

Scraping PC games from metacritic.com

...

A learning exercise masquerading as an investigation

Syntax and conventions

- *Italicized* column/field/feature names
- yellow underlined text are relevant hyperlinks
- shorthand:
 - “&”, “ea.” abide by their conventional usages
 - approximately: “~”
 - with, without, & because: “w/” , “w/o” , “b/c”
- Box plots
 - middle quartiles are shaded and have whiskers of magnitude 1.5x IQR (interquartile range) or terminate at most extreme value, whichever comes first
 - Year axis subtensions are quarters unless otherwise labelled

About the metacritic domain & site



- An aptly named aggregator of media criticism
 - cites published reviews from acknowledged critics & accepts more from site users
 - critical response is a confidential heuristic weighting critics' stature
 - user response is a statistical average



- features film, TV, music, & video games
 - organized by platform, from the current console offerings to “legacy” predecessors
 - herein focuses on just PC (computer) games

Software tools used

- Scrapy
 - python based web crawler for visiting & extracting data from websites
 - visits & reads the ~26k PC game web pages at ~275/min
- pandas
 - python based data structures supporting spreadsheet-like operations
 - check Scrapy generated data w/o risking human error & boredom
- Tableau
 - interactive data visualization
 - chosen over python offerings like seaborn & matplotlib for the GUI

**Under the hood:
How each tool is put to use**

Under the hood - Scrappy - visiting game pages

Start at alphanumeric header, check each game, exhaust all pages of games, go to next alphanumeric category & repeat

The screenshot shows a web page titled "GAME RELEASES BY TITLE" with a navigation bar for various platforms: PS4, Xbox One, Switch, PC (highlighted), Wii U, 3DS, PS Vita, and IOS. Below the navigation bar is an alphabetical header from A to Z. The main content area displays a list of games. Annotations include a yellow arrow pointing to the alphabetical header, a purple arrow pointing to the first game entry, a vertical double-headed arrow indicating a scroll action, and a blue arrow pointing to the pagination controls at the bottom. The pagination shows "Page: 1 2 3" with "2" selected, and "prev" and "next" buttons.

GAME RELEASES BY TITLE																										
PS4 Xbox One Switch PC Wii U 3DS PS Vita IOS																										
#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
tbd	! That Bastard Is Trying To Steal Our Gold ! (PC)																								User: tbd	May 11, 2014
tbd	Insane (PC)																								User: tbd	Mar 30, 2016
tbd	1982 (PC)																								User: tbd	Jun 28, 2017
tbd	1992 Space Machine (PC)																								User: 5.5	Mar 28, 2016

Page: 1 2 3 prev next

Under the hood - Scrappy - reading game pages for features

title

COMPANY OF HEROES: OPPOSING FRONTS PC



THQ | Release Date: Sep 24, 2007

publisher

Summary

Critic Reviews

User Reviews

Details & Credits

Trailers & Videos

releaseDate



87

Metascore

Generally favorable reviews
based on 36 Critics

What's this?

8.6

User Score

Generally favorable reviews
based on 236 Ratings

Your Score

0

developer

Summary: Company of Heroes: Opposing Fronts, the next chapter in the real-time strategy franchise from Relic Entertainment is powered by an upgraded version of Relic's proprietary Essence Engine with full Microsoft DirectX 10 support. Opposing Fronts features two unique armies with full length... [Expand](#)

Developer: Relic

Genre(s): Strategy, Real-Time, Military, General, Military

of players: 6 Online

Cheats: On GameFAQs

Rating: M


[More Details and Credits »](#)

genres

Buy On [amazon.com](#)

esrb

CRITIC REVIEWS

Positive:  36
Mixed:  0
Negative:  0

USER REVIEWS

[Write a Review](#)

Positive:  36
Mixed:  4
Negative:  7

Under the hood - Scrappy - reading game pages for features

numCriticReviews

meanCriticReview

fractionCriticPositive

fractionCriticMixed


fractionCriticNegative

COMPANY OF HEROES: OPPOSING FRONTS PC

THQ | Release Date: Sep 24, 2007

f t p g

Summary Critic Reviews User Reviews Details & Credits Trailers & Videos



87


Metascore
Generally favorable reviews based on **36 Critics**

What's this?

Summary: Company of Heroes: Opposing Fronts, the next chapter in the real-time strategy franchise from Relic Entertainment is powered by an upgraded version of Relic's proprietary Essence Engine with full Microsoft DirectX 10 support. Opposing Fronts features two unique armies with full length... [Expand](#)




[Buy On amazon.com](#)

User Score
8.6
Generally favorable reviews based on **236 Ratings**




Your Score  0

Developer: Relic
Genre(s): Strategy, Real-Time, Military, General, Military
of players: 6 Online
Cheats: On GameFAQs
Rating: M
[More Details and Credits](#)

CRITIC REVIEWS

Positive:  36
Mixed:  0
Negative:  0

USER REVIEWS [Write a Review](#)

Positive:  36
Mixed:  4
Negative:  7

Under the hood - Scrappy - reading game pages for features

numUserReviews

meanUserReview

fractionUserPositive

fractionUserMixed


fractionUserNegative

COMPANY OF HEROES: OPPOSING FRONTS PC

THQ | Release Date: Sep 24, 2007


f t p g

Summary Critic Reviews User Reviews Details & Credits Trailers & Videos



Metascore
87
Generally favorable reviews based on 36 Critics
What's this?

User Score
8.6
Generally favorable reviews based on 236 Ratings




Your Score  0

Summary: Company of Heroes: Opposing Fronts, the next chapter in the real-time strategy franchise from Relic Entertainment is powered by an upgraded version of Relic's proprietary Essence Engine with full Microsoft DirectX 10 support. Opposing Fronts features two unique armies with full length... [Expand](#)




Developer: Relic
Genre(s): Strategy, Real-Time, Military, General, Military
of players: 6 Online
Cheats: On GameFAQs
Rating: M
[More Details and Credits](#)

[Buy On amazon.com](#)

CRITIC REVIEWS

Positive:  36
Mixed:  0
Negative:  0

USER REVIEWS [Write a Review](#)

Positive:  36
Mixed:  4
Negative:  7

Under the hood - pandas - data validation

- build a table (pd.dataframe) with features in columns & games in rows
 - add new helper features for use during manual inspection
 - *url*: the address of the web page; serves as a unique key
 - *scrapyStatus*: the status code Scrapy returns for the web page
 - *inspect*: any information to assist w/ manual data edits goes here
- test ea. row (game) for sensible values, noting issues in *inspect*
 - ensure values that ought to be integer, float, datetime, etc. are the correct type
 - check values of a homogenous type for bounds, e.g. $0 \leq x \leq 1$ or $\sum x_i = 1$
 - reduces manual inspection effort in the validation pipeline
- export a CSV for a human to verify and Tableau to import

Under the hood - pandas - data validation

- Scrapy generated 26647 entries; 522 contained 589 things to *inspect*
 - 435 of 522: duplicate *title* - entries w/ a distinct *url*, but otherwise identical, are removed
 - 122 of 522: missing *developer* & *publisher* - false positives where no info was provided
 - 17 of 522: special character in *title* - occurrences of ‘;’ sometimes as part of parsing ‘<’ or ‘>’
 - 15 of 522: non datetime *releaseDate* - unannounced release dates revised to null
 - no issues with quantitative fields
- the final dataset: 26278 entries w/ 17 dimensions:
 - titles from fringe to popular, hit to dud, spanning 1988 to Sept. 2017 and beyond
 - qualitative game descriptors
 - quantitative critic data
 - quantitative user data

Under the hood - Tableau - interrogating the data

Answer questions & gain insight by plotting features & their derived statistics:

- Is game quality improving?
- Is there a "summer slump"?
- Are critics biased?
- Do users form angry mobs?
- Are some developers predictable?
- Are they independent of publishers?
- What is the genre distribution?

It's stable.

Not objectively.

Blame the system.

A fair characterization.

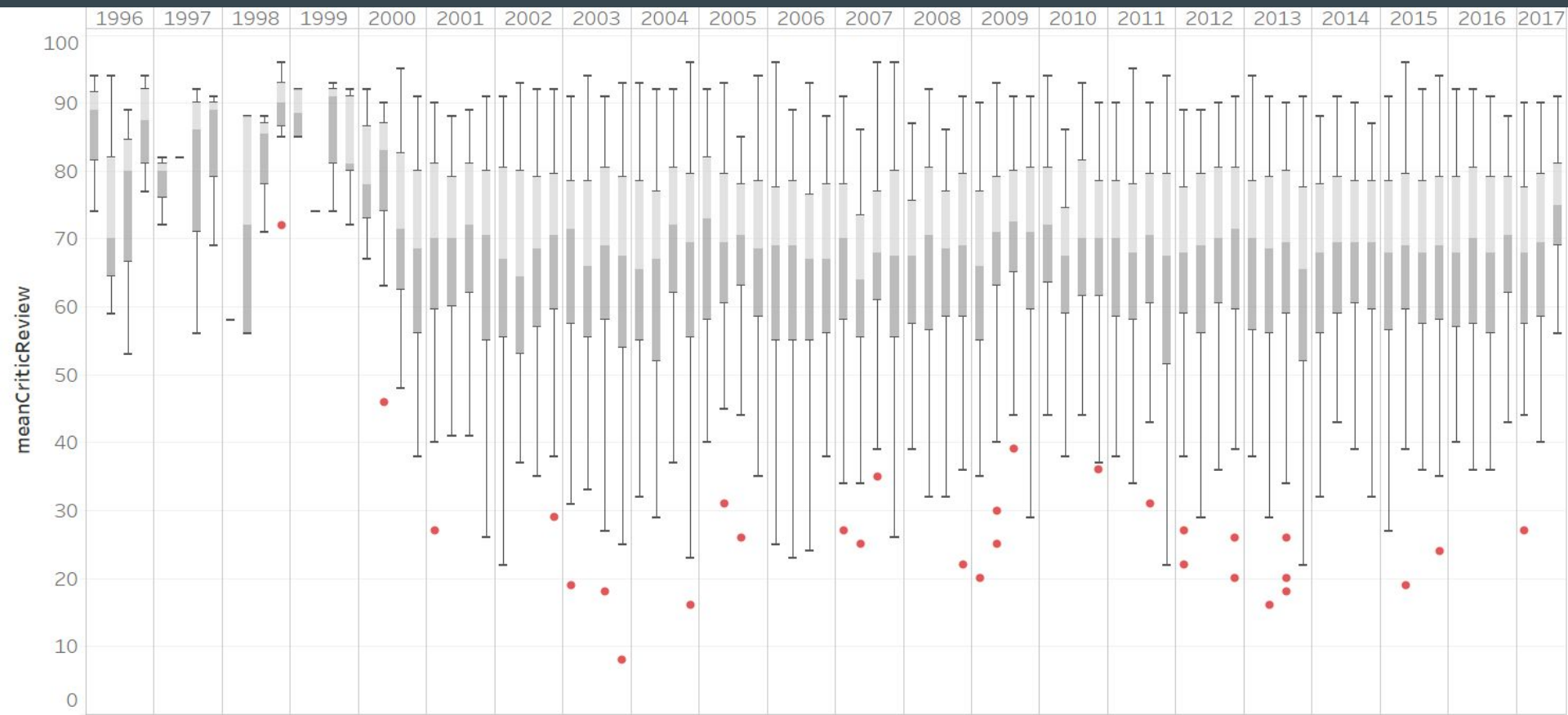
In both good & bad ways.

