin, Charlie Brown
10840	Funeral in Berlin
10841	The Shooting
10842	Winnie the Pooh and the Honey Tree
10843	Khartoum
10844	Our Man Flint
10845	Carry On Cowboy
10846	Dracula: Prince of Darkness
10847	Island of Terror
10849	Gambit
10850	Harper
10851	Born Free
10852	A Big Hand for the Little Lady
10853	Alfie

```
10854                                               The Chase
10855                                 The Ghost & Mr. Chicken
10856                                      The Ugly Dachshund
10857                                            Nevada Smith
10858  The Russians Are Coming, The Russians Are Coming
10859                                                 Seconds
10860                                    Carry On Screaming!
10861                                      The Endless Summer
10862                                              Grand Prix
10863                                    Beregis Avtomobilya
10864                                  What's Up, Tiger Lily?
10865                                 Manos: The Hands of Fate

                                                    cast  \
48     Jason Statham|Michael Angarano|Milo Ventimigli...
67     Pierce Brosnan|Milla Jovovich|Dylan McDermott|...
74     Melanie Stone|Kevin Sorbo|Adam Johnson|Jake St...
75     Thomas Mann|RJ Cyler|Olivia Cooke|Connie Britt...
92     Melanie Stone|Adam Johnson|Kevin Sorbo|Nicola ...
93     Ambyr Childers|Thomas Jane|Bryan Greenberg|Bru...
100    Kristen Bell|Idina Menzel|Jonathan Groff|Josh ...
101    Tom Hiddleston|Sienna Miller|Jeremy Irons|Luke...
103    Peter Firth|Kit Harington|Jennifer Ehle|Lara P...
116    Victor Webster|Ellen Hollman|Barry Bostwick|Wi...
122    Salma Hayek|Hiroyuki Watanabe|Jennifer Blanc|T...
133    Gabriel Byrne|Isabelle Huppert|Jesse Eisenberg...
140    Julian Morris|Tamzin Merchant|Jassa Ahluwalia|...
143     Manuel Camacho|Jean Reno|Tobias Moretti|Eva Kuen
145    Kurt Russell|Richard Jenkins|Matthew Fox|Lili ...
147    Tobey Maguire|Lily Rabe|Peter Sarsgaard|Liev S...
149    Olga Kurylenko|Morgan Freeman|James Purefoy|Je...
151    Nicolas Cage|Sarah Wayne Callies|Veronica Ferr...
152    Ryan Reynolds|Gemma Arterton|Anna Kendrick|Jac...
154    Salma Hayek|Vincent Cassel|John C. Reilly|Toby...
156    Nicole Kidman|James Franco|Robert Pattinson|Da...
158    Drew Barrymore|Toni Collette|Dominic Cooper|Ja...
159    David Sandberg|Jorma Taccone|Leopold Nilsson|A...
164    Anthony Hopkins|Jim Sturgess|Sam Worthington|R...
165    Adam Sandler|Taylor Lautner|Steve Buscemi|Terr...
166    Teresa Palmer|Simon Pegg|Luke Hemsworth|Sulliv...
169    Charlotte Rampling|Tom Courtenay|Dolly Wells|G...
174    Katherine Heigl|Tom Wilkinson|Alexis Bledel|Gr...
175    Booboo Stewart|Dove Cameron|Keegan Connor Trac...
176    Joe Manganiello|Jason Sudeikis|Blythe Danner|R...
...                                                  ...
10834  Yul Brynner|Robert Fuller|JuliÃąn Mateos|Warre...
10836  Cary Grant|Samantha Eggar|Jim Hutton|John Stan...
10837  George Peppard|James Mason|Ursula Andress|Jere...
```

```
10838  Burt Lancaster|Lee Marvin|Robert Ryan|Woody St...
10839  Christopher Shea|Sally Dryer|Kathy Steinberg|A...
10840  Michael Caine|Paul Hubschmid|Oskar Homolka|Eva...
10841  Will Hutchins|Millie Perkins|Jack Nicholson|Wa...
10842  Sterling Holloway|Junius Matthews|Sebastian Ca...
10843  Charlton Heston|Laurence Olivier|Richard Johns...
10844  James Coburn|Lee J. Cobb|Gila Golan|Edward Mul...
10845  Sid James|Jim Dale|Angela Douglas|Kenneth Will...
10846  Christopher Lee|Barbara Shelley|Andrew Keir|Fr...
10847  Peter Cushing|Edward Judd|Carole Gray|Eddie By...
10849  Michael Caine|Shirley MacLaine|Herbert Lom|Joh...
10850  Paul Newman|Lauren Bacall|Julie Harris|Arthur ...
10851  Virginia McKenna|Bill Travers|Geoffrey Keen|Pe...
10852  Henry Fonda|Joanne Woodward|Jason Robards|Paul...
10853  Michael Caine|Shelley Winters|Millicent Martin...
10854  Marlon Brando|Jane Fonda|Robert Redford|E.G. M...
10855  Don Knotts|Joan Staley|Liam Redmond|Dick Sarge...
10856  Dean Jones|Suzanne Pleshette|Charles Ruggles|K...
10857  Steve McQueen|Karl Malden|Brian Keith|Arthur K...
10858  Carl Reiner|Eva Marie Saint|Alan Arkin|Brian K...
10859  Rock Hudson|Salome Jens|John Randolph|Will Gee...
10860  Kenneth Williams|Jim Dale|Harry H. Corbett|Joa...
10861  Michael Hynson|Robert August|Lord 'Tally Ho' B...
10862  James Garner|Eva Marie Saint|Yves Montand|Tosh...
10863  Innokentiy Smoktunovskiy|Oleg Efremov|Georgi Z...
10864  Tatsuya Mihashi|Akiko Wakabayashi|Mie Hama|Joh...
10865  Harold P. Warren|Tom Neyman|John Reynolds|Dian...

                                                 homepage  \
48                                                    NaN
67                                http://survivormovie.com/
74                http://www.mythicamovie.com/#!blank/wufvh
75     http://www.foxsearchlight.com/meandearlandthed...
92                http://www.mythicamovie.com/#!blank/y9ake
93                                                    NaN
100                                                   NaN
101                                                   NaN
103    http://www.shinepictures.co.uk/films/9/spooks-...
116                                                   NaN
122                                                   NaN
133                  http://www.motlys.com/louder-than-bombs
140                                                   NaN
143    http://www.terramater.at/cinema/brothers-of-th...
145                                                   NaN
147                                                   NaN
149                                                   NaN
151                                                   NaN
152                                                   NaN
```

```
154                                                 NaN
156                                                 NaN
158                                                 NaN
159                             http://www.kungfury.com/
164                 http://kidnappingmrheinekenmovie.com/
165                 http://www.netflix.com/title/80039517
166                                                 NaN
169                                                 NaN
174         https://www.facebook.com/jennysweddingmovie
175                                                 NaN
176                                                 NaN
...                                                 ...
10834                                               NaN
10836                                               NaN
10837                                               NaN
10838                                               NaN
10839                                               NaN
10840                                               NaN
10841                                               NaN
10842                                               NaN
10843                                               NaN
10844                                               NaN
10845                                               NaN
10846                                               NaN
10847                                               NaN
10849                                               NaN
10850                                               NaN
10851                                               NaN
10852                                               NaN
10853                                               NaN
10854                                               NaN
10855                                               NaN
10856                                               NaN
10857                                               NaN
10858                                               NaN
10859                                               NaN
10860                                               NaN
10861                                               NaN
10862                                               NaN
10863                                               NaN
10864                                               NaN
10865                                               NaN

                         director  \
48                     Simon West
67                 James McTeigue
74                  Anne K. Black
75            Alfonso Gomez-Rejon
```

