

Project: Investigate The Movie Database (TMDb) Dataset

Table of Contents

- [Introduction](#)
- [Data Wrangling](#)
- [Exploratory Data Analysis](#)
- [Research Question 1 - Do movie stars move to bigger budget films as their career progresses?](#)
- [Research Question 2 - Is the TMDb popularity linked to the revenue taken by the film?](#)
- [Conclusions](#)

Introduction

I have chose to investigate the TMDb (The Movie Database) dataset, as this is the one which I have the most previous knowledge and interest in. The database includes information on over 10,000 movies, including details such as the cast and the budget.

This Jupyter notebook was created with **python version 2.7.15**.

Out[1]:

The raw code for this IPython notebook is by default hidden for easier reading. To toggle on/off the raw code, click [here](#).

Data Wrangling

First of all I will load the csv (Comma Separated Values) dataset into pandas, before looking at the start and end of the dataset.

General Properties

Out[2]:

<bound method DataFrame.head of udget	method					b
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	
5	281957	tt1663202	9.110700	135000000	532950503	
6	87101	tt1340138	8.654359	155000000	440603537	
7	286217	tt3659388	7.667400	108000000	595380321	
8	211672	tt2293640	7.404165	74000000	1156730962	
9	150540	tt2096673	6.326804	175000000	853708609	
10	206647	tt2379713	6.200282	245000000	880674609	
11	76757	tt1617661	6.189369	176000003	183987723	
12	264660	tt0470752	6.118847	15000000	36869414	
13	257344	tt2120120	5.984995	88000000	243637091	
14	99861	tt2395427	5.944927	280000000	1405035767	
15	273248	tt3460252	5.898400	44000000	155760117	
16	260346	tt2446042	5.749758	48000000	325771424	
17	102899	tt0478970	5.573184	130000000	518602163	
18	150689	tt1661199	5.556818	95000000	542351353	
19	131634	tt1951266	5.476958	160000000	650523427	
20	158852	tt1964418	5.462138	190000000	209035668	
21	307081	tt1798684	5.337064	30000000	91709827	
22	254128	tt2126355	4.907832	110000000	470490832	
23	216015	tt2322441	4.710402	40000000	569651467	
24	318846	tt1596363	4.648046	28000000	133346506	
25	177677	tt2381249	4.566713	150000000	682330139	
26	214756	tt2637276	4.564549	68000000	215863606	
27	207703	tt2802144	4.503789	81000000	403802136	
28	314365	tt1895587	4.062293	20000000	88346473	
29	294254	tt4046784	3.968891	61000000	311256926	
...	
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	
10848	2161	tt0060397	0.207257	5115000	120000000	
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 \		
0				Jurassic World		
1				Mad Max: Fury Road		
2				Insurgent		
3				Star Wars: The Force Awakens		
4				Furious 7		
5				The Revenant		
6				Terminator Genisys		
7				The Martian		
8				Minions		
9				Inside Out		
10				Spectre		
11				Jupiter Ascending		
12				Ex Machina		
13				Pixels		
14				Avengers: Age of Ultron		
15				The Hateful Eight		
16				Taken 3		
17				Ant-Man		
18				Cinderella		
19				The Hunger Games: Mockingjay - Part 2		
20				Tomorrowland		
21				Southpaw		
22				San Andreas		
23				Fifty Shades of Grey		
24				The Big Short		
25				Mission: Impossible - Rogue Nation		
26				Ted 2		
27				Kingsman: The Secret Service		
28				Spotlight		
29				Maze Runner: The Scorch Trials		
...				...		
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		
10848				Fantastic Voyage		
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 \

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 ...
5	Leonardo DiCaprio Tom Hardy Will Poulter Domhn...
6	Arnold Schwarzenegger Jason Clarke Emilia Clar...
7	Matt Damon Jessica Chastain Kristen Wiig Jeff ...
8	Sandra Bullock Jon Hamm Michael Keaton Allison...
9	Amy Poehler Phyllis Smith Richard Kind Bill Ha...
10	Daniel Craig Christoph Waltz LÃ©a Seydoux Ralp...
11	Mila Kunis Channing Tatum Sean Bean Eddie Redm...
12	Domhnall Gleeson Alicia Vikander Oscar Isaac S...
13	Adam Sandler Michelle Monaghan Peter Dinklage ...
14	Robert Downey Jr. Chris Hemsworth Mark Ruffalo...
15	Samuel L. Jackson Kurt Russell Jennifer Jason ...
16	Liam Neeson Forest Whitaker Maggie Grace Famke...
17	Paul Rudd Michael Douglas Evangeline Lilly Cor...
18	Lily James Cate Blanchett Richard Madden Helen...
19	Jennifer Lawrence Josh Hutcherson Liam Hemswor...
20	Britt Robertson George Clooney Raffey Cassidy ...
21	Jake Gyllenhaal Rachel McAdams Forest Whitaker...
22	Dwayne Johnson Alexandra Daddario Carla Gugino...
23	Dakota Johnson Jamie Dornan Jennifer Ehle Eloi...
24	Christian Bale Steve Carell Ryan Gosling Brad ...
25	Tom Cruise Jeremy Renner Simon Pegg Rebecca Fe...
26	Mark Wahlberg Seth MacFarlane Amanda Seyfried ...
27	Taron Egerton Colin Firth Samuel L. Jackson Mi...
28	Mark Ruffalo Michael Keaton Rachel McAdams Lie...
29	Dylan O'Brien Kaya Scodelario Thomas Brodie-Sa...
...	...
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...
10848	Stephen Boyd Raquel Welch Edmond O'Brien Donal...
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 \
0	http://www.jurassicworld.com/
1	http://www.madmaxmovie.com/
2	http://www.thedivergentseries.movie/#insurgent
3	http://www.starwars.com/films/star-wars-episod...
4	http://www.furious7.com/
5	http://www.foxmovies.com/movies/the-revenant
6	http://www.terminatormovie.com/
7	http://www.foxmovies.com/movies/the-martian
8	http://www.minionsmovie.com/
9	http://movies.disney.com/inside-out
10	http://www.sonypictures.com/movies/spectre/
11	http://www.jupiterascending.com
12	http://exmachina-movie.com/
13	http://www.pixels-movie.com/
14	http://marvel.com/movies/movie/193/avengers_ag...
15	http://thehatefuleight.com/
16	http://www.taken3movie.com/
17	http://marvel.com/movies/movie/180/ant-man
18	NaN
19	http://www.thehungergames.movie/
20	http://movies.disney.com/tomorrowland
21	NaN
22	http://www.sanandreasmovie.com/
23	https://www.facebook.com/fiftyshadesofgreymovie
24	http://www.thebigshortmovie.com/
25	http://www.missionimpossible.com
26	NaN
27	http://www.kingsmanmovie.com/
28	http://www.spotlightthefilm.com
29	http://mazerunnermovies.com
...	...
10836	NaN
10837	NaN
10838	NaN
10839	NaN
10840	NaN
10841	NaN
10842	NaN
10843	NaN
10844	NaN
10845	NaN
10846	NaN
10847	NaN
10848	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	\
0	Colin Trevorrow	
1	George Miller	
2	Robert Schwentke	
3	J.J. Abrams	
4	James Wan	
5	Alejandro González Iñárritu	
6	Alan Taylor	
7	Ridley Scott	
8	Kyle Balda Pierre Coffin	
9	Pete Docter	
10	Sam Mendes	
11	Lana Wachowski Lilly Wachowski	
12	Alex Garland	
13	Chris Columbus	
14	Joss Whedon	
15	Quentin Tarantino	
16	Olivier Megaton	
17	Peyton Reed	
18	Kenneth Branagh	
19	Francis Lawrence	
20	Brad Bird	
21	Antoine Fuqua	
22	Brad Peyton	
23	Sam Taylor-Johnson	
24	Adam McKay	
25	Christopher McQuarrie	
26	Seth MacFarlane	
27	Matthew Vaughn	
28	Tom McCarthy	
29	Wes Ball	
...	...	
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	
10848	Richard Fleischer	
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	...	\
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	...	
5	(n. One who has returned, as if from the dead.)	...	
6	Reset the future	...	
7	Bring Him Home	...	
8	Before Gru, they had a history of bad bosses	...	
9	Meet the little voices inside your head.	...	
10	A Plan No One Escapes	...	
11	Expand your universe.	...	
12	There is nothing more human than the will to s...	...	
13	Game On.	...	
14	A New Age Has Come.	...	
15	No one comes up here without a damn good reason.	...	
16	It Ends Here	...	
17	Heroes Don't Get Any Bigger	...	
18	Midnight is just the beginning.	...	
19	The fire will burn forever.	...	
20	Imagine a world where nothing is impossible.	...	
21	Believe in Hope.	...	
22	A rescue pilot survived an earthquake, this is...	...	
23	Are you curious?	...	
24	This is a true story.	...	
25	Desperate Times. Desperate Measures.	...	
26	Ted is Coming, Again.	...	
27	Manners maketh man.	...	
28	Break the story. Break the silence.	...	
29	The Maze Was Just the Beginning.	...	
...	
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...	...	
10848	A Fantastic and Spectacular Voyage... Through	
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	\
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	
5	In the 1820s, a frontiersman, Hugh Glass, sets...	156	
6	The year is 2029. John Connor, leader of the r...	125	
7	During a manned mission to Mars, Astronaut Mar...	141	
8	Minions Stuart, Kevin and Bob are recruited by...	91	
9	Growing up can be a bumpy road, and it's no ex...	94	
10	A cryptic message from Bondâ€™s past sends him...	148	
11	In a universe where human genetic material is ...	124	
12	Caleb, a 26 year old coder at the world's larg...	108	
13	Video game experts are recruited by the milita...	105	
14	When Tony Stark tries to jumpstart a dormant p...	141	
15	Bounty hunters seek shelter from a raging bliz...	167	
16	Ex-government operative Bryan Mills finds his ...	109	
17	Armed with the astonishing ability to shrink i...	115	
18	When her father unexpectedly passes away, youn...	112	
19	With the nation of Panem in a full scale war, ...	136	
20	Bound by a shared destiny, a bright, optimisti...	130	
21	Billy "The Great" Hope, the reigning junior mi...	123	
22	In the aftermath of a massive earthquake in Ca...	114	
23	When college senior Anastasia Steele steps in ...	125	
24	The men who made millions from a global econom...	130	
25	Ethan and team take on their most impossible m...	131	
26	Newlywed couple Ted and Tami-Lynn want to have...	115	
27	The story of a super-secret spy organization t...	130	
28	The true story of how The Boston Globe uncover...	128	
29	Thomas and his fellow Gladers face their great...	132	
...	
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	
10848	The science of miniaturization has been unlock...	100	
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 \