Just under half are.

It caters to short attention spans.

What the data have to say:

Is game quality improving?

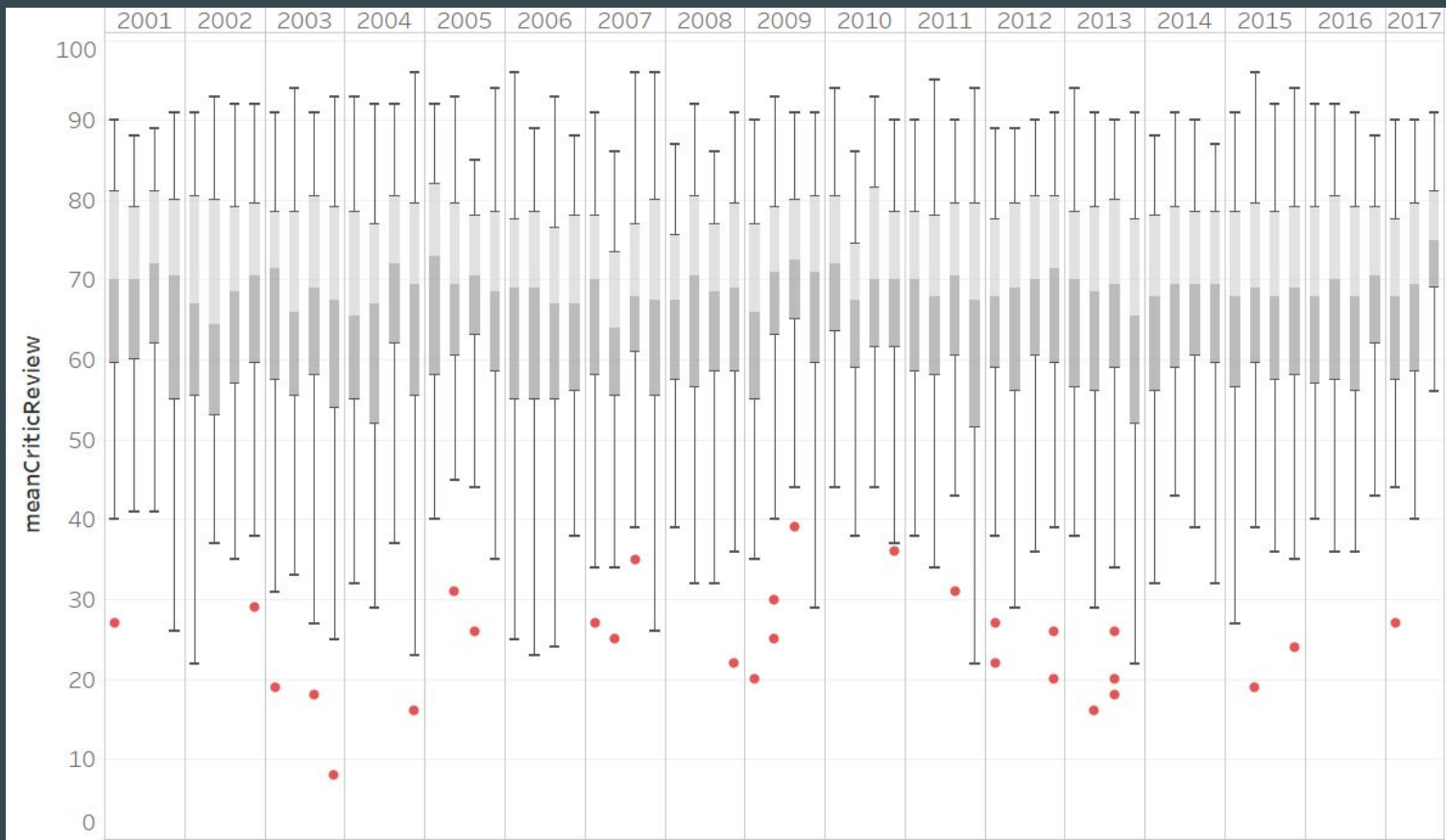


Is game quality improving?

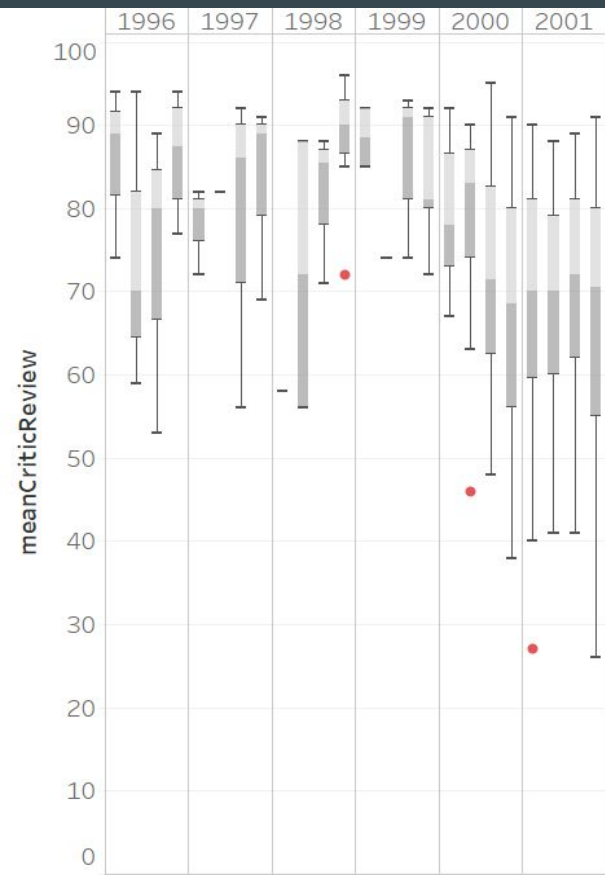
Since 2001, critics would say “no” overall.

This past quarter (Q3 2017) was unusually good, however.

What about the decline of 2000?



Did game quality decline after 2000?

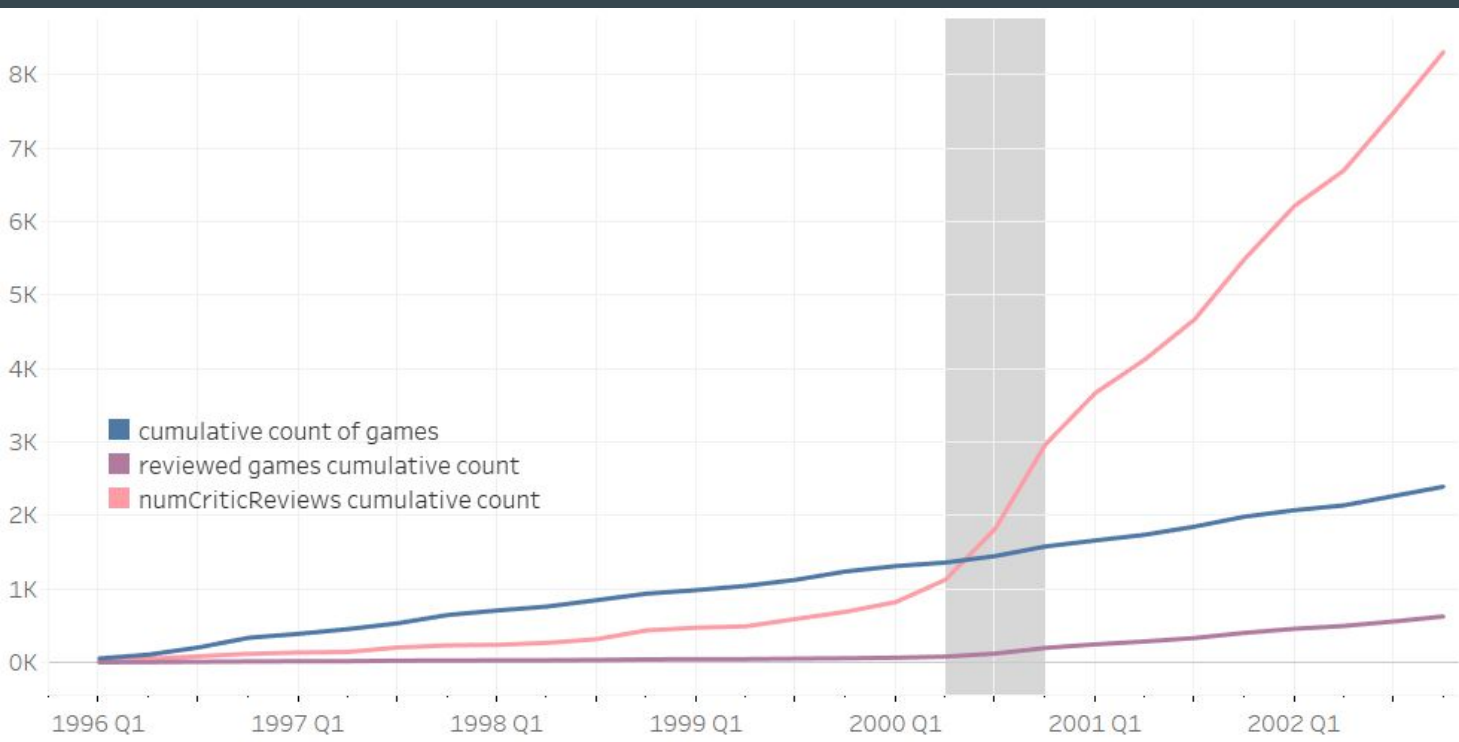


Consider that the underlying data is the set for which *meanCriticReview* exists b/c *numCriticReviews* > 3.

Are there other qualities of the data that change at the same time, possibly contributing to this trend?