92	A. Todd Smith
93	Brian A Miller
100	Chris Buck|Jennifer Lee
101	Ben Wheatley
103	Bharat Nalluri
116	Mike Elliott
122	Joe Lynch
133	Joachim Trier
140	Colin Teague
143	Gerado Olivares|Otmar Penker
145	S. Craig Zahler
147	Edward Zwick
149	Stephen S. Campanelli
151	Uli Edel
152	Marjane Satrapi
154	Matteo Garrone
156	Werner Herzog
158	Catherine Hardwicke
159	David Sandberg
164	Daniel Alfredson
165	Frank Coraci
166	Kriv Stenders
169	Andrew Haigh
174	Mary Agnes Donoghue
175	Kenny Ortega
176	Sean Mewshaw
...	...
10834	Burt Kennedy
10836	Charles Walters
10837	John Guillermin
10838	Richard Brooks
10839	Bill Melendez
10840	Guy Hamilton
10841	Monte Hellman
10842	Wolfgang Reitherman
10843	Basil Dearden|Eliot Elisofon
10844	Daniel Mann
10845	Gerald Thomas
10846	Terence Fisher
10847	Terence Fisher
10849	Ronald Neame
10850	Jack Smight
10851	James Hill
10852	Fielder Cook
10853	Lewis Gilbert
10854	Arthur Penn
10855	Alan Rafkin
10856	Norman Tokar

```
10857             Henry Hathaway
10858             Norman Jewison
10859         John Frankenheimer
10860              Gerald Thomas
10861                Bruce Brown
10862         John Frankenheimer
10863             Eldar Ryazanov
10864                Woody Allen
10865           Harold P. Warren

                                                 tagline     ...        \
48          Never bet against a man with a killer hand.     ...
67                   His Next Target is Now Hunting Him     ...
74                                                  NaN     ...
75             A Little Friendship Never Killed Anyone.     ...
92                                                  NaN     ...
93                       Where the future is your past.     ...
100                                                 NaN     ...
101                         Leave the real world behind     ...
103                                                 NaN     ...
116                                    Action Adventure     ...
122                                  Enter if you dare.     ...
133                                                 NaN     ...
140                                                 NaN     ...
143                Sometimes a friendship sets you free     ...
145   May the Lord have mercy and grant you a swift ...     ...
147   On the board he fought the Cold War. In his mi...     ...
149                                                 NaN     ...
151                                 Evil walks among us.     ...
152                      Hearing voices can be murder.     ...
154                          Desire. Envy. Obsession.     ...
156                                                 NaN     ...
158     When life falls apart, friends keep it together     ...
159   It takes a cop from the future to fight an ene...     ...
164   It was the perfect crime until they got away w...     ...
165                                                 NaN     ...
166                               Once is never enough.     ...
169                                                 NaN     ...
174                      Family is worth fighting for.     ...
175        They're not bad. They're just born that way.     ...
176                Turn the page. Start a new chapter.     ...
...                                                 ...     ...
10834 Between the law and the lawless - SEVEN again...    ...
10836           Run, don't walk to see Walk, Don't Run.     ...
10837          There was no quiet on the Western Front!     ...
10838                            Rough, tough and ready.     ...
10839      Every year he rises from the pumpkin patch...     ...
10840                                               NaN     ...
```

```
10841  Suspenseful desert pursuit in the "High Noon" ...     ...
10842                                                NaN     ...
10843  Where the Nile divides, the great Cinerama adv...     ...
10844                       The ORIGINAL man of mystery!     ...
10845                             How the west was lost!     ...
10846  DEAD for Ten Years DRACULA, Prince of Darkness...     ...
10847  How could they stop the devouring death...that...     ...
10849            Shirley MacLaine raises Michael Caine!     ...
10850    Harper takes a case - and the payoff is murder.     ...
10851  From The Pages Of The Beloved Best Seller... A...     ...
10852  All the action you can take...all the adventur...     ...
10853                Is any man an Alfie? Ask any girl!     ...
10854                                  The chase is on!     ...
10855  G-G-GUARANTEED! YOU'LL BE SCARED UNTIL YOU LAU...     ...
10856  A HAPPY HONEYMOON GOES TO THE DOGS!...When a G...     ...
10857  Some called him savage- and some called him sa...     ...
10858   IT'S A PLOT! ...to make the world die laughing!!     ...
10859                                                NaN     ...
10860  Carry On Screaming with the Hilarious CARRY ON...     ...
10861                                                NaN     ...
10862  Cinerama sweeps YOU into a drama of speed and ...     ...
10863                                                NaN     ...