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
5	Western Drama Adventure Thriller
6	Science Fiction Action Thriller Adventure
7	Drama Adventure Science Fiction
8	Family Animation Adventure Comedy
9	Comedy Animation Family
10	Action Adventure Crime
11	Science Fiction Fantasy Action Adventure
12	Drama Science Fiction
13	Action Comedy Science Fiction
14	Action Adventure Science Fiction
15	Crime Drama Mystery Western
16	Crime Action Thriller
17	Science Fiction Action Adventure
18	Romance Fantasy Family Drama
19	War Adventure Science Fiction
20	Action Family Science Fiction Adventure Mystery
21	Action Drama
22	Action Drama Thriller
23	Drama Romance
24	Comedy Drama
25	Action
26	Comedy
27	Crime Comedy Action Adventure
28	Drama Thriller History
29	Action Science Fiction Thriller
...	...
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

10848	Adventure Science Fiction
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	\
0	Universal Studios Amblin Entertainment Legenda...	6/9/15	
1	Village Roadshow Pictures Kennedy Miller Produ...	5/13/15	
2	Summit Entertainment Mandeville Films Red Wago...	3/18/15	
3	Lucasfilm Truenorth Productions Bad Robot	12/15/15	
4	Universal Pictures Original Film Media Rights ...	4/1/15	
5	Regency Enterprises Appian Way CatchPlay Anony...	12/25/15	
6	Paramount Pictures Skydance Productions	6/23/15	
7	Twentieth Century Fox Film Corporation Scott F...	9/30/15	
8	Universal Pictures Illumination Entertainment	6/17/15	
9	Walt Disney Pictures Pixar Animation Studios W...	6/9/15	
10	Columbia Pictures Danjaq B24	10/26/15	
11	Village Roadshow Pictures Dune Entertainment A...	2/4/15	
12	DNA Films Universal Pictures International (UP...	1/21/15	
13	Columbia Pictures Happy Madison Productions	7/16/15	
14	Marvel Studios Prime Focus Revolution Sun Studios	4/22/15	
15	Double Feature Films The Weinstein Company Fil...	12/25/15	
16	Twentieth Century Fox Film Corporation M6 Film...	1/1/15	
17	Marvel Studios	7/14/15	
18	Walt Disney Pictures Genre Films Beagle Pug Fi...	3/12/15	
19	Studio Babelsberg StudioCanal Lionsgate Walt D...	11/18/15	
20	Walt Disney Pictures Babieka A113	5/19/15	
21	Escape Artists Riche-Ludwig Productions	6/15/15	
22	New Line Cinema Village Roadshow Pictures Warn...	5/27/15	
23	Focus Features Trigger Street Productions Mich...	2/11/15	
24	Paramount Pictures Plan B Entertainment Regenc...	12/11/15	
25	Paramount Pictures Skydance Productions China ...	7/23/15	
26	Universal Pictures Media Rights Capital Fuzzy ...	6/25/15	
27	Twentieth Century Fox Film Corporation Marv Fi...	1/24/15	
28	Participant Media Open Road Films Anonymous Co...	11/6/15	
29	Gotham Group Temple Hill Entertainment TSG Ent...	9/9/15	
...	
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
10848	Twentieth Century Fox Film Corporation		8/24/66
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
0	5562	6.5	2015	1.379999e+08	1.392446e+09
1	6185	7.1	2015	1.379999e+08	3.481613e+08
2	2480	6.3	2015	1.012000e+08	2.716190e+08
3	5292	7.5	2015	1.839999e+08	1.902723e+09
4	2947	7.3	2015	1.747999e+08	1.385749e+09
5	3929	7.2	2015	1.241999e+08	4.903142e+08
6	2598	5.8	2015	1.425999e+08	4.053551e+08
7	4572	7.6	2015	9.935996e+07	5.477497e+08
8	2893	6.5	2015	6.807997e+07	1.064192e+09
9	3935	8.0	2015	1.609999e+08	7.854116e+08
10	3254	6.2	2015	2.253999e+08	8.102203e+08
11	1937	5.2	2015	1.619199e+08	1.692686e+08
12	2854	7.6	2015	1.379999e+07	3.391985e+07
13	1575	5.8	2015	8.095996e+07	2.241460e+08
14	4304	7.4	2015	2.575999e+08	1.292632e+09
15	2389	7.4	2015	4.047998e+07	1.432992e+08
16	1578	6.1	2015	4.415998e+07	2.997096e+08
17	3779	7.0	2015	1.195999e+08	4.771138e+08
18	1495	6.8	2015	8.739996e+07	4.989630e+08
19	2380	6.5	2015	1.471999e+08	5.984813e+08
20	1899	6.2	2015	1.747999e+08	1.923127e+08
21	1386	7.3	2015	2.759999e+07	8.437300e+07
22	2060	6.1	2015	1.012000e+08	4.328514e+08
23	1865	5.3	2015	3.679998e+07	5.240791e+08
24	1545	7.3	2015	2.575999e+07	1.226787e+08
25	2349	7.1	2015	1.379999e+08	6.277435e+08
26	1666	6.3	2015	6.255997e+07	1.985944e+08
27	3833	7.6	2015	7.451997e+07	3.714978e+08
28	1559	7.8	2015	1.839999e+07	8.127872e+07
29	1849	6.4	2015	5.611998e+07	2.863562e+08
...
10836	11	5.8	1966	0.000000e+00	0.000000e+00
10837	12	5.5	1966	0.000000e+00	0.000000e+00
10838	21	6.0	1966	0.000000e+00	0.000000e+00
10839	49	7.2	1966	0.000000e+00	0.000000e+00
10840	13	5.7	1966	0.000000e+00	0.000000e+00
10841	12	5.5	1966	5.038511e+05	0.000000e+00
10842	12	7.9	1966	0.000000e+00	0.000000e+00
10843	12	5.8	1966	0.000000e+00	0.000000e+00

10844	13	5.6	1966	0.000000e+00	0.000000e+00
10845	15	5.9	1966	0.000000e+00	0.000000e+00
10846	16	5.7	1966	0.000000e+00	0.000000e+00
10847	13	5.3	1966	0.000000e+00	0.000000e+00
10848	42	6.7	1966	3.436265e+07	8.061618e+07
10849	14	6.1	1966	0.000000e+00	0.000000e+00
10850	14	6.0	1966	0.000000e+00	0.000000e+00
10851	15	6.6	1966	0.000000e+00	0.000000e+00
10852	11	6.0	1966	0.000000e+00	0.000000e+00
10853	26	6.2	1966	0.000000e+00	0.000000e+00
10854	17	6.0	1966	0.000000e+00	0.000000e+00
10855	14	6.1	1966	4.702610e+06	0.000000e+00
10856	14	5.7	1966	0.000000e+00	0.000000e+00
10857	10	5.9	1966	0.000000e+00	0.000000e+00
10858	11	5.5	1966	0.000000e+00	0.000000e+00
10859	22	6.6	1966	0.000000e+00	0.000000e+00
10860	13	7.0	1966	0.000000e+00	0.000000e+00
10861	11	7.4	1966	0.000000e+00	0.000000e+00
10862	20	5.7	1966	0.000000e+00	0.000000e+00
10863	11	6.5	1966	0.000000e+00	0.000000e+00
10864	22	5.4	1966	0.000000e+00	0.000000e+00
10865	15	1.5	1966	1.276423e+05	0.000000e+00

[10866 rows x 21 columns]>

Looking at the start and end of the data, it can be seen that the budget and revenue data is missing for older films, but as this data may be useful, it is worth keeping the data series. The homepage, tagline, keywords and overview columns are long columns which are unlikely to provide usable data, so I plan to drop these to make the table easier to read. I am also unlikely to use the imdb_id column, so this will also be dropped. I will also output the first row of data to see a clearer sample of the headings and respective data.

Note - Only run this cell once, otherwise you will get an error as it tries to delete the columns again.

Data Cleaning

id		135397
popularity		32.9858
budget		150000000
revenue		1513528810
original_title		Jurassic World
cast	Chris Pratt Bryce Dallas Howard Irrfan Khan Vi...	
director		Colin Trevorrow
runtime		124
genres	Action Adventure Science Fiction Thriller	
production_companies	Universal Studios Amblin Entertainment Legenda...	
release_date		6/9/15
vote_count		5562
vote_average		6.5
release_year		2015
budget_adj		1.38e+08
revenue_adj		1.39245e+09
Name:	0, dtype:	object

The cast, genres and production_companies look like interesting datasets to probe, so I am going to add total counts of the data in these columns to 3 new columns. First of all I need to check for any empty series, as I will be adding 1 to the count of '|' characters, which will only be present when there is more than one item.

```
Cast length: 10866  
Genres length: 10866  
Production companies length: 10866
```

Then I count the separators and add one column to the dataset for the count of items in each series.

Note - Only run this cell once, otherwise you will get an error as it tries to create the columns again.

Again print the first datapoint and the data types statistics for the dataset.

```

id                                135397
popularity                         32.9858
budget                            1500000000
revenue                           1513528810
original_title                     Jurassic World
cast                               Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...
director                          Colin Trevorrow
runtime                            124
genres                             Action|Adventure|Science Fiction|Thriller
production_companies               Universal Studios|Amblin Entertainment|Legenda...
release_date                        6/9/15
vote_count                          5562
vote_average                         6.5
release_year                         2015
budget_adj                           1.38e+08
revenue_adj                           1.39245e+09
cast_count                           5
genre_count                           4
companies_count                      5
Name: 0, dtype: object
id                                int64
popularity                         float64
budget                            int64
revenue                           int64
original_title                     object
cast                               object
director                          object
runtime                            int64
genres                             object
production_companies               object
release_date                        object
vote_count                          int64
vote_average                         float64
release_year                         int64
budget_adj                           float64
revenue_adj                           float64
cast_count                           float64
genre_count                           float64
companies_count                      float64
dtype: object

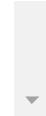
```

Now change the data types in the data frame, to their respective data types - integers, floats and dates.

```
<bound method Series.head of 0      137999939
1      137999939
2      101199955
3      183999919
4      174799923
5      124199945
6      142599937
7      99359956
8      68079970
9      160999929
10     225399900
11     161919931
12     13799993
13     80959964
14     257599886
15     40479982
16     44159980
17     119599947
18     87399961
19     147199935
20     174799923
21     27599987
22     101199955
23     36799983
24     25759988
25     137999939
26     62559972
27     74519967
28     18399991
29     56119975
...
10836     0
10837     0
10838     0
10839     0
10840     0
10841     503851
10842     0
10843     0
10844     0
10845     0
10846     0
10847     0
10848     34362645
10849     0
10850     0
10851     0
10852     0
10853     0
10854     0
10855     4702610
10856     0
10857     0
10858     0
10859     0
10860     0
10861     0
10862     0
10863     0
10864     0
```

```
10865      127642
Name: budget_adj, Length: 10866, dtype: int64>
<bound method Series.head of 0          1392445892
1      348161292
2      271619025
3      1902723129
4      1385748801
5      490314247
6      405355075
7      547749654
8      1064192016
9      785411574
10     810220283
11     169268630
12     33919845
13     224146025
14     1292632336
15     143299244
16     299709578
17     477113780
18     498963025
19     598481289
20     192312729
21     84373003
22     432851374
23     524079119
24     122678731
25     627743451
26     198594430
27     371497801
28     81278719
29     286356245
...
10836      0
10837      0
10838      0
10839      0
10840      0
10841      0
10842      0
10843      0
10844      0
10845      0
10846      0
10847      0
10848      80616176
10849      0
10850      0
10851      0
10852      0
10853      0
10854      0
10855      0
10856      0
10857      0
10858      0
10859      0
10860      0
10861      0
10862      0
10863      0
```

```
10864      0
10865      0
Name: revenue_adj, Length: 10866, dtype: int64>
```