Critic behavior also changed during 2000.

meanCriticReview came to represent a greater diversity of opinions.

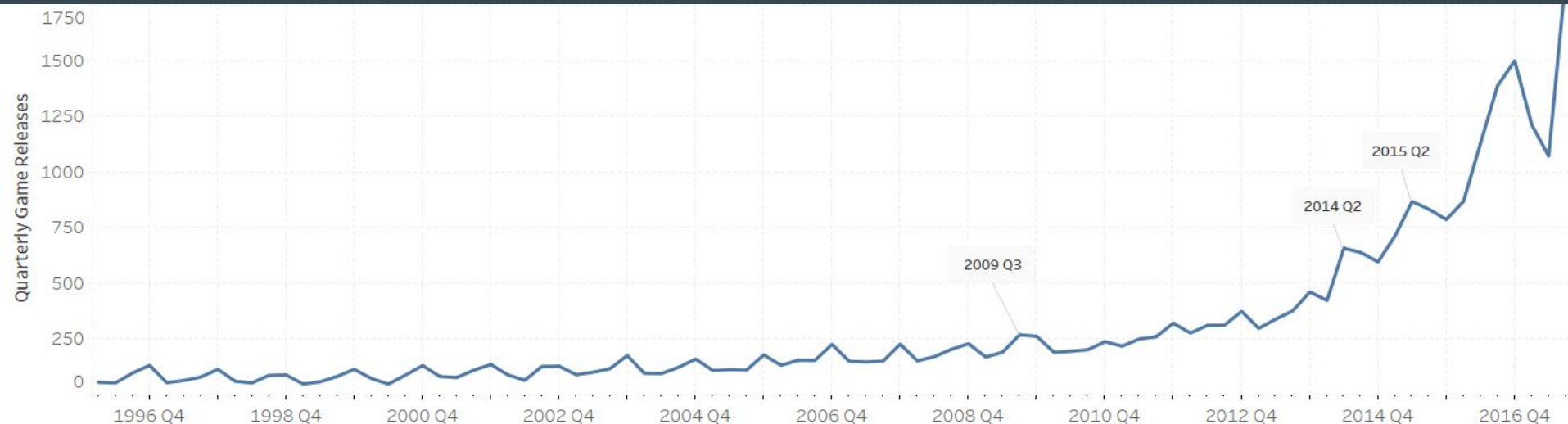


Perhaps some critics retroactively reviewed titles they enjoyed or incoming critics were harder to please.

Whatever the cause, it casts doubt that games became objectively worse after 2000.

Is there a summer slump?

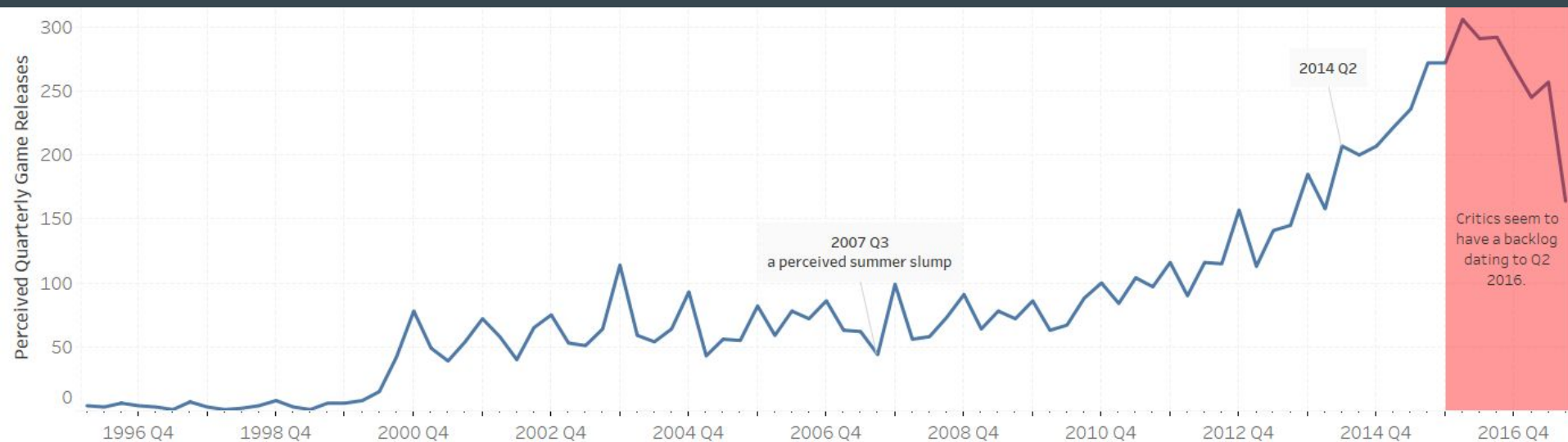
- Neither in quality (from earlier), nor the availability of games



- The true slumps are in the wake of the Q4 holiday shopping season
 - sans a few exceptions, a year's releases peaks at Q4, w/ a low point the following Q1 or Q2
 - It's a popular sentiment that summer slumps exist. Is something being overlooked?

The summer slump is a matter of perception.

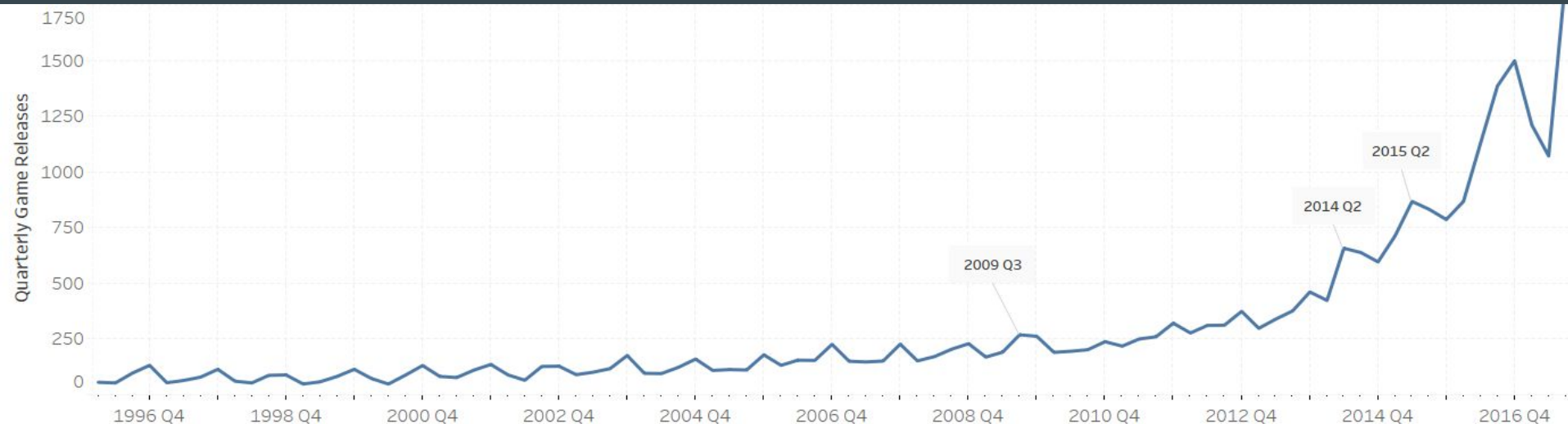
- Controlling for the perceived availability of games doesn't change much.
 - perhaps only games w/ *numCriticReviews* > 0 may be known to people:



- It's darkest before dawn: the slump's over by summer, but the sentiment persists.

PC games have also been surging in popularity.

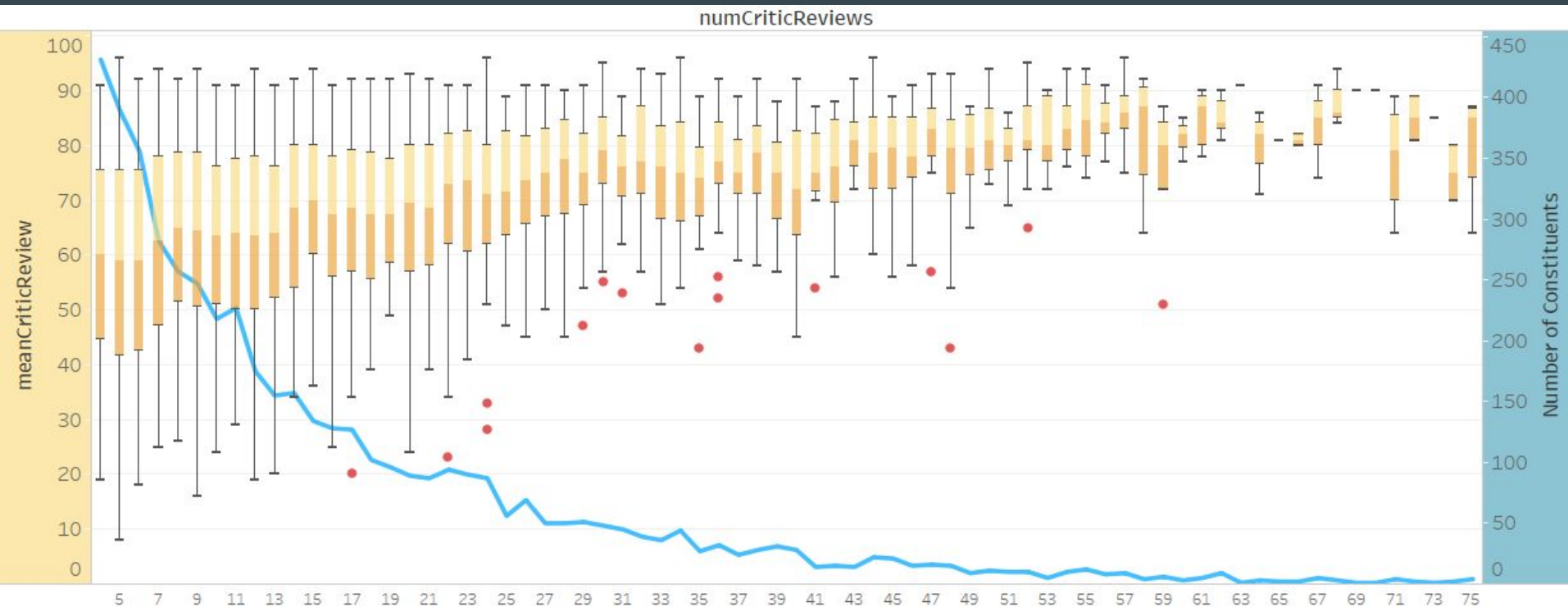
- “Nerd” culture, digital youth, & indie games’ work, it seems.