10864                          WOODY ALLEN STRIKES BACK!     ...
10865       It's Shocking! It's Beyond Your Imagination!     ...

                                                overview runtime  \
48     When a Las Vegas bodyguard with lethal skills ...       92
67     A Foreign Service Officer in London tries to p...       96
74     When Teelaâs sister is murdered and a powerf...    108
75     Greg is coasting through senior year of high s...      105
92     Mallister takes Thane prisoner and forces Mare...        0
93     Julian Michaels has designed the ultimate reso...       96
100    On Anna's birthday, Elsa and Kristoff are dete...        8
101    Dr. Robert Laing is the newest resident of a l...      119
103    During a handover to the head of counter-terro...      104
116    When he is betrayed by a trusted friend, Matha...      105
122    After she betrays a powerful mob boss, a woman...       90
133    Three years after his wife, acclaimed photogra...      109
140    When aspiring knight Gareth goes in search of ...       97
143    The way of the eagle is to raise two chicks. T...       98
145    During a shootout in a saloon, Sheriff Hunt in...      132
147    American chess champion Bobby Fischer prepares...      114
149    When Alex, an infiltration expert with a secre...       96
151    One year after his young son disappeared durin...       94
152    A mentally unhinged factory worker must decide...      101
154    A fantasy film with horror elements, "The Tale...      125
156    A chronicle of Gertrude Bell's life, a travele...      128
158    The friendship between two life-long girlfrien...      112
```

```
159    During an unfortunate series of events, a frie...     31
164    The true story of the kidnapping of Freddy Hei...     95
165    When his long-lost outlaw father returns, Tomm...    119
166    While on a seemingly routine job, a jaded hit ...     90
169    There is just one week until Kate Mercer's 45t...     95
174    Jenny Farrell is getting married. But how will...     94
175    A present-day idyllic kingdom where the benevo...    112
176    A young woman struggles to move on with her li...    105
...                                                  ...    ...
10834  Chico one of the remaining members of The Magn...     95
10836  British industrialist Sir William Rutland - "B...    114
10837  A young pilot in the German air force of 1918,...    156
10838  The Professionals is a 1966 American Western f...    117
10839  This classic "Peanuts" tale focuses on the thu...     25
10840  Colonel Stok, a Soviet intelligence officer re...    102
10841  A hired gun seeks to enact revenge on a group ...     82
10842  Christopher Robin's bear attempts to raid a be...     25
10843  English General Charles George Gordon, a devou...    134
10844  When scientists use eco-terrorism to impose th...    108
10845  Stodge City is in the grip of the Rumpo Kid an...     93
10846  Whilst vacationing in the Carpathian Mountain,...     90
10847  A small island community is overrun with creep...     89
10849  Harry Dean (Michael Caine) has a perfect plan ...    109
10850  Harper is a cynical private eye in the best tr...    121
10851  Born Free (1966) is an Open Road Films Ltd./Co...     95
10852  A naive traveler in Laredo gets involved in a ...     95
10853  The film tells the story of a young man who le...    114
10854  Most everyone in town thinks that Sheriff Cald...    135
10855  Luther Heggs aspires to being a reporter for h...     90
10856  The Garrisons (Dean Jones and Suzanne Pleshett...     93
10857  Nevada Smith is the young son of an Indian mot...    128
10858  Without hostile intent, a Soviet sub runs agro...    126
10859  A secret organisation offers wealthy people a ...    100
10860  The sinister Dr Watt has an evil scheme going...     87
10861  The Endless Summer, by Bruce Brown, is one of ...     95
10862  Grand Prix driver Pete Aron is fired by his te...    176
10863  An insurance agent who moonlights as a carthie...     94
10864  In comic Woody Allen's film debut, he took the...     80
10865  A family gets lost on the road and stumbles up...     74

                                        genres  \
48                       Thriller|Crime|Drama
67                      Crime|Thriller|Action
74                  Action|Adventure|Fantasy
75                               Comedy|Drama
92                  Fantasy|Action|Adventure
93     Thriller|Science Fiction|Action|Adventure
100               Adventure|Animation|Family
```

```
101                   Action|Drama|Science Fiction
103                                Thriller|Action
116                     Action|Fantasy|Adventure
122                                Thriller|Action
133                                          Drama
140                     Action|Adventure|Fantasy
143                       Adventure|Drama|Family
145             Horror|Western|Adventure|Drama
147                                          Drama
149                                Thriller|Action
151                                Horror|Thriller