Note the dates are in US format - months/days/years

Default 2digit dates in python for 19xx are from 1969, so we will need to change the dates before 1969, from 2066, 2067 & 2068 to 1966, 1967 & 1968 respectively. As the dates are in the range 1966 to 2015, I know all dates after 2015 should infact begin with 19xx, so I will subtract 100 years off of all dates with a year greater than today (2018).

```
<bound method Series.head of 0      2015-06-09
1      2015-05-13
2      2015-03-18
3      2015-12-15
4      2015-04-01
5      2015-12-25
6      2015-06-23
7      2015-09-30
8      2015-06-17
9      2015-06-09
10     2015-10-26
11     2015-02-04
12     2015-01-21
13     2015-07-16
14     2015-04-22
15     2015-12-25
16     2015-01-01
17     2015-07-14
18     2015-03-12
19     2015-11-18
20     2015-05-19
21     2015-06-15
22     2015-05-27
23     2015-02-11
24     2015-12-11
25     2015-07-23
26     2015-06-25
27     2015-01-24
28     2015-11-06
29     2015-09-09
...
10836    1966-01-01
10837    1966-06-21
10838    1966-11-01
10839    1966-10-27
10840    1966-12-22
10841    1966-10-23
10842    1966-01-01
10843    1966-06-09
10844    1966-01-16
10845    1966-03-01
10846    1966-01-09
10847    1966-06-20
10848    1966-08-24
10849    1966-12-16
10850    1966-02-23
10851    1966-06-22
10852    1966-05-31
10853    1966-03-29
10854    1966-02-17
10855    1966-01-20
10856    1966-02-16
10857    1966-06-10
10858    1966-05-25
10859    1966-10-05
10860    1966-05-20
10861    1966-06-15
10862    1966-12-21
10863    1966-01-01
10864    1966-11-02
```

```
10865 1966-11-15
Name: release_date, Length: 10866, dtype: datetime64[ns]>
<bound method Series.head of 0          2015
1      2015
2      2015
3      2015
4      2015
5      2015
6      2015
7      2015
8      2015
9      2015
10     2015
11     2015
12     2015
13     2015
14     2015
15     2015
16     2015
17     2015
18     2015
19     2015
20     2015
21     2015
22     2015
23     2015
24     2015
25     2015
26     2015
27     2015
28     2015
29     2015
...
10836 1966
10837 1966
10838 1966
10839 1966
10840 1966
10841 1966
10842 1966
10843 1966
10844 1966
10845 1966
10846 1966
10847 1966
10848 1966
10849 1966
10850 1966
10851 1966
10852 1966
10853 1966
10854 1966
10855 1966
10856 1966
10857 1966
10858 1966
10859 1966
10860 1966
10861 1966
10862 1966
10863 1966
```

```
10864    1966
10865    1966
Name: release_year, Length: 10866, dtype: int64>
```

I then sorted the dataset in date format, before then displaying the start and end of the dataset, to ensure that I parsed all of the dates correctly.

<bound method DataFrame.head of			id	popularity	budget
revenue \					
10150	6643	0.421043	0	0	
10152	23220	0.333643	0	0	
10162	18780	0.194948	0	0	
10166	21240	0.136525	0	0	
10156	1673	0.267118	0	0	
10172	1941	0.256779	0	0	
10147	11167	0.569424	0	0	
10155	43040	0.269428	0	0	
10158	31102	0.254124	0	0	
10144	284	0.947307	3000000	25000000	
10165	22013	0.138777	0	0	
10153	23439	0.323180	270000	0	
10163	1908	0.162753	0	0	
10167	11773	0.127219	200000	0	
10164	2982	0.144106	0	0	
10168	15788	0.114188	0	10000000	
10169	29756	0.110511	0	0	
10149	299	0.423531	0	0	
10141	539	2.610362	806948	32000000	
10146	2134	0.712389	750000	0	
10154	37230	0.271858	0	0	
10171	39890	0.065808	0	0	
10160	24452	0.225218	30000	0	
10143	967	1.136943	12000000	60000000	
10145	24014	0.875173	0	0	
10159	12617	0.232574	0	0	
10142	966	1.872132	2000000	4905000	
10161	11209	0.310398	0	0	
10170	18973	0.055821	3000000	7100000	
10157	18639	0.266435	0	0	
...	
284	364067	0.439598	0	0	
293	277713	0.421790	0	0	
528	369058	0.133921	0	0	
431	325173	0.239496	0	0	
364	329829	0.390188	0	0	
24	318846	4.648046	28000000	133346506	
165	347969	0.913085	60000000	0	
479	367551	0.283650	0	0	
462	309298	0.204356	0	0	
559	373977	0.031635	0	0	
3	140607	11.173104	200000000	2068178225	
95	258509	1.841779	0	233755553	
117	266294	1.376605	30000000	105011053	
173	326425	0.872149	0	16775	
92	370687	1.876037	0	0	
450	366018	0.215386	0	0	
455	373558	0.209036	0	0	
290	352208	0.427116	0	0	
515	395560	0.142759	1300000	0	
53	274479	2.793297	60000000	101134059	
143	378373	1.128081	0	0	
5	281957	9.110700	135000000	532950503	
91	274167	1.904898	50000000	240360392	
15	273248	5.898400	44000000	155760117	
553	371759	0.068741	0	0	
540	371758	0.112284	0	0	
569	371833	0.088764	0	0	

144	291270	1.125239	8000000	3759286
322	362057	0.372889	0	0
190	382517	0.773909	0	0
			original_title \	
10150			The Unforgiven	
10152			The Brides of Dracula	
10162			Sink the Bismarck!	
10166			Carry On Constable	
10156			Comanche Station	
10172			Exodus	
10147			Peeping Tom	
10155			Sergeant Rutledge	
10158			Pollyanna	
10144			The Apartment	
10165			Elmer Gantry	
10153			House of Usher	
10163			Inherit the Wind	
10167			Village of the Damned	
10164			The Lost World	
10168			The Bellboy	
10169			13 Ghosts	
10149			Ocean's Eleven	
10141			Psycho	
10146			The Time Machine	
10154			Saturday Night and Sunday Morning	
10171			The City of the Dead	
10160			The Little Shop of Horrors	
10143			Spartacus	
10145			Let's Make Love	
10159			Midnight Lace	
10142			The Magnificent Seven	
10161			The Alamo	
10170			Cinderfella	
10157			Flaming Star	
...			...	
284			A Very Murray Christmas	
293			Sunset Song	
528			Merry Kissmas	
431			Close Range	
364			The Girl King	
24			The Big Short	
165			The Ridiculous 6	
479			American Hero	
462			Don Verdean	
559			Childhood's End	
3			Star Wars: The Force Awakens	
95			Alvin and the Chipmunks: The Road Chip	
117			Sisters	
173			Extraction	
92			Mythica: The Necromancer	
450			A Christmas Melody	
455			Taylor Swift: The 1989 World Tour - Live	
290			Where to Invade Next	
515			Capsule	
53			Joy	
143			Brothers of the Wind	
5			The Revenant	
91			Daddy's Home	
15			The Hateful Eight	
553			Doctor Who: The Husbands of River Song	

540	And Then There Were None
569	Harry Price: Ghost Hunter
144	Anomalisa
322	Martyrs
190	Open Season: Scared Silly

	cast \
10150	Burt Lancaster Audrey Hepburn Audie Murphy Joh...
10152	Peter Cushing Martita Hunt Yvonne Monlaur Fred...
10162	Kenneth More Dana Wynter Carl MÃ¶hner Laurence...
10166	Kenneth Connor Charles Hawtrey Sid James Kenne...
10156	Randolph Scott Nancy Gates Skip Homeier Dyke J...
10172	Paul Newman Eva Marie Saint Lee J. Cobb Ralph ...
10147	Karlheinz BÃ¶hm Moira Shearer Anna Massey Maxi...
10155	Jeffrey Hunter Woody Strode Constance Towers B...
10158	Hayley Mills Jane Wyman Richard Egan Karl Mald...
10144	Jack Lemmon Shirley MacLaine Fred MacMurray Ra...
10165	Burt Lancaster Jean Simmons Arthur Kennedy Dea...
10153	Vincent Price Mark Damon Myrna Fahey Harry Ell...
10163	Spencer Tracy Fredric March Gene Kelly Dick Yo...
10167	George Sanders Barbara Shelley Martin Stephens...
10164	Michael Rennie Jill St. John David Hedison Cla...
10168	Jerry Lewis Alex Gerry Bob Clayton Sonnie Sand...
10169	Charles Herbert Jo Morrow Martin Milner Rosema...
10149	Frank Sinatra Dean Martin Peter Lawford Angie ...
10141	Anthony Perkins Vera Miles John Gavin Janet Le...
10146	Rod Taylor Alan Young Yvette Mimieux Sebastian...
10154	Albert Finney Shirley Anne Field Rachel Robert...
10171	Christopher Lee Dennis Lotis Patricia Jessel T...
10160	Jackie Joseph Mel Welles Jack Nicholson Leola ...
10143	Kirk Douglas Laurence Olivier Jean Simmons Cha...
10145	Marilyn Monroe Yves Montand Tony Randall Frank...
10159	Doris Day Rex Harrison John Gavin Myrna Loy Ro...
10142	Yul Brynner Eli Wallach Steve McQueen Charles ...
10161	John Wayne Richard Widmark Laurence Harvey Fra...
10170	Jerry Lewis Ed Wynn Judith Anderson Henry Silv...
10157	Elvis Presley Steve Forrest Barbara Eden Dolor...
...	...
284	Bill Murray Paul Shaffer George Clooney Miley ...
293	Peter Mullan Agyness Deyn Kevin Guthrie Hugh R...
528	Karissa Staples Brant Daugherty Brittany Under...
431	Scott Adkins Nick Chinlund Caitlin Keats Jake ...
364	Malin Buska Sarah Gadon Michael Nyqvist Lucas ...
24	Christian Bale Steve Carell Ryan Gosling Brad ...
165	Adam Sandler Taylor Lautner Steve Buscemi Terr...
479	Stephen Dorff Eddie Griffin Bill Millions Jona...
462	Sam Rockwell Jemaine Clement Amy Ryan Danny Mc...
559	Mike Vogel Osy Ikhile Daisy Betts Georgina Hai...
3	Harrison Ford Mark Hamill Carrie Fisher Adam D...
95	Jason Lee Justin Long Bella Thorne Matthew Gra...
117	Tina Fey Amy Poehler John Leguizamo Ike Barinh...
173	Bruce Willis Gina Carano Kellan Lutz D. B. Swe...
92	Melanie Stone Adam Johnson Kevin Sorbo Nicola ...
450	Lacey Chabert Mariah Carey Brennan Elliott Kat...
455	Taylor Swift Ellen DeGeneres Cara Delevingne L...
290	Michael Moore Krista Kiuru Tim Walker
515	Edmund Kingsley David Wayman Nigel Barber Lisa...
53	Jennifer Lawrence Bradley Cooper Robert De Niro...
143	Manuel Camacho Jean Reno Tobias Moretti Eva Kuen
5	Leonardo DiCaprio Tom Hardy Will Poulter Domhn...
91	Will Ferrell Mark Wahlberg Linda Cardellini Th...