- Surprisingly, growth endured the 2008 recession.

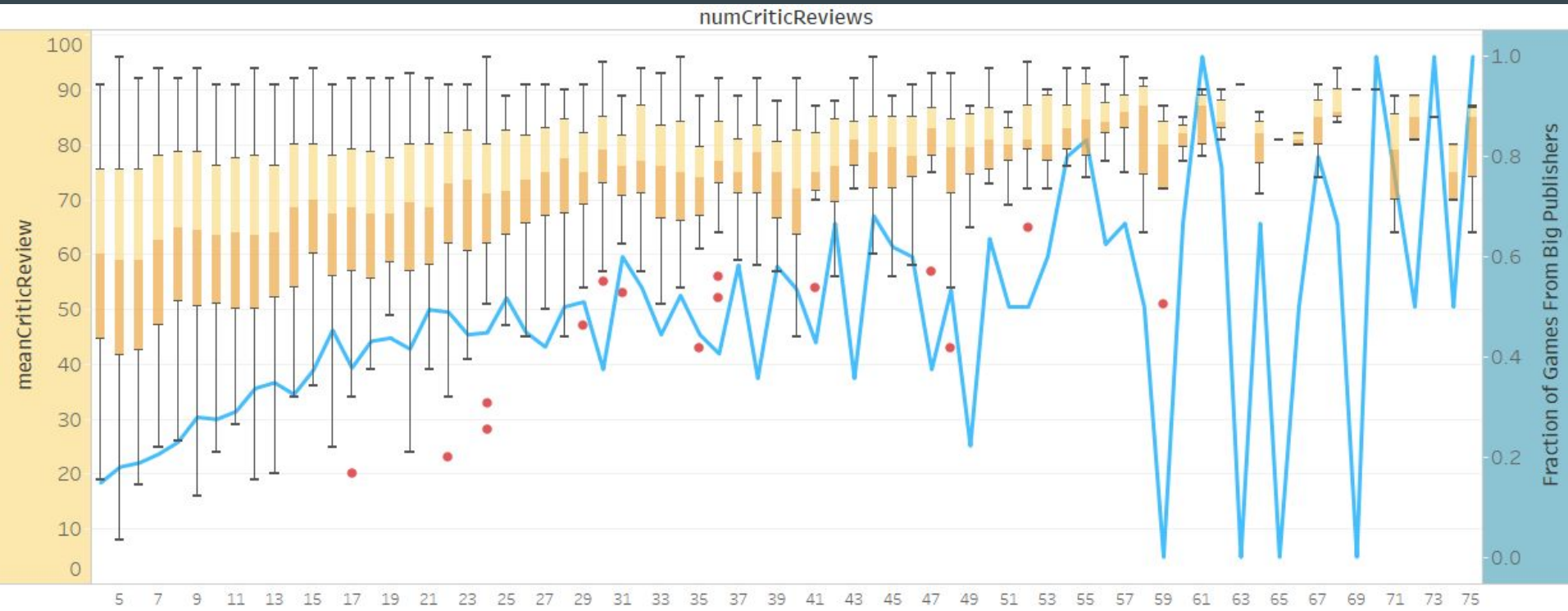
Are critics biased?

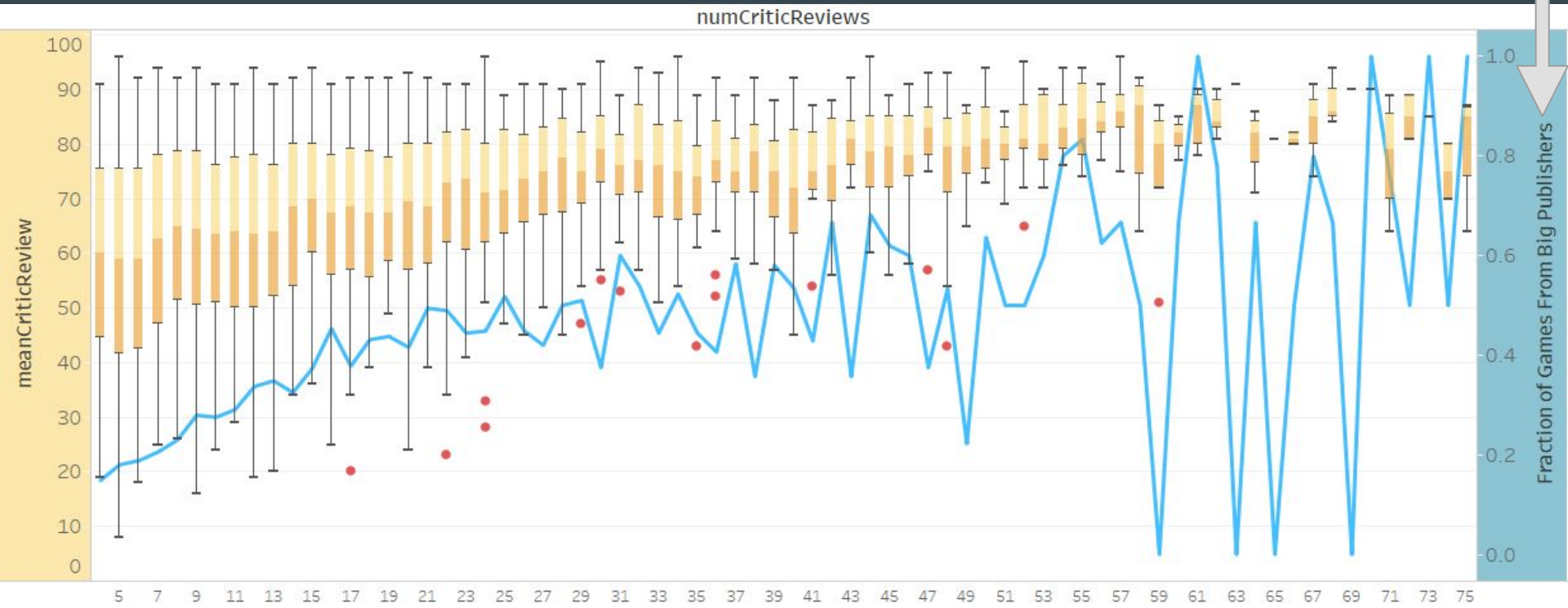
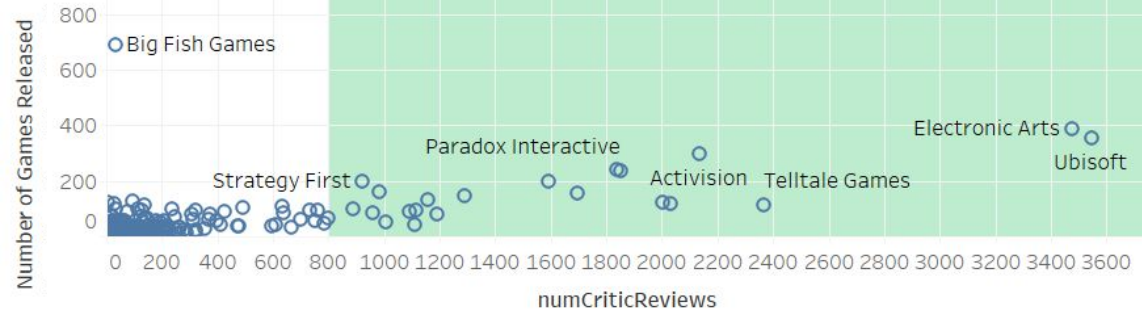
Popular games are better received. Is there causation behind this correlation?



Critic bias is one possible mechanism at work.

Popular games come from just a few places.



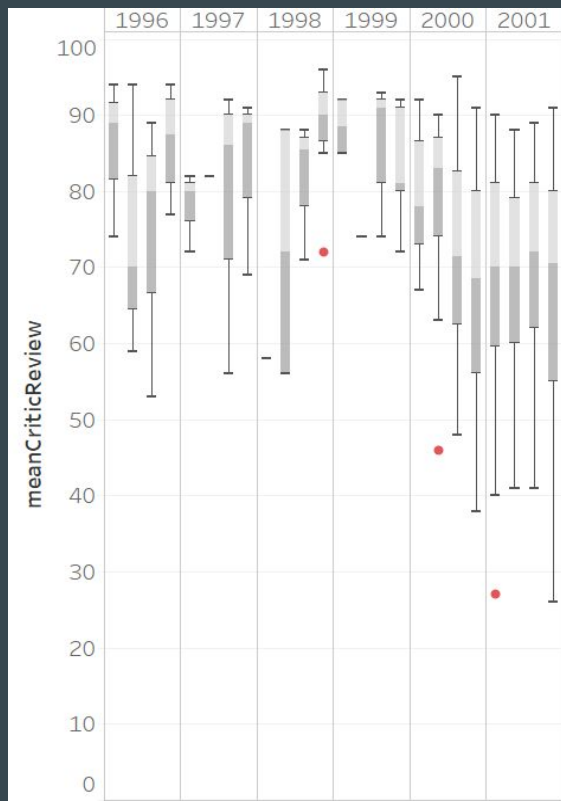
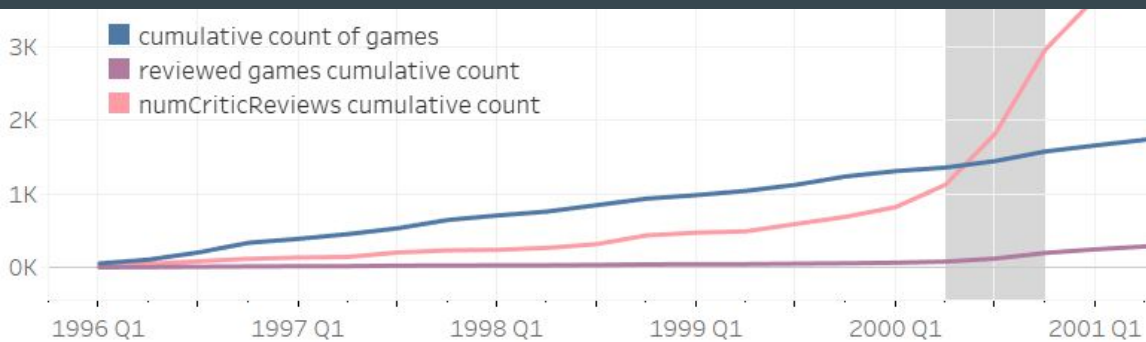


Critics or publishers, who is to blame?