152              Horror|Thriller|Comedy|Crime
154                       Romance|Fantasy|Horror
156                                  Drama|History
158                          Comedy|Drama|Romance
159        Action|Comedy|Science Fiction|Fantasy
164                Drama|Action|Crime|Thriller
165                                 Comedy|Western
166                                Comedy|Thriller
169                                          Drama
174                                   Comedy|Drama
175         Music|Action|Adventure|Comedy|Family
176                          Music|Romance|Comedy
...                                            ...
10834                               Action|Western
10836                               Comedy|Romance
10837                War|Action|Adventure|Drama
10838                   Action|Adventure|Western
10839                             Family|Animation
10840                                     Thriller
10841                                      Western
10842                             Animation|Family
10843            Adventure|Drama|War|History|Action
10844    Adventure|Comedy|Fantasy|Science Fiction
10845                               Comedy|Western
10846                                       Horror
10847                       Science Fiction|Horror
10849                          Action|Comedy|Crime
10850           Action|Drama|Thriller|Crime|Mystery
10851        Adventure|Drama|Action|Family|Foreign
10852                                      Western
10853                         Comedy|Drama|Romance
10854                         Thriller|Drama|Crime
10855              Comedy|Family|Mystery|Romance
10856                          Comedy|Drama|Family
10857                               Action|Western
10858                                   Comedy|War
10859       Mystery|Science Fiction|Thriller|Drama
```

```
10860                                          Comedy
10861                                     Documentary
10862                          Action|Adventure|Drama
10863                                  Mystery|Comedy
10864                                   Action|Comedy
10865                                          Horror

                                    production_companies release_date  \
48     Current Entertainment|Lionsgate|Sierra / Affin...      1/14/15
67     Nu Image Films|Winkler Films|Millennium Films|...      5/21/15
74                              Arrowstorm Entertainment      6/24/15
75                                     Indian Paintbrush      6/12/15
92     Arrowstorm Entertainment|Camera 40 Productions...     12/19/15
93     Grindstone Entertainment Group|K5 Internationa...      1/16/15
100    Walt Disney Pictures|Walt Disney Animation Stu...       3/9/15
101    Ingenious Media|HanWay Films|Scope Pictures|Re...      9/26/15
103    BBC Films|Isle of Man Film|Shine Pictures|Kudo...      4/11/15
116                                   Universal Pictures       1/9/15
122    Crime Scene Pictures|Radius-TWC|Anonymous Cont...      1/23/15
133            Motlys|Arte France CinÃĨma|Animal Kingdom      5/18/15
140                                Raffaella Productions      2/24/15
143                         Terra Mater Factual Studios      12/24/15
145    Caliber Media Company|The Fyzz Facility|Realbu...     10/23/15
147    Material Pictures|MICA Entertainment|PalmStar ...      9/16/15
149                                Thaba Media|Azari Media       8/1/15
151          Voltage Films|Midnight Kitchen Productions      9/16/15
152    Studio Babelsberg|Mandalay Vision|1984 Private...       2/6/15
154    HanWay Films|Rai Cinema|Le Pacte|Fonds Eurimag...      5/14/15
156    Benaroya Pictures|H Films|Raslan Company of Am...       9/3/15
158                                S Films|New Sparta Films      9/12/15
159                                       Laser Unicorns      5/28/15
164    Umedia|Informant Europe SPRL|European Film Com...      3/12/15
165                            Happy Madison Productions     12/11/15
166    Parabolic Pictures|Stable Way Entertainment|Ca...      4/10/15
169                                           The Bureau      8/28/15
174    MM Productions|Merced Media Partners|PalmStar ...      7/31/15
175                              Walt Disney Television      7/31/15
176                   Echo Films|Bron Studios|Hahnscape      4/18/15
...                                                  ...          ...
10834     C.B. Films S.A.|The Mirisch Production Company     10/19/66
10836                      Columbia Pictures Corporation       1/1/66
10837              Twentieth Century Fox Film Corporation      6/21/66
10838                                  Columbia Pictures      11/1/66
10839                            Warner Bros. Home Video     10/27/66
10840                        Lowndes Productions Limited     12/22/66
10841                                      Proteus Films     10/23/66
10842                                                NaN       1/1/66
10843                  Julian Blaustein Productions Ltd.       6/9/66
```