15 Samuel L. Jackson|Kurt Russell|Jennifer Jason ...
 553 Peter Capaldi|Alex Kingston|Greg Davies|Matt L...
 540 Maeve Dermody|Aidan Turner|Charles Dance|Dougl...
 569 Rafe Spall|Cara Theobold|Tom Ward|Sophie Stant...
 144 Jennifer Jason Leigh|David Thewlis|Tom Noonan
 322 Troian Bellisario|Bailey Noble|Kate Burton|Cai...
 190 Garry Chalk|Kathleen Barr|Willa Townsend|Melis...

	director	runtime	\
10150	John Huston	125	
10152	Terence Fisher	85	
10162	Lewis Gilbert	97	
10166	Gerald Thomas	86	
10156	Budd Boetticher	73	
10172	Otto Preminger	208	
10147	Michael Powell	101	
10155	John Ford	111	
10158	David Swift	134	
10144	Billy Wilder	125	
10165	Richard Brooks	146	
10153	Roger Corman	79	
10163	Stanley Kramer	128	
10167	Wolf Rilla	77	
10164	Irwin Allen	97	
10168	Jerry Lewis	72	
10169	William Castle	85	
10149	Lewis Milestone	127	
10141	Alfred Hitchcock	109	
10146	George Pal	103	
10154	Karel Reisz	89	
10171	John Llewellyn Moxey	76	
10160	Roger Corman	70	
10143	Stanley Kubrick	197	
10145	George Cukor	114	
10159	David Miller	110	
10142	John Sturges	128	
10161	John Wayne	167	
10170	Frank Tashlin	91	
10157	Don Siegel	101	
...	
284	Sofia Coppola	56	
293	Terence Davies	135	
528	Michael Feifer	90	
431	Isaac Florentine	80	
364	Mika Kaurismäki	106	
24	Adam McKay	130	
165	Frank Coraci	119	
479	Nick Love	86	
462	Jared Hess	96	
559	Nick Hurran	246	
3	J.J. Abrams	136	
95	Walt Becker	92	
117	Jason Moore	118	
173	Steven C. Miller	83	
92	A. Todd Smith	0	
450	Mariah Carey	83	
455	Jonas Åkerlund	132	
290	Michael Moore	120	
515	Andrew Martin	91	
53	David O. Russell	124	
143	Gerardo Olivares Otmar Penker	98	

5	Alejandro González Iñárritu	156
91	Sean Anders	96
15	Quentin Tarantino	167
553	Douglas Mackinnon	60
540	Craig Viveiros	168
569	Alex Pillai	90
144	Duke Johnson Charlie Kaufman	90
322	Kevin Goetz Michael Goetz	86
190	David Feiss	85

genres \

10150	Action Drama Western
10152	Horror
10162	Action Drama Foreign History War
10166	Comedy
10156	Action Western
10172	Action Drama History War
10147	Horror Thriller
10155	Crime Western
10158	Drama Family
10144	Comedy Drama Romance
10165	Drama
10153	Horror Thriller
10163	Drama History
10167	Horror Science Fiction
10164	Adventure Fantasy Science Fiction
10168	Comedy Family
10169	Horror
10149	Thriller Music Comedy Crime
10141	Drama Horror Thriller
10146	Thriller Adventure Fantasy Science Fiction Rom...
10154	Drama
10171	Horror
10160	Comedy
10143	Action Drama History
10145	Comedy Romance
10159	Thriller
10142	Action Adventure Western
10161	Action Adventure Drama History Western
10170	Comedy Romance
10157	Action Drama Romance Western
...	...
284	Comedy Music
293	Drama
528	Romance TV Movie
431	Crime Action
364	Drama History Romance
24	Comedy Drama
165	Comedy Western
479	Action Comedy Science Fiction
462	Comedy
559	Thriller TV Movie Science Fiction Drama
3	Action Adventure Science Fiction Fantasy
95	Adventure Animation Comedy Family
117	Comedy
173	Action Thriller
92	Fantasy Action Adventure
450	Romance Music TV Movie Comedy
455	Documentary Music
290	Documentary
515	Drama History Thriller Science Fiction

53		Comedy Drama	
143		Adventure Drama Family	
5		Western Drama Adventure Thriller	
91		Comedy	
15		Crime Drama Mystery Western	
553	Drama Science Fiction Comedy		
540	Mystery Drama		
569	Drama Horror		
144	Animation Comedy Drama Romance		
322	Drama Horror Thriller		
190	Animation Comedy Family Adventure		
production_companies release_date \			
10150	James Productions	1960-01-01	
10152	Hammer Film Productions Hotspur Film Productio...	1960-01-01	
10162	Twentieth Century Fox Film Corporation	1960-02-11	
10166	Peter Rogers Productions	1960-02-22	
10156	Columbia Pictures Corporation Ranown Pictures ...	1960-03-01	
10172	Carlyle Productions	1960-03-27	
10147	National Film Finance Corporation (NFFC) Anglo...	1960-04-06	
10155	John Ford Productions	1960-05-18	
10158	Walt Disney Productions	1960-05-19	
10144	United Artists The Mirisch Company	1960-06-15	
10165	United Artists	1960-06-19	
10153	Alta Vista Productions	1960-06-21	
10163	Stanley Kramer Productions	1960-06-24	
10167	Metro-Goldwyn-Mayer (MGM)	1960-07-01	
10164	Twentieth Century Fox Film Corporation	1960-07-13	
10168	Paramount Pictures Jerry Lewis Productions	1960-07-20	
10169	Columbia Pictures William Castle Productions	1960-08-05	
10149	Warner Bros. Dorchester	1960-08-10	
10141	Shamley Productions	1960-08-14	
10146	George Pal Productions Galaxy Films Inc.	1960-08-17	
10154	Woodfall Film Productions	1960-08-29	
10171	Vulcan Productions Inc.	1960-09-09	
10160	Santa Clara Productions	1960-09-14	
10143	Bryna Productions	1960-10-06	
10145	Twentieth Century Fox Film Corporation The Com...	1960-10-07	
10159	Arwin Productions Universal International Pict...	1960-10-13	
10142	The Mirisch Corporation Alpha Productions	1960-10-23	
10161	United Artists Batjac Productions The Alamo Co...	1960-10-24	
10170	Paramount Pictures Jerry Lewis Productions	1960-12-18	
10157	Twentieth Century Fox Film Corporation	1960-12-20	
...
284	American Zoetrope South Beach Productions Depa...	2015-12-04	
293	Iris productions Hurricane Films SellOutPictures	2015-12-04	
528		NaN	2015-12-05
431	Bleiberg Entertainment Compound B Underdogg En...	2015-12-11	
364	Starhaus Filmproduktion Anagram Produktion Gal...	2015-12-11	
24	Paramount Pictures Plan B Entertainment Regenc...	2015-12-11	
165	Happy Madison Productions	2015-12-11	
479	Screen Media Films Vertigo Films Protagonist P...	2015-12-11	
462	Lions Gate Films	2015-12-11	
559		NaN	2015-12-14
3	Lucasfilm Truenorth Productions Bad Robot	2015-12-15	
95	Regency Enterprises Fox 2000 Pictures SunSwept...	2015-12-17	
117	Little Stranger	2015-12-18	
173	Emmett / Furla Films Oasis Films Aperture Entert...	2015-12-18	
92	Arrowstorm Entertainment Camera 40 Productions...	2015-12-19	
450		NaN	2015-12-19
455	Apple Inc.	2015-12-20	

290	Dog Eat Dog Films IMG Films	2015-12-23
515	Ecaveo Capital Partners Hermes Space Industries	2015-12-23
53	Fox 2000 Pictures Davis Entertainment Annapurn...	2015-12-24
143	Terra Mater Factual Studios	2015-12-24
5	Regency Enterprises Appian Way CatchPlay Anony...	2015-12-25
91	Paramount Pictures Gary Sanchez Productions Go...	2015-12-25
15	Double Feature Films The Weinstein Company Fil...	2015-12-25
553	British Broadcasting Corporation (BBC)	2015-12-25
540	British Broadcasting Corporation (BBC) Mammoth...	2015-12-26
569	Bentley Productions	2015-12-27
144	Starburns Industries	2015-12-30
322	Blumhouse Productions The Safran Company Templ...	2015-12-31
190	Sony Pictures Animation	2015-12-31

	vote_count	vote_average	release_year	budget_adj	revenue_adj
10150	17	4.9	1960	0	0
10152	19	6.6	1960	0	0
10162	12	4.9	1960	0	0
10166	10	6.0	1960	0	0
10156	12	6.5	1960	0	0
10172	14	6.1	1960	0	0
10147	56	7.4	1960	0	0
10155	12	5.2	1960	0	0
10158	25	7.2	1960	0	0
10144	235	7.9	1960	22118902	184324186
10165	11	6.3	1960	0	0
10153	28	5.9	1960	1990701	0
10163	30	6.8	1960	0	0
10167	28	7.0	1960	1474593	0
10164	11	6.0	1960	0	0
10168	12	7.0	1960	0	73729674
10169	12	5.5	1960	0	0
10149	39	6.6	1960	0	0
10141	1180	8.0	1960	5949601	235934958
10146	101	7.3	1960	5529725	0
10154	12	5.6	1960	0	0
10171	13	6.1	1960	0	0
10160	23	6.0	1960	221189	0

10143	211	6.9	1960	88475609	442378047
10145	15	4.9	1960	0	0
10159	10	5.7	1960	0	0
10142	224	7.0	1960	14745934	36164405
10161	27	6.2	1960	0	0
10170	13	7.2	1960	22118902	52348068
10157	10	5.9	1960	0	0
...
284	101	5.4	2015	0	0
293	19	7.1	2015	0	0
528	10	7.6	2015	0	0
431	41	5.0	2015	0	0
364	12	6.1	2015	0	0
24	1545	7.3	2015	25759988	122678731
165	252	4.8	2015	55199975	0
479	31	5.2	2015	0	0
462	26	4.0	2015	0	0
559	21	6.2	2015	0	0
3	5292	7.5	2015	183999919	1902723129
95	278	5.7	2015	0	215055014
117	399	5.9	2015	27599987	96610126
173	57	4.2	2015	0	15432
92	11	5.4	2015	0	0
450	12	5.5	2015	0	0
455	15	7.6	2015	0	0
290	67	7.2	2015	0	0
515	11	5.3	2015	1195999	0
53	1001	6.4	2015	55199975	93043293
143	11	7.5	2015	0	0
5	3929	7.2	2015	124199945	490314247