- Absent data on individual critics, there is no certainty.
 - Lessening *meanCriticReview* IQR isn't enough to work with. It's likely this is due sample sizes rather than variability in critics' diversity of opinions.
 - A steady upper IQR bound hints at some critic objectivity. Unpopular, well received games exist.
 - Those few games receiving substantial critic attention are not representative of the population.
 - Often from big publishers, they'd also be subject to a unique, nonexistent, sampling strategy. They were likely given to media outlets for their critics to review. Most games are published through less influential avenues and chosen for review under different circumstances.

Are critics fair? - contradictory results?

- Didn't *meanCriticReview* decrease w/ *numCriticReviews*?
 - *meanCriticReview* changed w/ the slope of *numCriticReviews*.
 - This *numCriticReviews* differential is invariant for most of the previous plot's data members, that is, $dN/dt = \text{constant}$.
 - *meanCriticReview* is effectively independent of dN/dt so we look for dependence on N , the *numCriticReviews*.

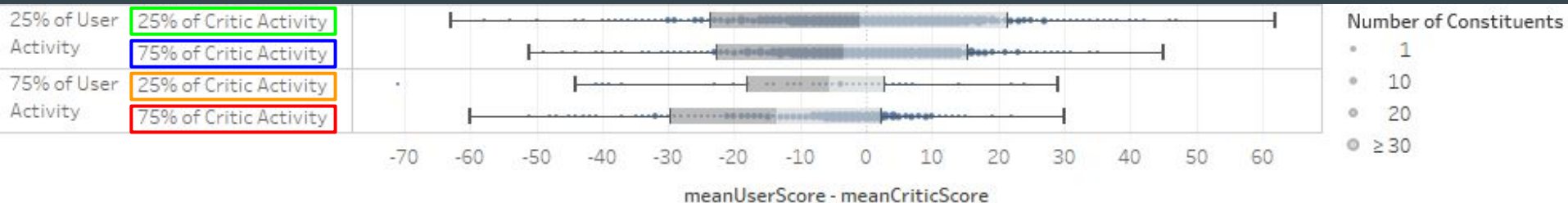


Are the users an angry mob?

- Not all of them, but they lean that way
 - Games that aren't the focus of users aren't as divisive as those that are.
 - When users amass, they're more pessimistic overall than critics.
 - Users are reactionary or drawn to the same titles as critics, given $\text{count} \ll \text{count}$
 - Critics exercise less selection bias, given $\text{count} + \text{count} \gg \text{count} + \text{count}$

The number of games receiving:

	25% of Critic Activity	75% of Critic Activity
25% of User Activity	1,748	1,838
75% of User Activity	56	616



The most explicit act of user wrath:

OUT OF THE PARK BASEBALL 2007 PC

Sports Interactive | Release Date: Mar 23, 2006



Summary

Critic Reviews

User Reviews

Details & Credits

Trailers & Videos



96

Metascore

Universal acclaim
based on 5 Critics

What's this?

2.5

User Score

Generally unfavorable reviews
based on 794 Ratings

Your Score

0

6

dana

Dec 9, 2009

Metacritic needs to have a limit of at least 20-30 reviews before a game can get awards like best PC game of the year or be the #2 PC game overall. This is ridiculous. This game is decent, but 96? No way.

39 of 49 users found this helpful

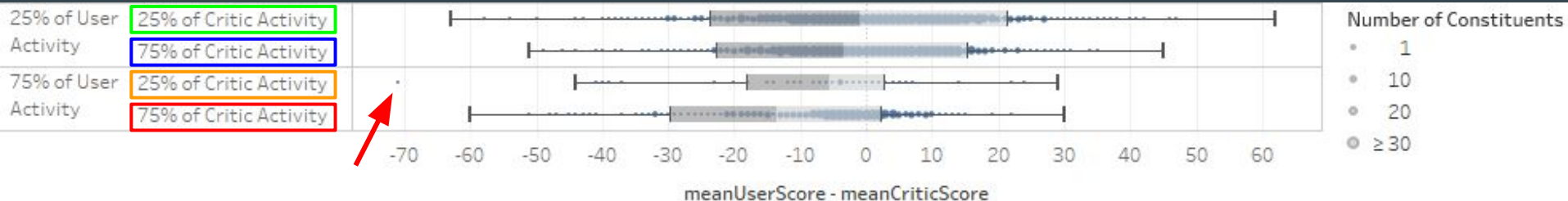
0

BobBob

Mar 28, 2009

this rating system is flawed. post all reviews for all games of every platform to get the real scores, not the ones that the owners of this site like so they can have an effect on the end score. this in no way deserves to have the 2nd highest rating ever.

63 of 79 users found this helpful



The most contentious game...

COMPANY OF HEROES 2 PC

Sega | Release Date: Jun 24, 2013



Summary

Critic Reviews

User Reviews

Details & Credits

Trailers & Videos



80

Metascore

Generally favorable reviews
based on **82 Critics**

What's this?

2.0

User Score

Generally unfavorable reviews
based on **5621 Ratings**

Your Score

0

0

alderkien

Aug 9, 2013

Not the slightest hint of historical authenticity. Very disapointing, uses the same graphics engine, the graphics look very similar but it brings my 3770K and 680GTX to a halt. IT has very few improvements over the first game. This just screams cash in to me, Stick with the oringal CoH its much superior

14 of 19 users found this helpful

[All this user's reviews](#)

0

Adjjezz1706

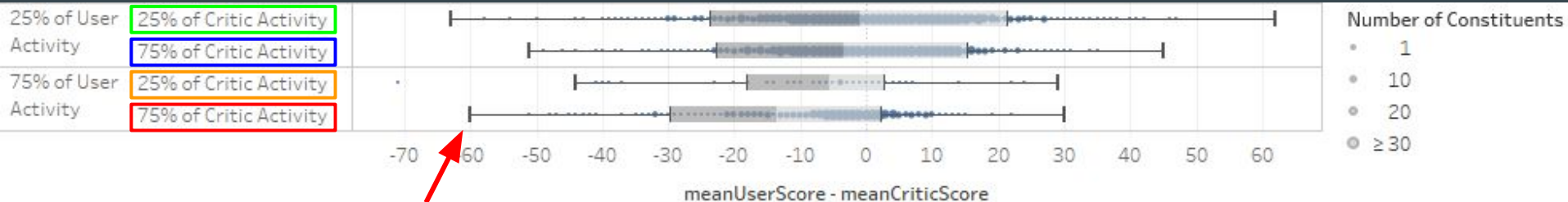
Apr 11, 2016

This game has 2 major aspects:

1. Gameplay: Nothing to write home about, the first one was quite a bit better.
- 2 Propaganda: This one is way better in portraying the Eastern front as a battle between the devil and the deep sea. Regrettable from Relic.

2 of 5 users found this helpful

[All this user's reviews](#)



... in more ways than one

COMPANY OF HEROES 2 PC

Sega | Release Date: Jun 24, 2013



Summary

Critic Reviews

User Reviews

Details & Credits

Trailers & Videos



80

Metascore

Generally favorable reviews
based on **82 Critics**

What's this?

2.0

User Score

Generally unfavorable reviews
based on **5621 Ratings**

Your Score

0

0

Uber-Urk

Aug 8, 2013

This game is created by survivor nazis. Gebbels would have standing ovation. I think Relic has deserved award from the Third Reich. We're waiting for games with Gebbels "truth" about Holocoste, extermination camp and racial theory. Next time, do not forget to measure the skull parameters of your customer

13 of 17 users found this helpful

All this user's reviews

0

antistatic

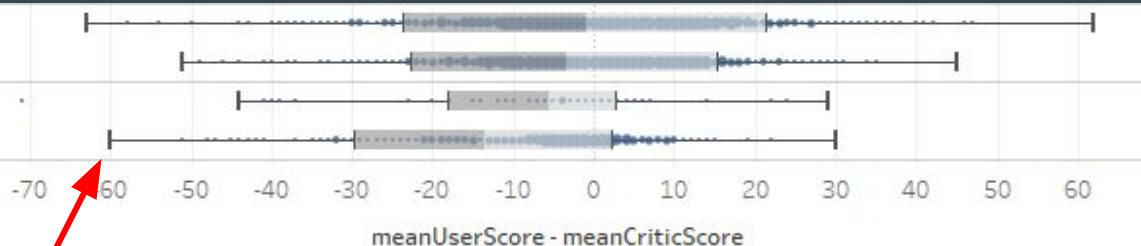
Aug 9, 2013

I resent how you portrayed in the game Russian soldiers. I resent how you portrayed in the game Russian soldiers. I resent how you portrayed in the game Russian soldiers. I resent how you portrayed in the game Russian soldiers. I resent how you portrayed in the game Russian soldiers.

18 of 23 users found this helpful

All this user's reviews

25% of User Activity	25% of Critic Activity
	75% of Critic Activity
75% of User Activity	25% of Critic Activity
	75% of Critic Activity



Number of Constituents

- 1
- 10
- 20
- ≥ 30

There are also fanboys on the defense....

AMERICAN MCGEE PRESENTS BAD DAY LA PC

Aspyr | Release Date: Sep 6, 2006



Summary

Critic Reviews

User Reviews

Details & Credits

Trailers & Videos



28

Metascore

Generally unfavorable reviews based on 24 Critics

What's this?

7.3

User Score

Mixed or average reviews based on 81 Ratings

Your Score

0

7

alk.:

Nov 19, 2006

Oh c'mon! It's a good game! Of course - it isn't wonderful, but I personally like a sense of humor in it.

0 of 0 users found this helpful

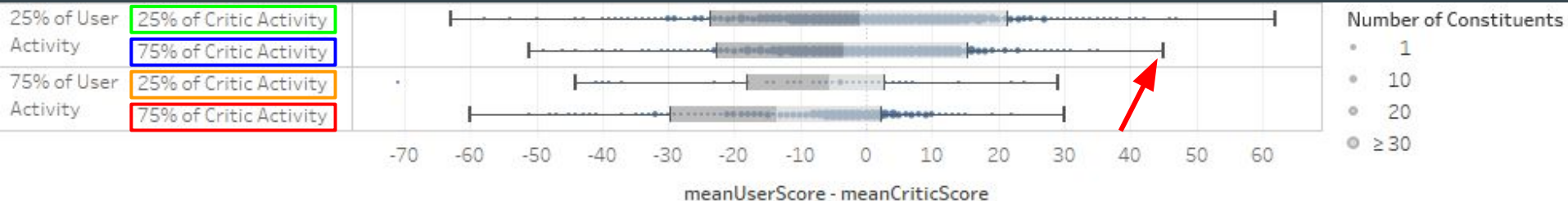
7

JulianM.