10844	20th Century Fox	1/16/66
10845	Peter Rogers Productions	3/1/66
10846	Seven Arts Productions|Hammer Film Productions	1/9/66
10847	Planet Film Productions|Protelco	6/20/66
10849	Universal Pictures	12/16/66
10850	Warner Bros.	2/23/66
10851	High Road	6/22/66
10852	Eden Productions Inc.	5/31/66
10853	NaN	3/29/66
10854	Horizon Pictures|Columbia Pictures Corporation	2/17/66
10855	Universal Pictures	1/20/66
10856	Walt Disney Pictures	2/16/66
10857	Paramount Pictures|Solar Productions|Embassy P...	6/10/66
10858	The Mirisch Corporation	5/25/66
10859	Gibraltar Productions|Joel Productions|John Fr...	10/5/66
10860	Peter Rogers Productions|Anglo-Amalgamated Fil...	5/20/66
10861	Bruce Brown Films	6/15/66
10862	Cherokee Productions|Joel Productions|Douglas ...	12/21/66
10863	Mosfilm	1/1/66
10864	Benedict Pictures Corp.	11/2/66
10865	Norm-Iris	11/15/66

	vote_count	vote_average	release_year	budget_adj	revenue_adj
48	481	5.3	2015	2.759999e+07	0.0
67	280	5.4	2015	1.839999e+07	0.0
74	27	5.1	2015	0.000000e+00	0.0
75	569	7.7	2015	0.000000e+00	0.0
92	11	5.4	2015	0.000000e+00	0.0
93	181	4.1	2015	9.199996e+06	0.0
100	475	7.0	2015	0.000000e+00	0.0
101	161	5.4	2015	0.000000e+00	0.0
103	114	5.6	2015	0.000000e+00	0.0
116	22	4.5	2015	0.000000e+00	0.0
122	169	5.1	2015	0.000000e+00	0.0
133	43	6.3	2015	1.012000e+07	0.0
140	59	4.5	2015	0.000000e+00	0.0
143	11	7.5	2015	0.000000e+00	0.0
145	220	6.3	2015	1.655999e+06	0.0
147	148	6.6	2015	0.000000e+00	0.0
149	100	5.8	2015	1.839999e+07	0.0
151	114	5.3	2015	0.000000e+00	0.0
152	371	6.0	2015	0.000000e+00	0.0
154	211	5.7	2015	1.104000e+07	0.0
156	30	6.0	2015	1.379999e+07	0.0
158	139	7.2	2015	0.000000e+00	0.0
159	487	7.7	2015	5.796172e+05	0.0
164	131	5.8	2015	0.000000e+00	0.0
165	252	4.8	2015	5.519998e+07	0.0

```
166          96        5.1        2015  0.000000e+00        0.0
169         167        6.0        2015  0.000000e+00        0.0
174          92        5.2        2015  0.000000e+00        0.0
175         262        6.7        2015  0.000000e+00        0.0
176          22        6.6        2015  0.000000e+00        0.0
...         ...        ...         ...           ...        ...
10834        14        5.1        1966  0.000000e+00        0.0
10836        11        5.8        1966  0.000000e+00        0.0
10837        12        5.5        1966  0.000000e+00        0.0
10838        21        6.0        1966  0.000000e+00        0.0
10839        49        7.2        1966  0.000000e+00        0.0
10840        13        5.7        1966  0.000000e+00        0.0
10841        12        5.5        1966  5.038511e+05        0.0
10842        12        7.9        1966  0.000000e+00        0.0
10843        12        5.8        1966  0.000000e+00        0.0
10844        13        5.6        1966  0.000000e+00        0.0
10845        15        5.9        1966  0.000000e+00        0.0
10846        16        5.7        1966  0.000000e+00        0.0
10847        13        5.3        1966  0.000000e+00        0.0
10849        14        6.1        1966  0.000000e+00        0.0
10850        14        6.0        1966  0.000000e+00        0.0
10851        15        6.6        1966  0.000000e+00        0.0
10852        11        6.0        1966  0.000000e+00        0.0
10853        26        6.2        1966  0.000000e+00        0.0
10854        17        6.0        1966  0.000000e+00        0.0
10855        14        6.1        1966  4.702610e+06        0.0
10856        14        5.7        1966  0.000000e+00        0.0
10857        10        5.9        1966  0.000000e+00        0.0
10858        11        5.5        1966  0.000000e+00        0.0
10859        22        6.6        1966  0.000000e+00        0.0
10860        13        7.0        1966  0.000000e+00        0.0
10861        11        7.4        1966  0.000000e+00        0.0
10862        20        5.7        1966  0.000000e+00        0.0
10863        11        6.5        1966  0.000000e+00        0.0
10864        22        5.4        1966  0.000000e+00        0.0
10865        15        1.5        1966  1.276423e+05        0.0

[6016 rows x 21 columns]
```