91	568	5.7	2015	45999979	221131463
15	2389	7.4	2015	40479982	143299244
553	31	8.0	2015	0	0
540	37	7.7	2015	0	0
569	10	5.3	2015	0	0
144	259	6.7	2015	7359996	3458541
322	43	5.0	2015	0	0
190	33	5.6	2015	0	0

	cast_count	genre_count	companies_count
10150	5.0	3.0	1.0
10152	5.0	1.0	2.0
10162	5.0	5.0	1.0
10166	5.0	1.0	1.0
10156	5.0	2.0	2.0
10172	5.0	4.0	1.0
10147	5.0	2.0	2.0
10155	5.0	2.0	1.0
10158	5.0	2.0	1.0
10144	5.0	3.0	2.0
10165	5.0	1.0	1.0
10153	5.0	2.0	1.0
10163	5.0	2.0	1.0
10167	5.0	2.0	1.0
10164	5.0	3.0	1.0
10168	5.0	2.0	2.0
10169	5.0	1.0	2.0
10149	5.0	4.0	2.0
10141	5.0	3.0	1.0
10146	5.0	5.0	2.0
10154	5.0	1.0	1.0
10171	5.0	1.0	1.0
10160	5.0	1.0	1.0
10143	5.0	3.0	1.0
10145	5.0	2.0	2.0
10159	5.0	1.0	2.0
10142	5.0	3.0	2.0
10161	5.0	5.0	3.0
10170	5.0	2.0	2.0
10157	5.0	4.0	1.0
...
284	5.0	2.0	3.0
293	5.0	1.0	3.0
528	5.0	2.0	NaN
431	5.0	2.0	4.0
364	5.0	3.0	5.0
24	5.0	2.0	3.0
165	5.0	2.0	1.0
479	4.0	3.0	3.0
462	5.0	1.0	1.0
559	5.0	4.0	NaN
3	5.0	4.0	3.0
95	5.0	4.0	5.0

```

117      5.0      1.0      1.0
173      5.0      2.0      4.0
92       5.0      3.0      3.0
450      5.0      4.0      NaN
455      5.0      2.0      1.0
290      3.0      1.0      2.0
515      5.0      4.0      2.0
53       5.0      2.0      4.0
143      4.0      3.0      1.0
5        5.0      4.0      5.0
91       5.0      1.0      4.0
15       5.0      4.0      3.0
553      5.0      3.0      1.0
540      5.0      2.0      5.0
569      5.0      2.0      1.0
144      3.0      4.0      1.0
322      5.0      3.0      3.0
190      5.0      4.0      1.0

```

[10866 rows x 19 columns]>

Finally I use the pandas .describe function, to see the statistics of the data in each of the columns.

Out[11]:

	id	popularity	budget	revenue	runtime	vote_count
count	10866.000000	10866.000000	1.086600e+04	1.086600e+04	10866.000000	10866.000000
mean	66064.177434	0.646441	1.462570e+07	3.982332e+07	102.070863	217.3897
std	92130.136561	1.000185	3.091321e+07	1.170035e+08	31.381405	575.6190
min	5.000000	0.000065	0.000000e+00	0.000000e+00	0.000000	10.0000
25%	10596.250000	0.207583	0.000000e+00	0.000000e+00	90.000000	17.0000
50%	20669.000000	0.383856	0.000000e+00	0.000000e+00	99.000000	38.0000
75%	75610.000000	0.713817	1.500000e+07	2.400000e+07	111.000000	145.7500
max	417859.000000	32.985763	4.250000e+08	2.781506e+09	900.000000	9767.0000

Exploratory Data Analysis

Research Question 1 - Do movie stars move to bigger budget films as their career progresses?

As movie stars progress in their career, they normally take on more major parts in big films. I plan to see if there is an increase in the budget of the films over the top 20 stars careers. With the number of actors for each film restricted to 5, parts of an actors career will not show in the statistics. I will need to compare this first to the inflation adjusted budget of the film, and then also the takings of the film, as a high takings to budget ratio may show when an actor has made the film more money, which then enabled them to claim a higher wage.

First of all I will separate all of the actors out into a separate data frame, before then combining into a single list, to see who the top 20 actors are.

Actors from oldest film: actor1 Burt Lancaster

actor2 Audrey Hepburn

actor3 Audie Murphy

actor4 John Saxon

actor5 Lillian Gish

Name: 10150, dtype: object

Start and end of Actors dataframe: <bound method DataFrame.head of

	actor1	actor2	actor3 \
10150	Burt Lancaster	Audrey Hepburn	Audie Murphy
10152	Peter Cushing	Martita Hunt	Yvonne Monlaur
10162	Kenneth More	Dana Wynter	Carl MÃ¶hner
10166	Kenneth Connor	Charles Hawtrey	Sid James
10156	Randolph Scott	Nancy Gates	Skip Homeier
10172	Paul Newman	Eva Marie Saint	Lee J. Cobb
10147	Karlheinz BÃ¶hm	Moira Shearer	Anna Massey
10155	Jeffrey Hunter	Woody Strode	Constance Towers
10158	Hayley Mills	Jane Wyman	Richard Egan
10144	Jack Lemmon	Shirley MacLaine	Fred MacMurray
10165	Burt Lancaster	Jean Simmons	Arthur Kennedy
10153	Vincent Price	Mark Damon	Myrna Fahey
10163	Spencer Tracy	Fredric March	Gene Kelly
10167	George Sanders	Barbara Shelley	Martin Stephens
10164	Michael Rennie	Jill St. John	David Hedison
10168	Jerry Lewis	Alex Gerry	Bob Clayton
10169	Charles Herbert	Jo Morrow	Martin Milner
10149	Frank Sinatra	Dean Martin	Peter Lawford
10141	Anthony Perkins	Vera Miles	John Gavin
10146	Rod Taylor	Alan Young	Yvette Mimieux
10154	Albert Finney	Shirley Anne Field	Rachel Roberts
10171	Christopher Lee	Dennis Lotis	Patricia Jessel
10160	Jackie Joseph	Mel Welles	Jack Nicholson
10143	Kirk Douglas	Laurence Olivier	Jean Simmons
10145	Marilyn Monroe	Yves Montand	Tony Randall
10159	Doris Day	Rex Harrison	John Gavin
10142	Yul Brynner	Eli Wallach	Steve McQueen
10161	John Wayne	Richard Widmark	Laurence Harvey
10170	Jerry Lewis	Ed Wynn	Judith Anderson
10157	Elvis Presley	Steve Forrest	Barbara Eden
...
284	Bill Murray	Paul Shaffer	George Clooney
293	Peter Mullan	Agyness Deyn	Kevin Guthrie
528	Karissa Staples	Brant Daugherty	Brittany Underwood
431	Scott Adkins	Nick Chinlund	Caitlin Keats
364	Malin Buska	Sarah Gadon	Michael Nyqvist
24	Christian Bale	Steve Carell	Ryan Gosling
165	Adam Sandler	Taylor Lautner	Steve Buscemi
479	Stephen Dorff	Eddie Griffin	Bill Billions
462	Sam Rockwell	Jemaine Clement	Amy Ryan
559	Mike Vogel	Osy Ikhile	Daisy Betts
3	Harrison Ford	Mark Hamill	Carrie Fisher
95	Jason Lee	Justin Long	Bella Thorne
117	Tina Fey	Amy Poehler	John Leguizamo
173	Bruce Willis	Gina Carano	Kellan Lutz
92	Melanie Stone	Adam Johnson	Kevin Sorbo
450	Lacey Chabert	Mariah Carey	Brennan Elliott
455	Taylor Swift	Ellen DeGeneres	Cara Delevingne
290	Michael Moore	Krista Kiuru	Tim Walker
515	Edmund Kingsley	David Wayman	Nigel Barber
53	Jennifer Lawrence	Bradley Cooper	Robert De Niro
143	Manuel Camacho	Jean Reno	Tobias Moretti

5	Leonardo DiCaprio	Tom Hardy	Will Poulter
91	Will Ferrell	Mark Wahlberg	Linda Cardellini
15	Samuel L. Jackson	Kurt Russell	Jennifer Jason Leigh
553	Peter Capaldi	Alex Kingston	Greg Davies
540	Maeve Dermody	Aidan Turner	Charles Dance
569	Rafe Spall	Cara Theobold	Tom Ward
144	Jennifer Jason Leigh	David Thewlis	Tom Noonan
322	Troian Bellisario	Bailey Noble	Kate Burton
190	Garry Chalk	Kathleen Barr	Willa Townsend

	actor4	actor5	
10150	John Saxon	Lillian Gish	
10152	Freda Jackson	David Peel	
10162	Laurence Naismith	Geoffrey Keen	
10166	Kenneth Williams	Leslie Phillips	
10156	Dyke Johnson	Foster Hood	
10172	Ralph Richardson	Peter Lawford	
10147	Maxine Audley	Brenda Bruce	
10155	Billie Burke	Juano Hernández	
10158	Karl Malden	Nancy Olson	
10144	Ray Walston	Jack Kruschen	
10165	Dean Jagger	Shirley Jones	
10153	Harry Ellerbe	David Andar	
10163	Dick York	Donna Anderson	
10167	Michael Gwynn	Laurence Naismith	
10164	Claude Rains	Fernando Lamas	
10168	Sonnie Sands	Eddie Shaeffer	
10169	Rosemary DeCamp	Margaret Hamilton	
10149	Angie Dickinson	Richard Conte	
10141	Janet Leigh	Martin Balsam	
10146	Sebastian Cabot	Tom Helmore	
10154	Hylde Baker	Norman Rossington	
10171	Tom Naylor	Venetia Stevenson	
10160	Leola Wendorff	Myrtle Vail	
10143	Charles Laughton	Peter Ustinov	
10145	Frankie Vaughan	Wilfrid Hyde-White	
10159	Myrna Loy	Roddy McDowall	
10142	Charles Bronson	Robert Vaughn	
10161	Frankie Avalon	Patrick Wayne	
10170	Henry Silva	Robert Hutton	
10157	Dolores del Río	Rodolfo Acosta	
...	
284	Miley Cyrus	Michael Cera	
293	Hugh Ross	Douglas Rankine	
528	Ion Overman	Doris Roberts	
431	Jake La Botz	Tony Perez	
364	Lucas Bryant	Laura Birn	
24	Brad Pitt	Melissa Leo	
165	Terry Crews	Danny Trejo	
479	Jonathan Billions	None	
462	Danny McBride	Will Forte	
559	Georgina Haig	Charles Dance	
3	Adam Driver	Daisy Ridley	
95	Matthew Gray Gubler	Jesse McCartney	
117	Ike Barinholtz	Maya Rudolph	
173	D. B. Sweeney	Dan Bilzerian	
92	Nicola Posener	Jake Stormoen	
450	Kathy Najimy	Kevin Chamberlin	
455	Lena Dunham	Selena Gomez	
290	None	None	
515	Lisa Greenwood	Gil Koirin	

53	Dascha Polanco	Edgar RamÃ¡rez
143	Eva Kuen	None
5	Domhnall Gleeson	Paul Anderson
91	Thomas Haden Church	Hannibal Buress
15	Walton Goggins	DemiÃ¡n Bichir
553	Matt Lucas	Phillip Rhys
540	Douglas Booth	Sam Neill
569	Sophie Stanton	Zoe Boyle
144	None	None
322	Caitlin Carmichael	Melissa Tracy
190	Melissa Sturm	Trevor Devall