Mar 20, 2007

PEOPLE! Tell me, what do you want from this game? It's a nice action game with good atmosphere and crazy humor. If Anyone liked South Park and the Simpsons - take it. It's not a game for long midnights. One session should take about 30 min - 1 h. And then it gives lots of fun and laugh.

0 of 0 users found this helpful



...as well as trolls

NAVY SEALS: WEAPONS OF MASS DESTRUCTION PC

ValuSoft | Release Date: Jul 18, 2003



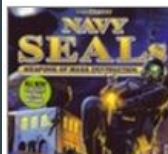
Summary

Critic Reviews

User Reviews

Details & Credits

Trailers & Videos



18

Metascore

Overwhelming dislike
based on 6 Critics

What's this?

8.0

User Score

Generally favorable reviews
based on 96 Ratings

Your Score

0

10

DrW.

Sep 14, 2003

Rivals half life in terms of quality.

6 of 12 users found this helpful

10

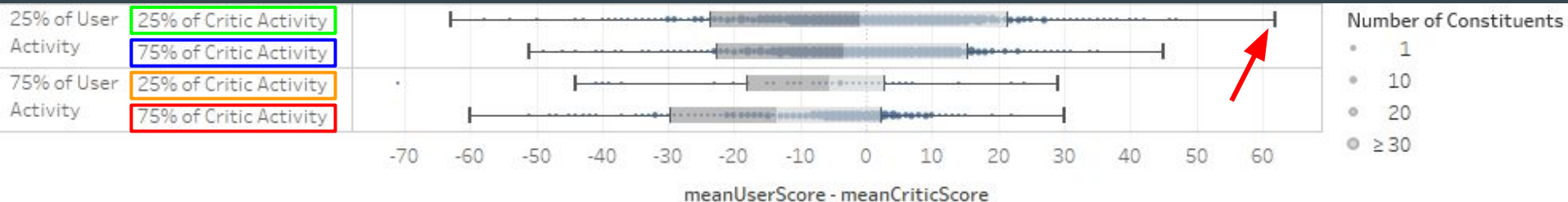
dev0_ru

Jun 20, 2012

This game changed my life... I never thought what 1 game can do this things with me... Awesome game...

4 of 7 users found this helpful

All this user's reviews



...as well as trolls
(again)

BIG RIGS: OVER THE ROAD RACING PC

GameMill Publishing | Release Date: Nov 20, 2003



Summary

Critic Reviews

User Reviews

Details & Credits

Trailers & Videos



8

Metascore

Overwhelming dislike
based on 5 Critics

What's this?

3.7

User Score

Generally unfavorable reviews
based on 1031 Ratings

Your Score

0

10

Rogueleader17

Aug 25, 2011

This game should be archived as the sole representation of our entire existence and cast out into the cold reaches of space to show all life in the universe what mankind was created to make, which is a masterpiece thousands of years in the making combining all of our collective knowledge and resources to make this game. It is so great of an achievement of mankind that I call it a game only for the lack of a word that would properly describe this gift from god. Collapse ▲

42 of 62 users found this helpful

All this user's reviews

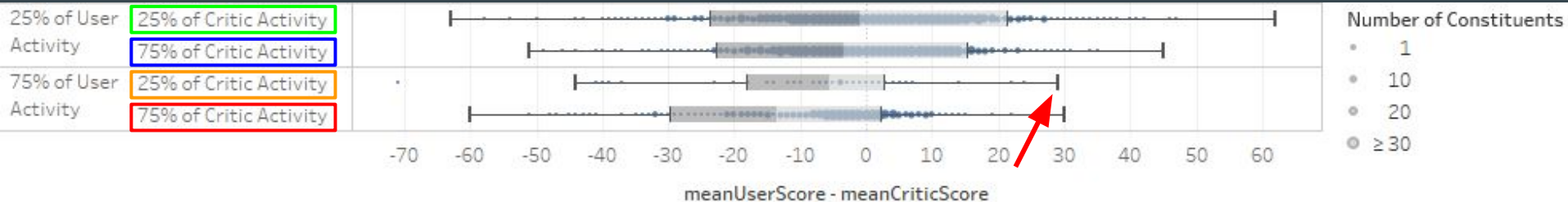
10

Walach

Aug 11, 2008

This game makes Half-Life 2 look bad.

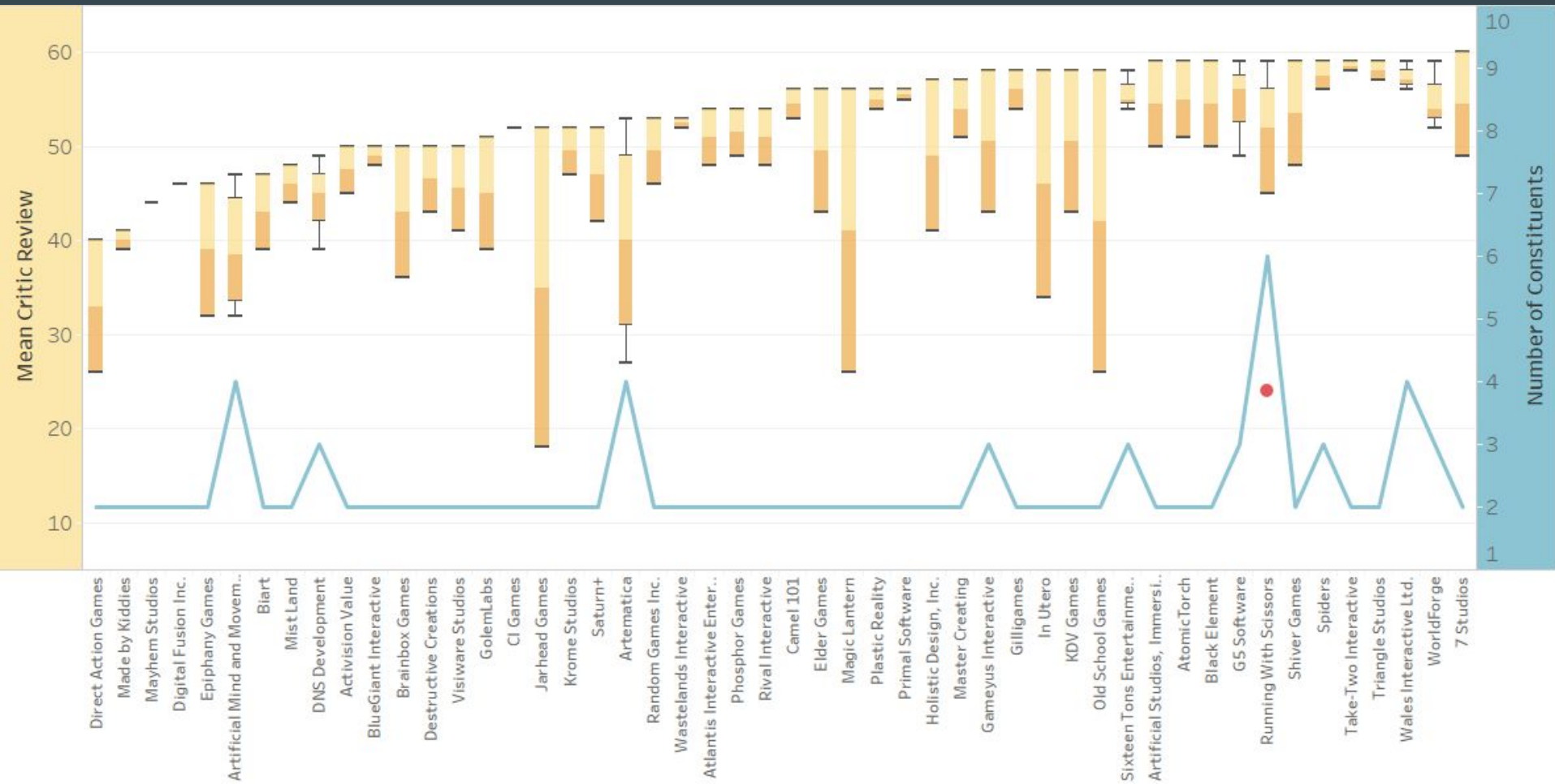
23 of 42 users found this helpful



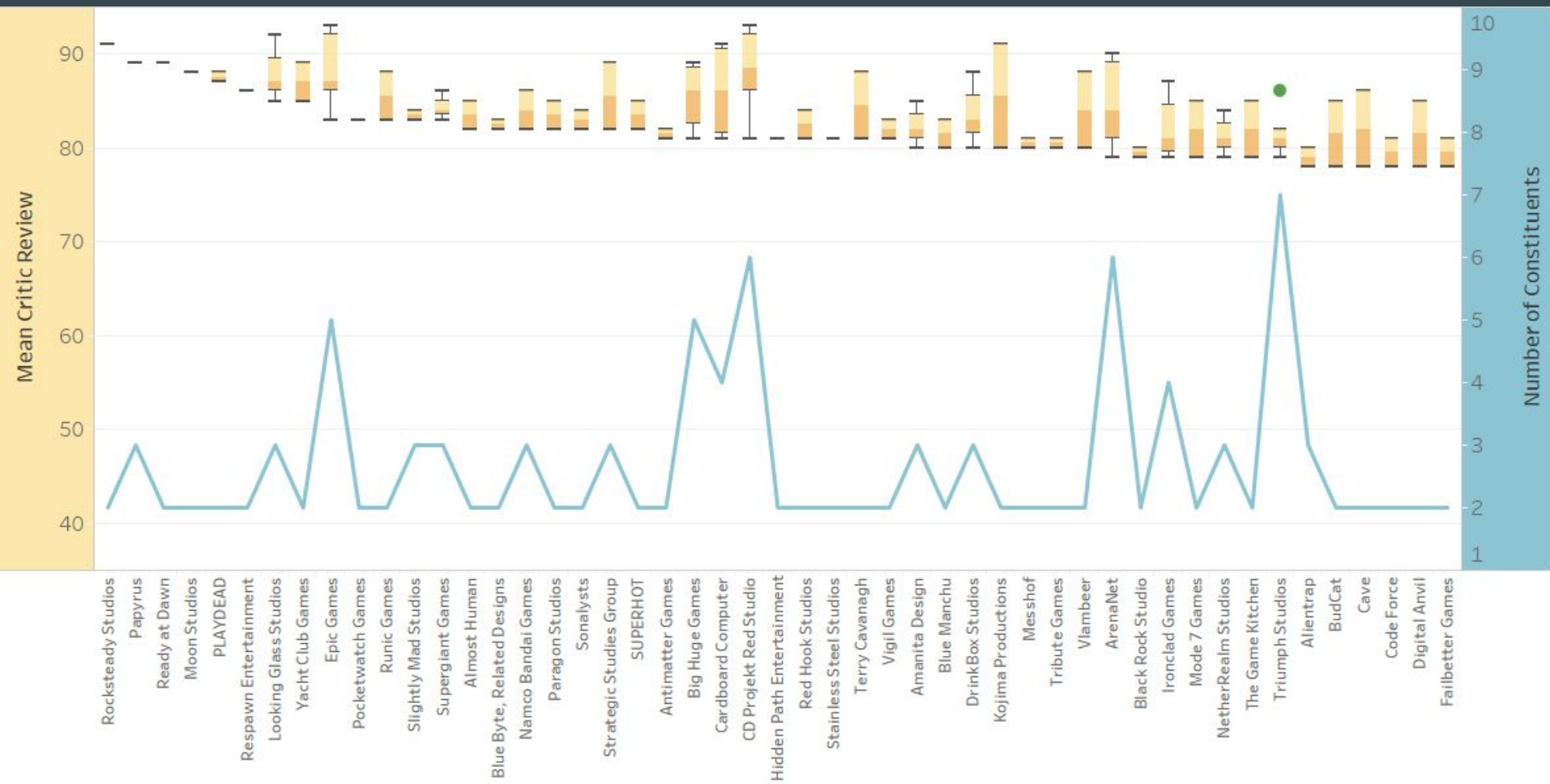
Are there predictable developers?