In [6]:
```
#Lookup specific imdb_ids to check for uniqueness and duplication
#check for invalid genre values
#df[df["imdb_id"] == "tt2231253"]

df[df["imdb_id"] == "tt3247714"]

df[df["genres"] == "..."]
```

Out[6]:
```
Empty DataFrame
Columns: [id, imdb_id, popularity, budget, revenue, original_title, cast, homepage, dire
```

```
Index: []

[0 rows x 21 columns]
```

1.2.2 Data Cleaning (Remove Data which has no revenue or neglible revenue reporting and remove data prior to 2010)

In [7]:
```
# Ignoring data with no revenue numbers or < 50K
df = df[df["revenue"] >= 50000 ]
# Ignoring data with no budget numbers or < 50K
df = df[df["budget"] >= 50000 ]
# Ignoring data with release date before 1/1/201
df = df[df["release_year"] > 2009 ]
df.describe()
```

Out[7]:
```
                  id   popularity        budget       revenue      runtime  \
count    1002.000000  1002.000000  1.002000e+03  1.002000e+03  1002.000000
mean   120569.783433     1.777016  4.737870e+07  1.437266e+08   108.443114
std     87454.687500     2.274152  5.513546e+07  2.283294e+08    18.152073
min       189.000000     0.010335  1.000000e+05  5.013600e+04    62.000000
25%     49691.750000     0.686944  1.183250e+07  1.379664e+07    96.000000
50%     82684.500000     1.120216  2.761000e+07  6.006237e+07   106.000000
75%    192139.750000     2.059014  6.000000e+07  1.647489e+08   118.000000
max    417859.000000    32.985763  4.250000e+08  2.068178e+09   338.000000

        vote_count  vote_average  release_year    budget_adj   revenue_adj
count  1002.000000   1002.000000   1002.000000  1.002000e+03  1.002000e+03
mean    901.056886      6.174152   2012.434132  4.504786e+07  1.362567e+08
std    1179.320957      0.776002      1.711497  5.251779e+07  2.155749e+08
min      10.000000      2.200000   2010.000000  9.199996e+04  4.612510e+04
25%     180.000000      5.700000   2011.000000  1.093544e+07  1.289249e+07
50%     445.000000      6.200000   2012.000000  2.577527e+07  5.648651e+07
75%    1141.250000      6.700000   2014.000000  5.698466e+07  1.570550e+08
max    9767.000000      8.200000   2015.000000  4.250000e+08  1.902723e+09
```

In [8]:
```
# After discussing the structure of the data and any problems that need to be
#   cleaned, perform those cleaning steps in the second part of this section.

#All data cleaning done above, checking null value fields again
pd.isnull(df).sum()
```

Out[8]:
```
id                        0
imdb_id                   0
popularity                0
budget                    0
revenue                   0
original_title            0
cast                      1
homepage                351
```

```
director                 0
tagline                 68
keywords                47
overview                 0
runtime                  0
genres                   0
production_companies     4
release_date             0
vote_count               0
vote_average             0
release_year             0
budget_adj               0
revenue_adj              0
dtype: int64
```