[10866 rows x 5 columns]>

Single list length: 54330
Start and end of single list: <bound method Series.head of 0 Burt Lancaster

1	Peter Cushing
2	Kenneth More
3	Kenneth Connor
4	Randolph Scott
5	Paul Newman
6	Karlheinz Böhm
7	Jeffrey Hunter
8	Hayley Mills
9	Jack Lemmon
10	Burt Lancaster
11	Vincent Price
12	Spencer Tracy
13	George Sanders
14	Michael Rennie
15	Jerry Lewis
16	Charles Herbert
17	Frank Sinatra
18	Anthony Perkins
19	Rod Taylor
20	Albert Finney
21	Christopher Lee
22	Jackie Joseph
23	Kirk Douglas
24	Marilyn Monroe
25	Doris Day
26	Yul Brynner
27	John Wayne
28	Jerry Lewis
29	Elvis Presley
	...
54300	Michael Cera
54301	Douglas Rankine
54302	Doris Roberts
54303	Tony Perez
54304	Laura Birn
54305	Melissa Leo
54306	Danny Trejo
54307	None
54308	Will Forte
54309	Charles Dance
54310	Daisy Ridley
54311	Jesse McCartney
54312	Maya Rudolph
54313	Dan Bilzerian
54314	Jake Stormoen
54315	Kevin Chamberlin
54316	Selena Gomez
54317	None
54318	Gil Kolirin
54319	Edgar Ramírez
54320	None
54321	Paul Anderson
54322	Hannibal Buress
54323	Demián Bichir
54324	Phillip Rhys
54325	Sam Neill
54326	Zoe Boyle

```
54327      None
54328  Melissa Tracy
54329    Trevor Devall
Length: 54330, dtype: object>
```

Then I will create a list with the total count of each actor appearing in the dataset.

Actors list length: 19026

Start and end of Top Actors: <bound method DataFrame.head of

Actor Count

	Actor	Count
0	Robert De Niro	72
1	Samuel L. Jackson	71
2	Bruce Willis	62
3	Nicolas Cage	61
4	Michael Caine	53
5	Robin Williams	51
6	John Cusack	50
7	Morgan Freeman	49
8	John Goodman	49
9	Liam Neeson	48
10	Susan Sarandon	48
11	Julianne Moore	47
12	Alec Baldwin	47
13	Christopher Walken	46
14	Tom Hanks	46
15	Johnny Depp	46
16	Gene Hackman	46
17	Sylvester Stallone	45
18	Willem Dafoe	45
19	Dennis Quaid	45
20	Donald Sutherland	44
21	Clint Eastwood	44
22	Meryl Streep	44
23	Ed Harris	44
24	Ewan McGregor	43
25	Antonio Banderas	43
26	Woody Harrelson	43
27	Robert Downey Jr.	43
28	Keanu Reeves	43
29	Danny Glover	42
...
18996	Wyclef Jean	1
18997	Nikki Leigh	1
18998	Steffani Brass	1
18999	William Takaku	1
19000	Andrew Fletcher	1
19001	Jack Klugman	1
19002	Ana ObregÃ³n	1
19003	Abigail Clayton	1
19004	Julia Winter	1
19005	Jonathan R. Scott	1
19006	Kyle Bornheimer	1
19007	Christopher St. John	1
19008	Reynaldo Gallegos	1
19009	Anna Deavere Smith	1
19010	Naomi Scott	1
19011	Don Omar	1
19012	Clare Grogan	1
19013	Julie Fain Lawrence	1
19014	Gala Besson	1
19015	Roz Hammond	1
19016	Pascal Rollin	1
19017	Pat Ryan	1
19018	Evan Helmuth	1
19019	Bruno Raffaelli	1
19020	Gina Ravera	1
19021	Melody C. Roscher	1

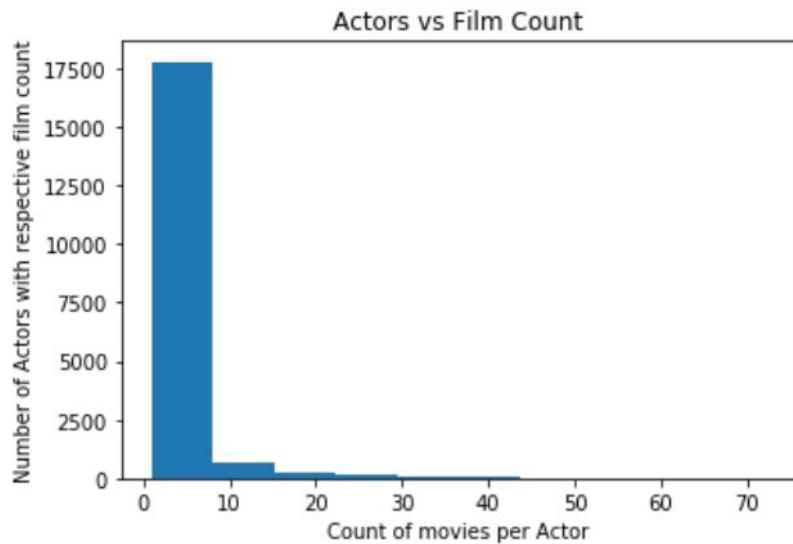
```
19022      William Ainscough      1
19023          India Hair       1
19024        Hudson Meek       1
19025    Takumi Yamazaki       1
```

[19026 rows x 2 columns]>

Then I will plot this list as a histogram to see if there is a good spread of data.

Out[15]:

```
Text(0.5,1,'Actors vs Film Count')
```

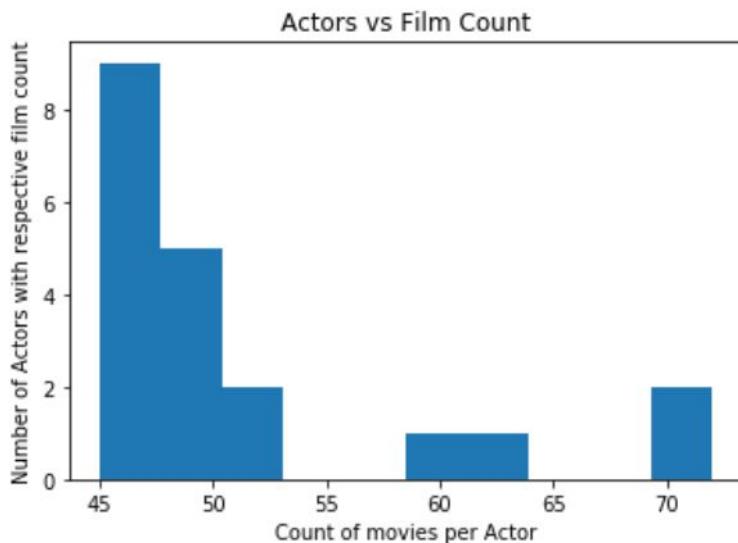


As the majority of actors have appeared in less than 10 films (as a top 5 actor), I will create a shorter list containing just the top 20 actors, so that I have the most data for each actor to compare against one another. I will also plot another histogram for these top 20 actors.

	Actor	Count
0	Robert De Niro	72
1	Samuel L. Jackson	71
2	Bruce Willis	62
3	Nicolas Cage	61
4	Michael Caine	53
5	Robin Williams	51
6	John Cusack	50
7	Morgan Freeman	49
8	John Goodman	49
9	Liam Neeson	48
10	Susan Sarandon	48
11	Julianne Moore	47
12	Alec Baldwin	47
13	Christopher Walken	46
14	Tom Hanks	46
15	Johnny Depp	46
16	Gene Hackman	46
17	Sylvester Stallone	45
18	Willem Dafoe	45
19	Dennis Quaid	45

Out[16]:

Text(0.5,1,'Actors vs Film Count')



Now that I have a useable Actor dataset, I can analyse the main TMDb dataset, and graph the out put.

I have defined two routines for creating the graphs, one to filter out the small and zero values I was seeing for the inflation adjusted budget, and the second to extract the data to graph for just one Actor.

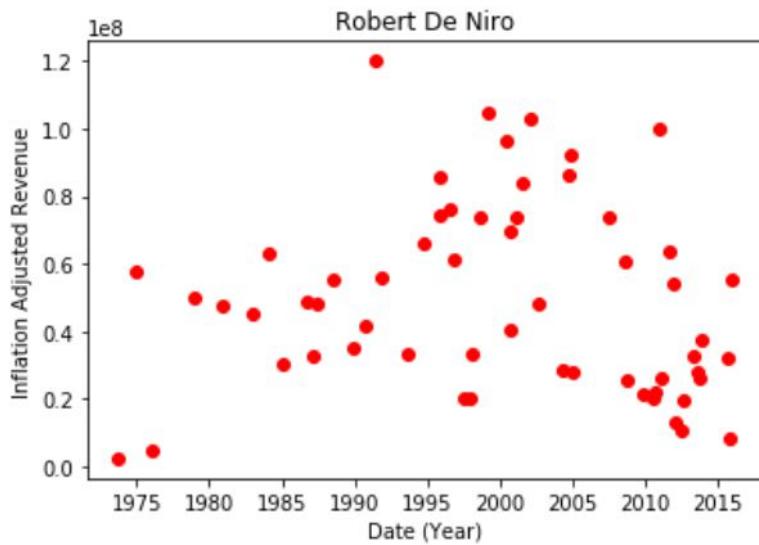
I then created graphs for the Top 20 Actors.