- You could expect a poor game to come from developers with histories of:
 - an upper IQR bound $<$ typical game *meanCriticReview* (~67)
 - multiple game releases (serial behavior is of concern here)
- Conversely, developers to place faith in can also be defined:
 - a lower IQR bound $>$ typical game *meanCriticReview* (~67)
 - multiple game releases (serial behavior is of concern here)

The worst 50 of 103 reliably poor developers:



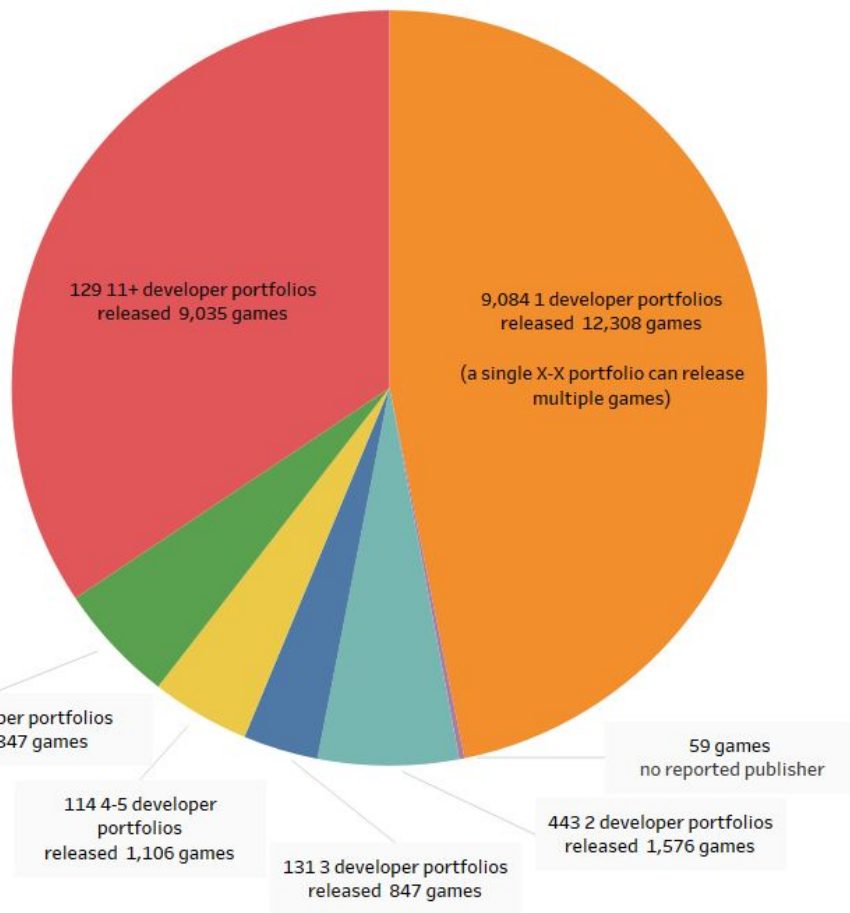
The top 50 of 232 reliably bankable developers:



On developer reliability...

- Developers in the positive category...
 - achieve more consistent review scores (smaller *meanCriticReview* IQRs)
 - release more games on average than developers in the negative category
 - outnumber the negative category more than 2:1
- Most developers (the remaining 438) defy categorization and another 11,997 have only released a single game.

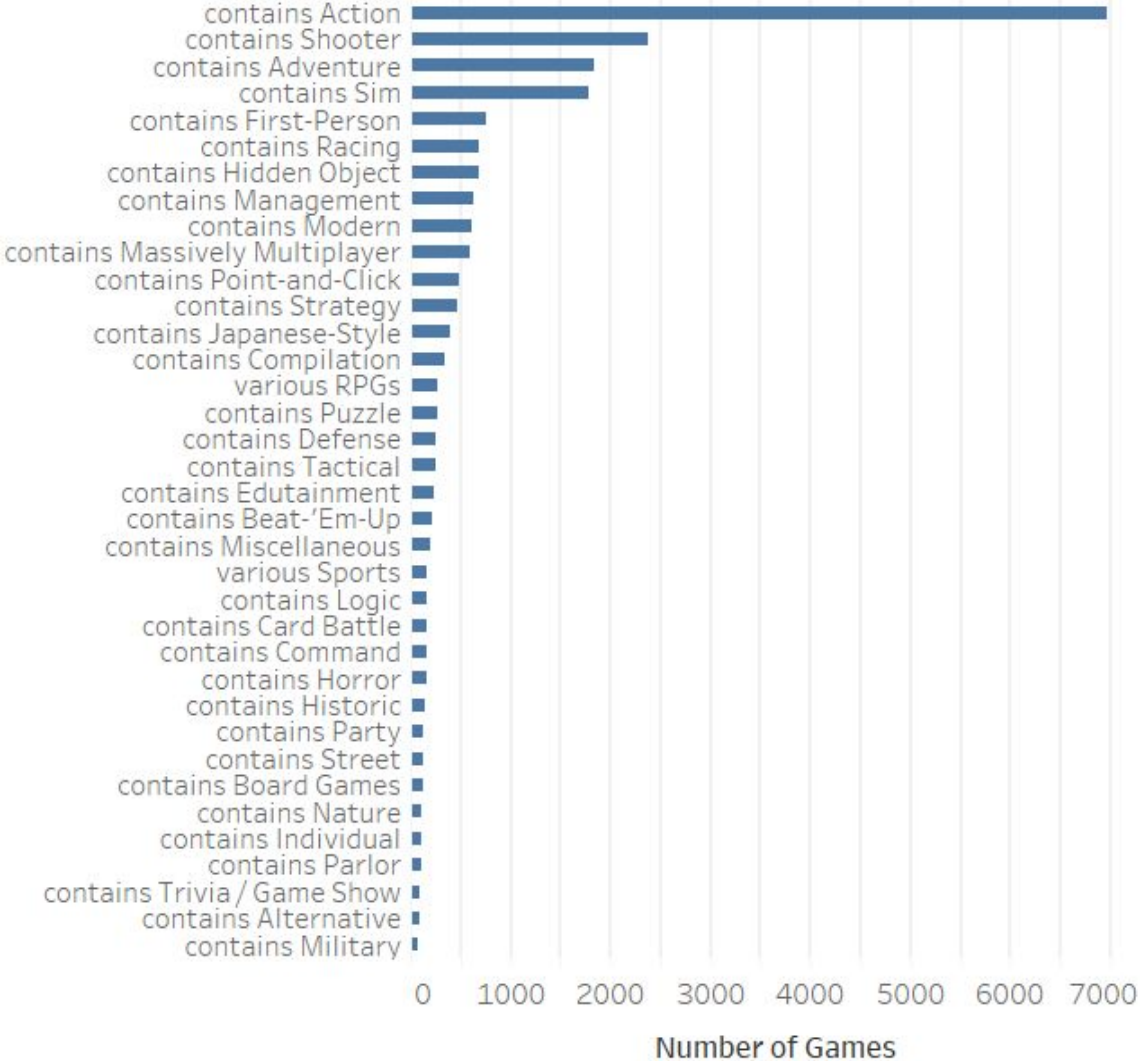
Are developers independent of publishers?



Publishers often have production & promotion advantages to help game sales.

This incentivizes a top heavy **hegemony** with substantial market presence by way of one publisher doing business with many developers.

Opting out of this by **self publishing** is quite common.