Exploratory Data Analysis

1.2.3 Research Question 1: Which genres of movies made the highest revenue for movies released from 2010 to 2015?

In [9]:
```
# Use this, and more code cells, to explore your data. Don't forget to add
#   Markdown cells to document your observations and findings.

#Which Genres of movie make the highest revenue?

a = df['genres'].unique()

#np.sort(a)

#print(a)
myGenreList = []

# collect data from and process each row
for item in a:
    # set up a dictionary to hold the values for the cleaned and trimmed
            # data point
    b = item.split("|")
    myGenreList.extend(b)

#print("out of loop")
#print (myGenreList)
myGListDF = pd.DataFrame(myGenreList, columns=["Genres"])
myUniqueGenres = myGListDF["Genres"].unique()
np.sort(myUniqueGenres)
print(myUniqueGenres)
myUniqueGListDF = pd.DataFrame(myUniqueGenres, columns=["UniqueGenres"])
myUniqueGListDF = myUniqueGListDF.assign(Revenue = 0.0)
myUniqueGListDF.head()
```

```
    #print('step1')


    # collect data from and process each row
    for index1, row1 in df.iterrows():
        item = row1['genres']
        genres = item.split("|")
        #print('iteration loop 1')
        for tmp in genres:
            #print('loop 2')
            for index2, row2 in myUniqueGListDF.iterrows():
                #print('loop 3')
                if row2['UniqueGenres'] == tmp:
                    #print('adding revenue')
                    myUniqueGListDF.at[index2, 'Revenue'] = myUniqueGListDF['Revenue'][index
    #print('step2')
    myUniqueGListDF = myUniqueGListDF.sort_values(by='Revenue', ascending=False)
    myUniqueGListDF.head()
```

```
['Action' 'Adventure' 'Science Fiction' 'Thriller' 'Fantasy' 'Crime'
 'Western' 'Drama' 'Family' 'Animation' 'Comedy' 'Mystery' 'Romance' 'War'
 'History' 'Music' 'Horror' 'Documentary' 'Foreign']
```

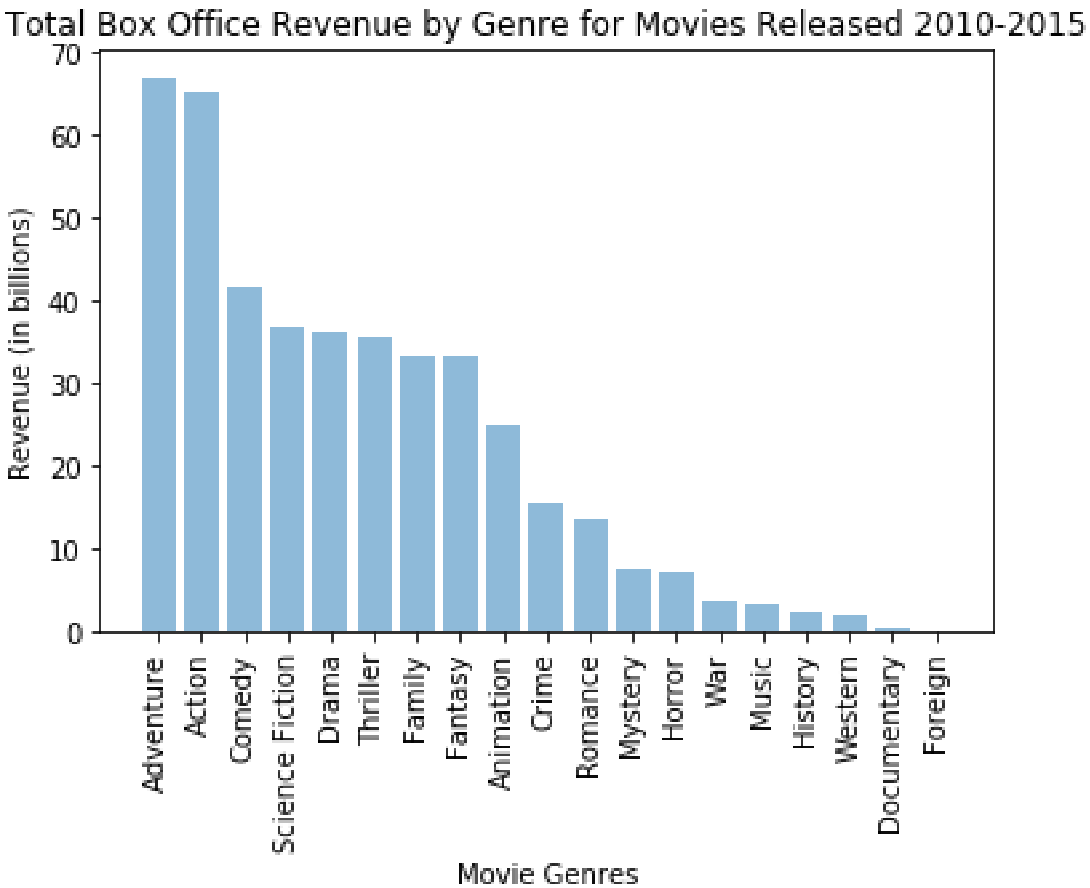
Out[9]:

	UniqueGenres	Revenue
1	Adventure	66.943523
0	Action	65.199387
10	Comedy	41.442885
2	Science Fiction	36.889324
7	Drama	36.199946

In [10]:

```
#Plot Bar Chart

import matplotlib.pyplot as plt; plt.rcdefaults()

# this is a 'magic word' that allows for plots to be displayed
# inline with the notebook. If you want to know more, see:
# http://ipython.readthedocs.io/en/stable/interactive/magics.html
%matplotlib inline

# example histogram, data taken from bay area sample
import numpy as np

objects = myUniqueGListDF['UniqueGenres']
count = objects.count()
y_pos = np.arange(count )
```