Graph for Actor: Robert De Niro

Top Actor Number: 1

Was a top 5 Actor in: 72 films

Number of plotted films: 58

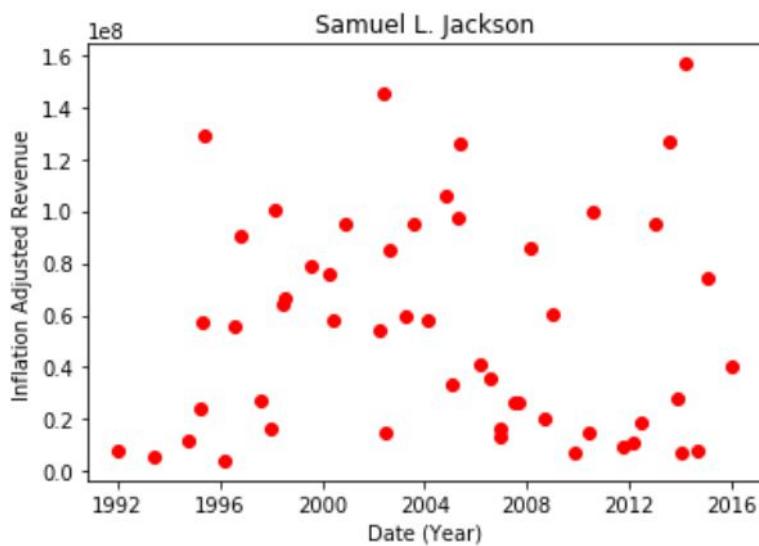


Graph for Actor: Samuel L. Jackson

Top Actor Number: 2

Was a top 5 Actor in: 71 films

Number of plotted films: 52

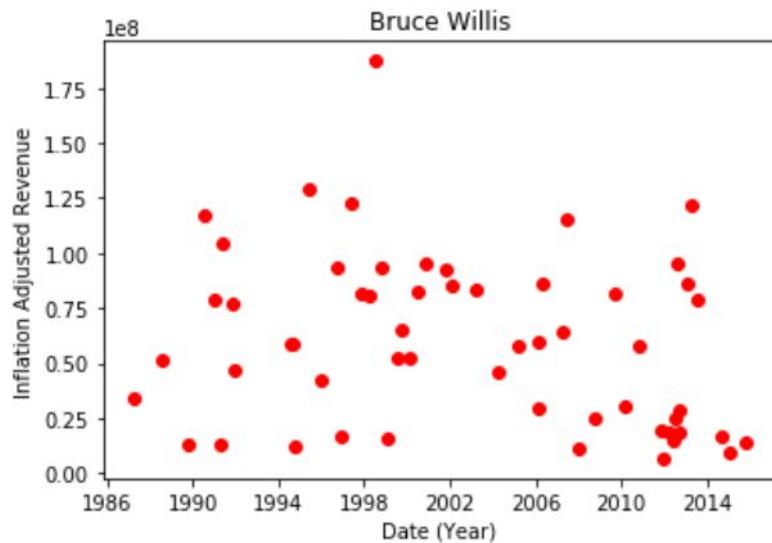


Graph for Actor: Bruce Willis

Top Actor Number: 3

Was a top 5 Actor in: 62 films

Number of plotted films: 56

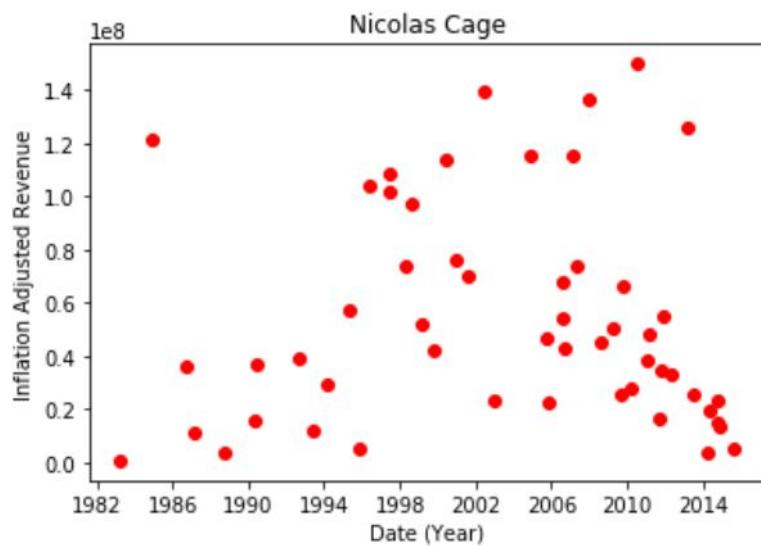


Graph for Actor: Nicolas Cage

Top Actor Number: 4

Was a top 5 Actor in: 61 films

Number of plotted films: 53

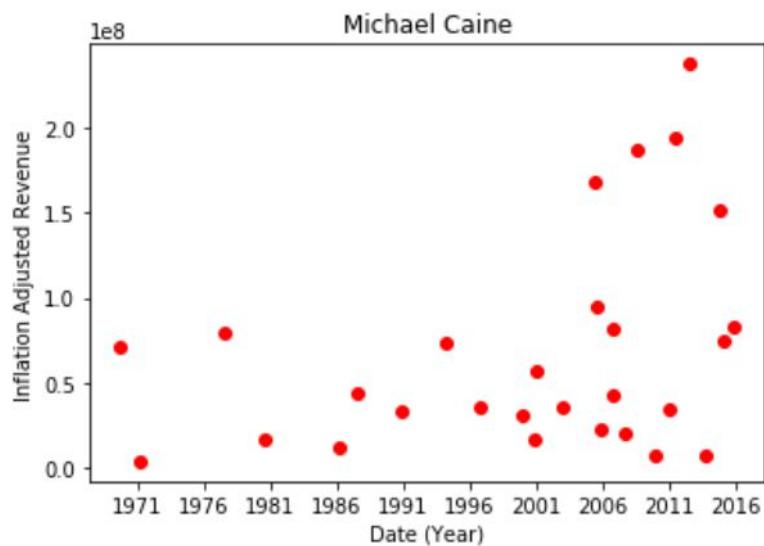


Graph for Actor: Michael Caine

Top Actor Number: 5

Was a top 5 Actor in: 53 films

Number of plotted films: 28

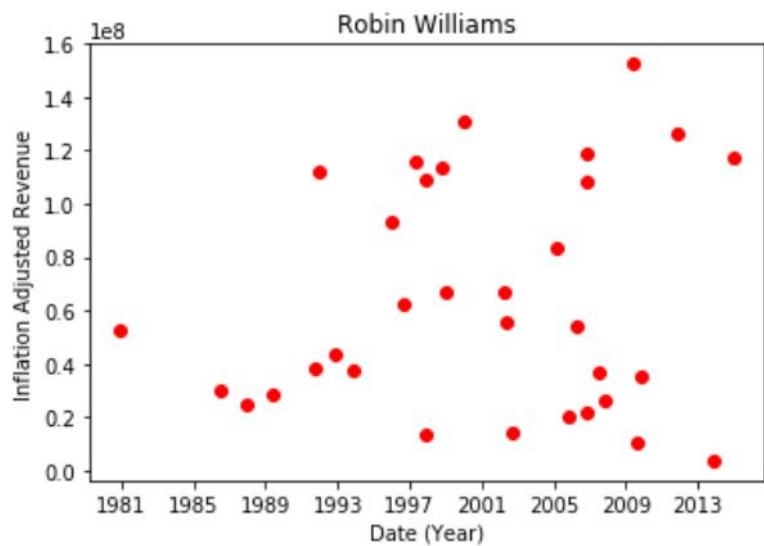


Graph for Actor: Robin Williams

Top Actor Number: 6

Was a top 5 Actor in: 51 films

Number of plotted films: 33

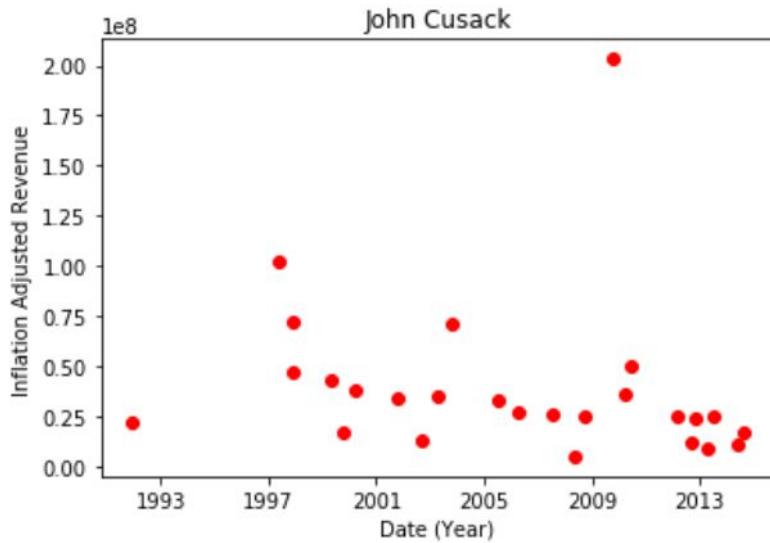


Graph for Actor: John Cusack

Top Actor Number: 7

Was a top 5 Actor in: 50 films

Number of plotted films: 26

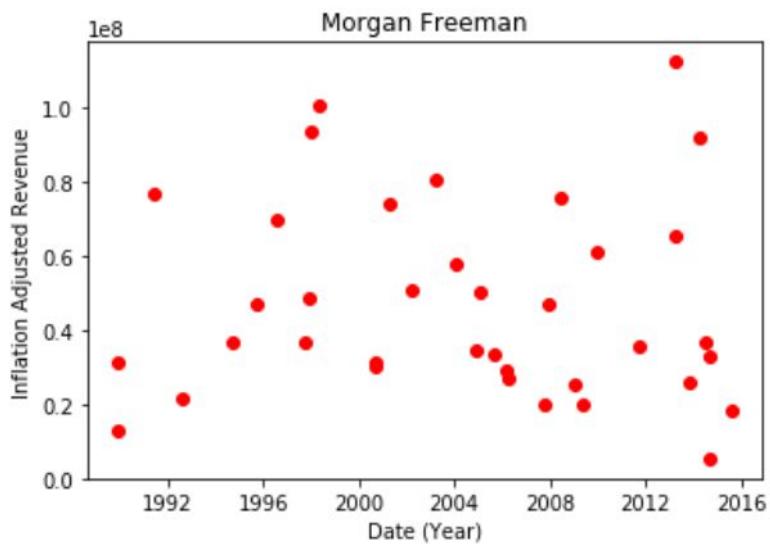


Graph for Actor: Morgan Freeman

Top Actor Number: 8

Was a top 5 Actor in: 49 films

Number of plotted films: 37

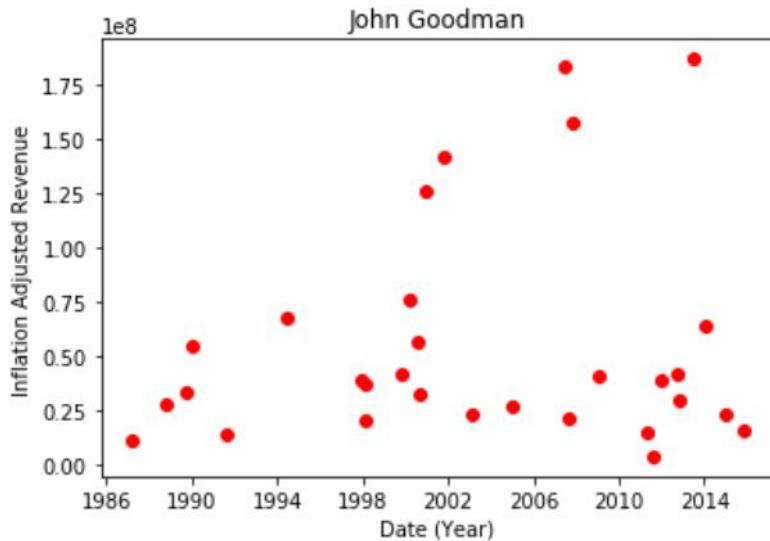


Graph for Actor: John Goodman

Top Actor Number: 9

Was a top 5 Actor in: 49 films

Number of plotted films: 30

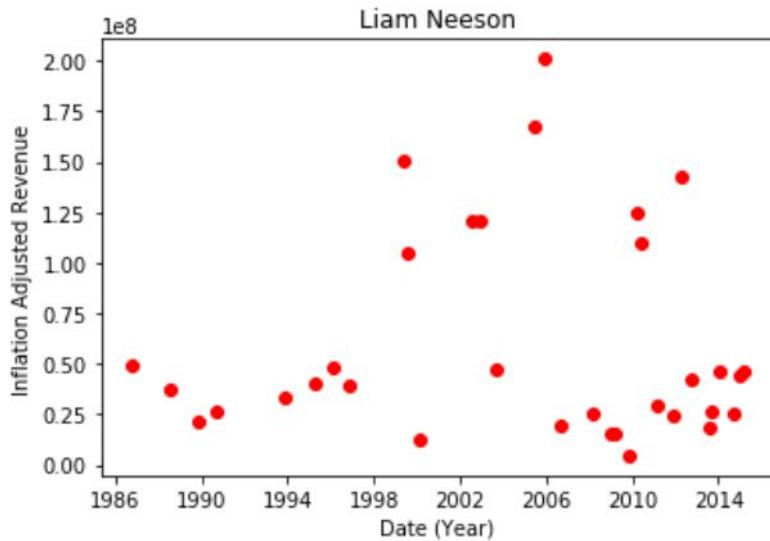


Graph for Actor: Liam Neeson

Top Actor Number: 10

Was a top 5 Actor in: 48 films

Number of plotted films: 33



Unfortunately I could not see a corellation between the budget of a film increasing as a movie stars career progresses. If the data had been available, this may have shown for actors at the start of their career, before they were mentioned in the top 5 actors.

Research Question 2 - Is the TMDb popularity linked to the revenue taken by the film?

I then wanted to see if the TMDb popularity was linked to the revenue taken by the film. I wanted to graph the inflation adjusted revenue against the average vote, however if this did not show a corellation, I could also look at the vote count and popularity data series. First I reminded myself of the first data point in my dataset.

```

id                               6643
popularity                      0.421043
budget                            0
revenue                           0
original_title                  The Unforgiven
cast        Burt Lancaster|Audrey Hepburn|Audie Murphy|Joh...
director                         John Huston
runtime                           125
genres                            Action|Drama|Western
production_companies            James Productions
release_date                     1960-01-01 00:00:00
vote_count                        17
vote_average                      4.9
release_year                      1960
budget_adj                         0
revenue_adj                         0
cast_count                         5
genre_count                        3
companies_count                   1
Name: 10150, dtype: object

```

I then created a new dataset where the value of the adjusted budget was at least 100, in order to remove any erroneous data points, before returning the size of the new dataset.

4814

I then again called up the first data point to ensure that the first data point with 0 for adjusted budget had been removed.

```

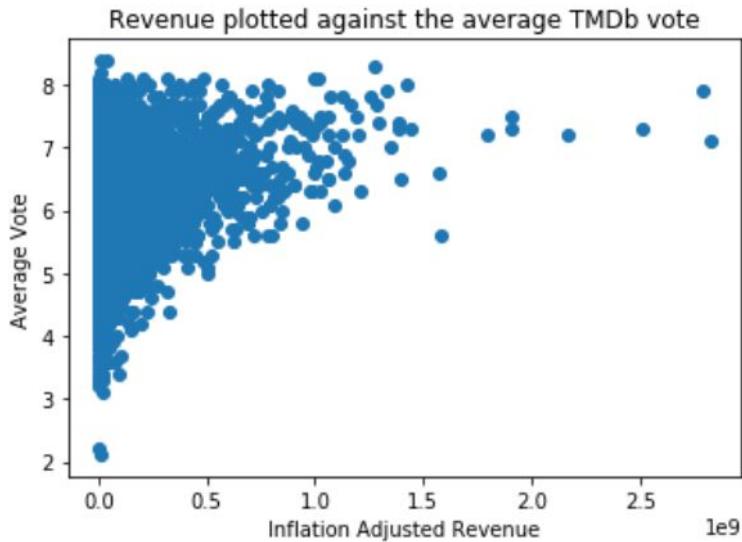
id                               284
popularity                      0.947307
budget                          3000000
revenue                          25000000
original_title                  The Apartment
cast        Jack Lemmon|Shirley MacLaine|Fred MacMurray|Ra...
director                         Billy Wilder
runtime                           125
genres                            Comedy|Drama|Romance
production_companies            United Artists|The Mirisch Company
release_date                     1960-06-15 00:00:00
vote_count                        235
vote_average                      7.9
release_year                      1960
budget_adj                        22118902
revenue_adj                        184324186
cast_count                         5
genre_count                        3
companies_count                   2
Name: 10144, dtype: object

```

I then separated out the data series I wanted to graph, along with the backup data series', before plotting the inflation adjusted revenue against the average vote.

Out[31]:

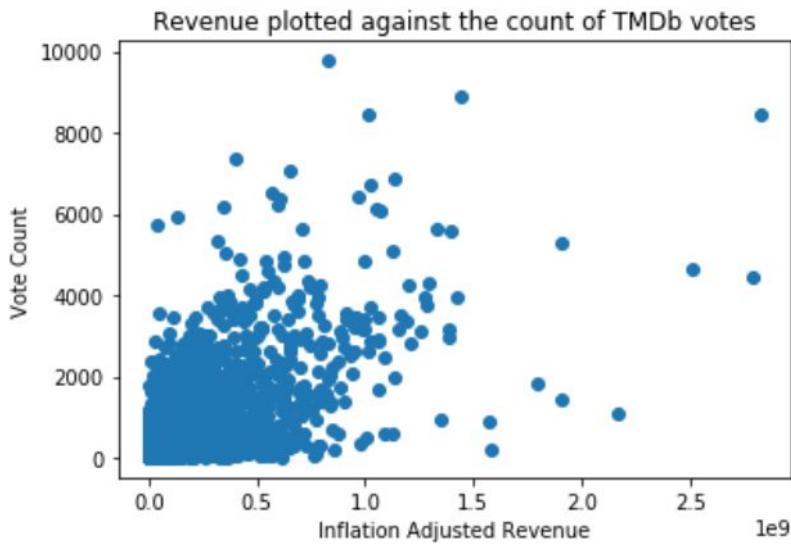
```
Text(0.5,1,'Revenue plotted against the average TMDb vote')
```



Whilst there was a positive skew showing that films with a larger budget tend to have a larger vote, I wanted to see if I had a better result off plotting the other data series, so I also plotted these.

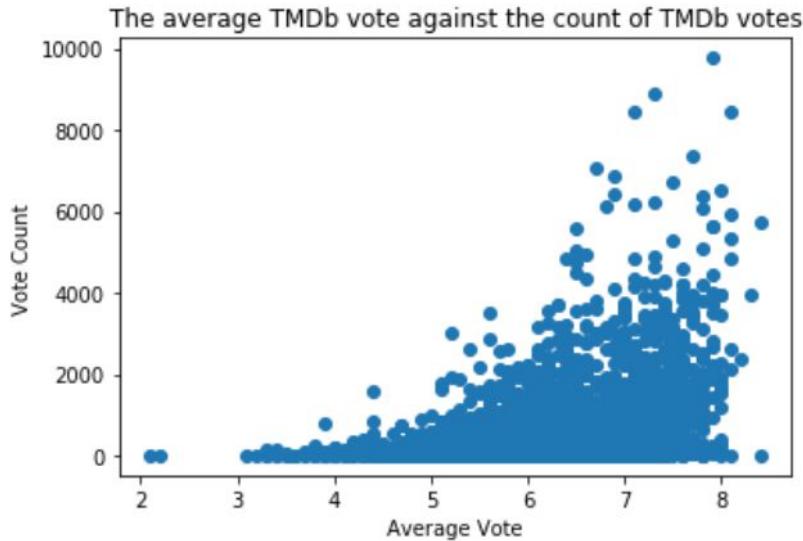
Out[32]:

```
Text(0.5,1,'Revenue plotted against the count of TMDb votes')
```



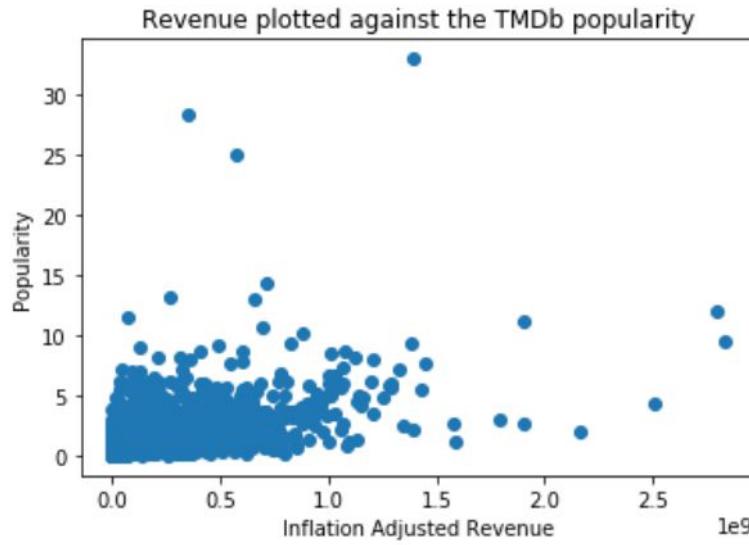
Out[33]:

Text(0.5,1,'The average TMDb vote against the count of TMDb votes')



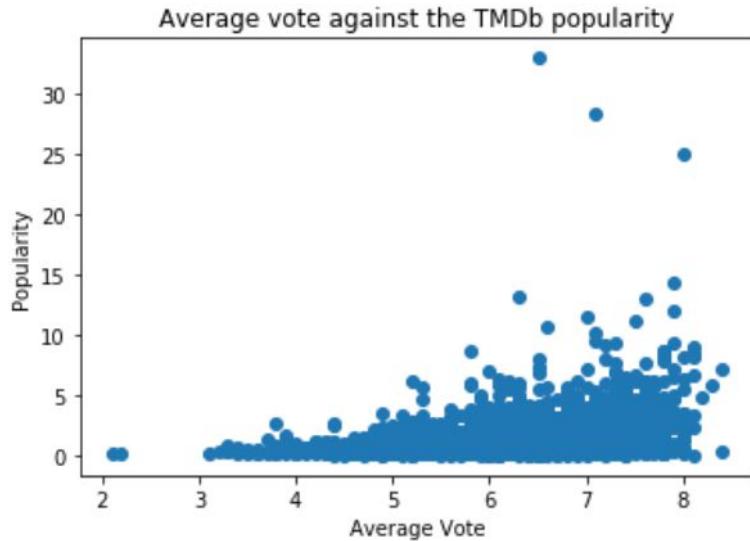
Out[34]:

Text(0.5,1,'Revenue plotted against the TMDb popularity')



Out[35]:

```
Text(0.5,1,'Average vote against the TMDb popularity')
```



Conclusions

Whilst I did not show any correlation between the top actors against the film budget, there was a correlation between the TMDb vote and the revenue taken by the film.

It was still worthwhile comparing the top actors against the budget, to see that this was not the case, but also the main benefit was to see how many films the top actors had actually starred in.

The positive correlation between the vote and revenue could benefit the production companies, as they may be able to select a test sample of reviewers to watch the film, in order to give a pre-release indication of the expected revenue of the film.