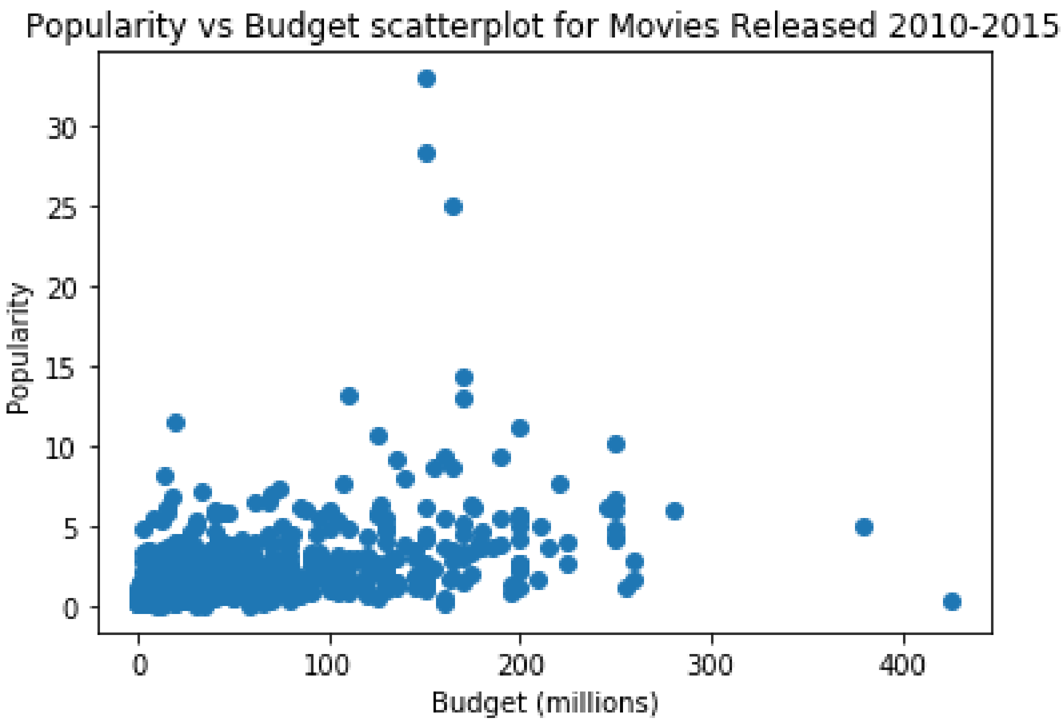
```
data = myUniqueGListDF['Revenue']

plt.bar(y_pos, data, align='center', alpha=0.5)
plt.xticks(y_pos, objects, rotation=90)
plt.ylabel('Revenue (in billions)')
plt.xlabel('Movie Genres')
plt.title('Total Box Office Revenue by Genre for Movies Released 2010-2015')

plt.show()
```



1.2.4 Is there any correlation between movie budget and popularity for movies released from 2010 to 2015

```
In [11]: #Plot Scatter Plot

         x = df['budget']/1000000.0;
         y = df['popularity']
         #colors = np.random.rand(N)
         #area = np.pi * (15 * np.random.rand(N))**2  # 0 to 15 point radii
```

```
plt.ylabel('Popularity')
plt.xlabel('Budget (millions)')
plt.title('Popularity vs Budget scatterplot for Movies Released 2010-2015')
#plt.scatter(x, y, s=area, c=colors, alpha=0.5)
plt.scatter(x, y)
plt.show()


#plt.xticks(y_pos, objects, rotation=90)

#plt.title('Total Box Office Revenue by Genre from 2013 to Present')
```



Limitations and Challenges

#1: The data is little outdated (not current), the last release year that data was available on was for 2015. So a more accurate treand analysis based cannot be performed with this data since we need current data..

#2: It is not clear if the poularity numbers are very accurate or comparable across movies, as the vote count has a large variation across the different titles and is low for many of the movies. The 50% percentile vote count is only 445 and the range is 10 - 9767.

#3: There were some challenges I faced in identifying the missing data and in some cases unrealistic low revenue numbers for cetain movies skewed the data and it was not clear of the data was erroneous or an outlier, so these data rows had to be identified and ignored.

Conclusions

#1: From the first research question we can conclude that the most popular genres of movies for the years 2010 to 2015 has been - Adventure, Action, Comedy, Science Fiction, Drama and Thriller The least popular genres of movies for the for the years 2010 to 2015 have been - Foreign, Documentary, Western, History and Music

#2: From the second research questions it is clear that there is no significant correlation between the budget of a movie and the popularity of the movie. However it can be seen that a higher percentage of low budget movies are unpopular. Also the most popular movies have been mid-budget movies.

```
In [12]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])

Out[12]: 0
```