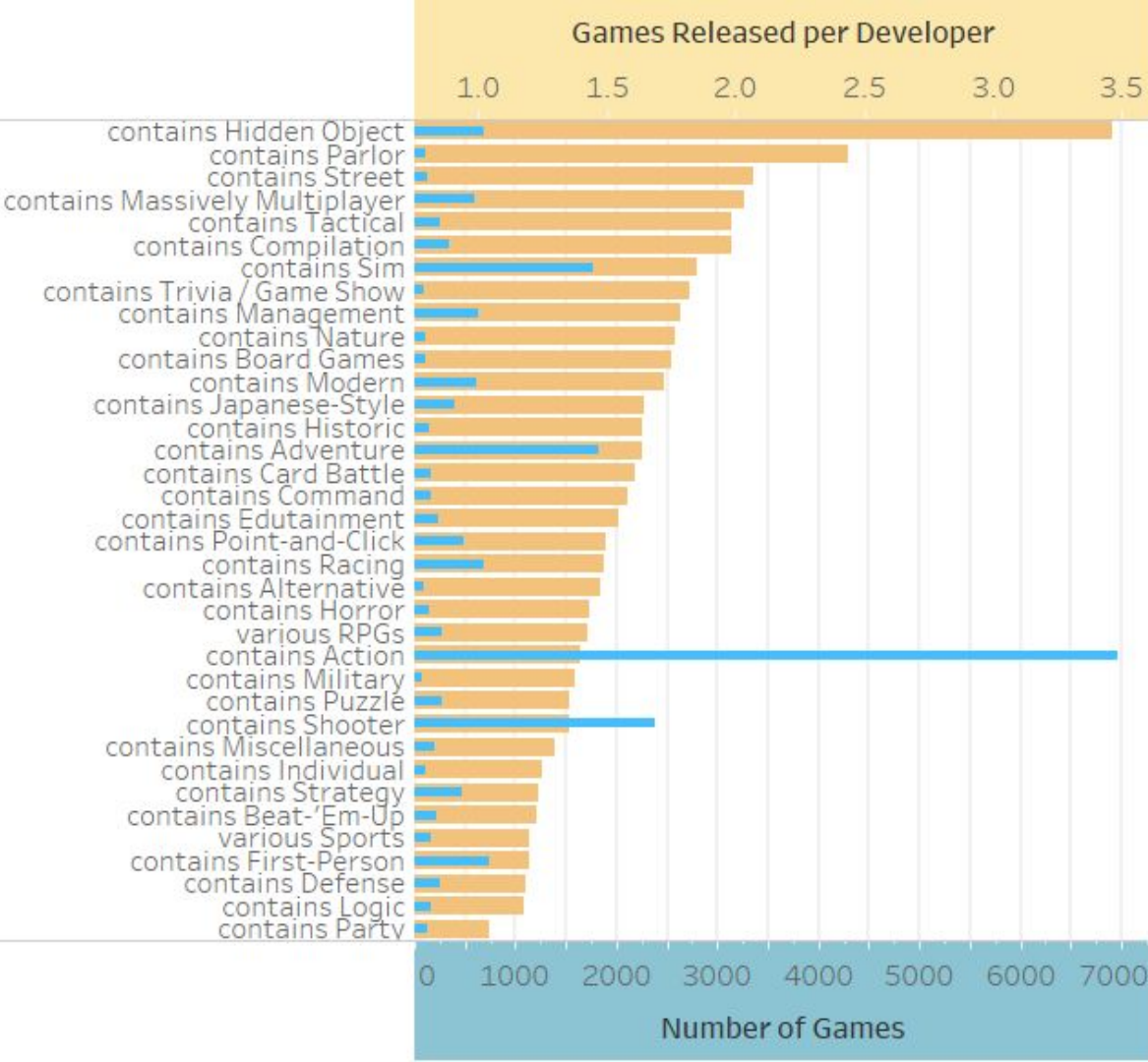


What genres are popular?

High stimulus ones.

‘Action’ leads by a wide margin, being a catch-all term, followed by ‘shooters’ & ‘adventure.’

~3800 games belong to a nondescript ‘general’ genre that we ignore, as well as those w/ < 75 games.



Some genres have low barriers of entry.

I hear little of 'Hidden Object' games despite their popularity.

Theorizing a few developers are produce them ad nauseam, I compute a game releases/developer ratio.

Watching gameplay corroborates that these require less effort to make. The level of interactivity is low, as is the demand for complex visuals, AI, etc.

This is also true of 'street' & 'parlor' games, though 'party' is oddly exempt.

Concluding thoughts

- Tableau is suboptimal for science friendly visuals expressing aggregation/ uncertainty
 - displaying means w/ 1σ error bars requires a workaround that removes dual axis functionality
 - boxplots are a poor, quirky alternative
 - continuous integer fields must be treated as discrete
 - reducing formatting options like x axis labels on top of plots
 - whisker magnitude is variable and can mislead
 - make use of grouping, filtering. etc. to quickly explore data before making cleaner plots elsewhere
- Advice on picking up these skills for yourself:
 - Scrapy: start with a [CrawlSpider](#) and [this](#) tutorial
 - pandas: [10 minutes to pandas](#)
 - Tableau (which is still pretty nice despite my complaining): [Coursera](#)

Concluding thoughts - a todo list

As was [already mentioned](#), no hardline conclusion can be drawn regarding critic bias w/o certain data.

The *fractionCritic_* & *fractionUser_* features are unutilized, but they may be of use in classification tasks.

An analogous [publisher pie chart](#) from console game data would isolate the difference, if any, in online distribution's disruption of the publisher hegemony.

Appendix

...

Supplementary information

Features

title: game name

*developer**

*publisher**

releaseDate

genres: the unique set of comma separated genres

esrb: ESRB content rating

meanCriticReview: Metascore

fractionCriticPositive: positive fraction of critic reviews

fractionCriticMixed: mixed fraction of critic reviews

fractionCriticNegative: negative fraction of critic reviews

*If one is present but not the other, their values become identical

Features

numCriticReviews: critic review count

meanUserReview: User Score multiplied by 10 for *meanCriticReview* comparisons

*fractionUserPositive**: positive fraction of user reviews

*fractionUserMixed**: mixed fraction of user reviews

*fractionUserNegative**: negative fraction of user reviews

numUserReviews: user review count

url: the game's web page (used as a key)

scrapyStatus: Scrapy status code for the URL's crawl

*Only user reviews w/ a written component contribute to this figure; these counts are constituents of

numUserReviews

Scrapy visitation details

Scrapy isn't browsing page elements like the diagram implies.

Scrapy queues URLs in depth first order using nested comparisons (`scrapy.spider.Rule`) to regular expressions.

What is represented is the logical structure and outcome of the nested regex comparisons.

The end result is that all URLs at those page elements will have been visited, albeit in a different order.

Scrapy visitation details - verbose

- list of start URLs: 'http://www.metacritic.com/browse/games/title/pc/{}'
 - {} = 'a' - 'z' + blank char " used for number sorted titles
- allow transitions among game lists within their alphanumeric parent dir
 - [http://www.metacritic.com/browse/games/title/pc?page=\[1-9\]{1,3}](http://www.metacritic.com/browse/games/title/pc?page=[1-9]{1,3})
 - [http://www.metacritic.com/browse/games/title/pc/\[a-z\]\?page=\[1-9\]{1,3}](http://www.metacritic.com/browse/games/title/pc/[a-z]\?page=[1-9]{1,3})
- visit game pages within each game list
 - allow game pages
 - [http://www.metacritic.com/game/pc/\[^\\\]+](http://www.metacritic.com/game/pc/[^\\]+)
 - deny children deeper than game's page
 - <http://www.metacritic.com/game/pc/\\S+/>
 - deny false positives (sorts, reviews, etc.) at game page's level
 - [http://www.metacritic.com/game/pc/\[a-z0-9\\-\\!\]+\[\\?\\=\]](http://www.metacritic.com/game/pc/[a-z0-9\\-\\!]+[\\?\\=])

The list of reliably poor developers

7 Studios, 7FX, 369 Interactive, Activision Value, Artematica, Artificial Mind and Movement, Artificial Studios, Atlantis Interactive Entertainment, AtomicTorch, B-Alive, Biart, bitComposer Games, Black Element, BlueGiant Interactive, Born Ready Games, Brainbox Games, Burut Software, Camel 101, Cat Daddy Games, Centauri Production, CI Games, Crocodile Entertainment, Daydream Software, Destructive Creations, Digital Fusion Inc., Digital Leisure, Direct Action Games, DNS Development, EA Bright Light, Elder Games, Epiphany Games, Etranges Libellules, Eurocom Entertainment Software, Eutechnyx, Future Games, G5 Software, Galilea Multimedia, Gameyus Interactive, Gaming Corps AB, Gilligames, GolemLabs, Hangar 13, Holistic Design, Inc., Hothouse Creations, Hydravision, Ice Water Games, In Utero, Jarhead Games, Juice Games, KDV Games, Krome Studios, Kylotonn, Made by Kiddies, Magic Lantern, Master Creating, Mayhem Studios, Metropolis Software, Mindware Studios, Mist Land, Muse Games, Old School Games, Phosphor Games, Pivotal Games, Plastic Reality, Primal Software, Propaganda Games, Provox Games, Random Games Inc., Realmforge Studios, Rebelmind, Reflections Interactive, Revistonic, Revistronic, Rival Interactive, Running With Scissors, Saturn+, Shiver Games, Shrapnel Games, Silent Dreams, Similis, Sixteen Tons Entertainment, Slitherine, Slitherine Group, Small Rockets, Sonic Team, souvenir circ., Spark Unlimited, Spiders, Streko-Graphics Inc., Stream On Studio, Take-Two Interactive, Three Fields Entertainment, Tri Synergy, Triangle Studios, ValuSoft, Vile Monarch, Visiware Studios, VSTEP, Wales Interactive Ltd., Wastelands Interactive, WorldForge, Zeal Game Studios, ZootFly

The list of reliably bankable developers

2By3 Games, 2K Australia, 3 Sprockets, 5TH Cell, 10tons Entertainment, 10tons, 11 bit studios, 17-Bit, 2015, Abbey Games, Alientrap, Alkemi Games, Almost Human, Amanita Design, Ankama Games, Antimatter Games, ArenaNet, Arkane Studios, Avalanche Software, Barking Dog, Benjamin Rivers, BestGameEver.com, Big Fish Games, Big Huge Games, Black Box, Black Forest Games, Black Isle Studios, Black Rock Studio, Blendo Games, Blizzard Entertainment, Bliober Team, Blue Byte, Related Designs, Blue Manchu, Browsome, BreakAway Games, Bright Future GmbH, BudCat, Bugbear, Bungie, Cadenza, Capcom, QLOC, Cappy Games, Cardboard Computer, Cave, Cavedog Entertainment, CCP, CD Projekt Red Studio, ChessBase, Code Force, Colossal Order, Crackshell, CreativeForge Games, Croteam, Crystal Dynamics, Nixxes Software, Darkling Room, Davey Wreden, Dennaton, Digital Anvil, Digital Eclipse, Digital Eel, Digital Illusions, Double Eleven, Dreadbit, DrinkBox Studios, Dynamix, EA DICE, EA Phenomic, Eagle Dynamics, Eidos Interactive, Eidos Montreal, Epic Games, Erik Svedang, Eugen Systems, Exordium Games, Failbetter Games, Falcom, Flying Lab Software, Frog City Software, Futuremark Games Studio, Gaijin Games, Ghost Games, Gigawatt Studios, Gray Matter, Haggard Games, Harvester Games, Her Interactive, Hi-Rez Studios, Hidden Path Entertainment, House of Tales, Illwinter Design Group, Image & Form, Infinite Interactive, Infinity Ward, Introversion, Ion Storm, Iron Galaxy Studios, Iron Lore Entertainment, Ironclad Games, Ironward, Irrational Games, Johamm Tael & Mihkel Tael, K-D Lab, Katauri Interactive, KCET, Koios Works, Kojima Productions, Laminar Research, Larian Studios, Legend Entertainment, Lexis Numerique, Lionhead Studios, Logic Artists, Looking Glass Studios, LucasArts, BioWare, MachineGames, Malfador Machinations, Massive Damage, Inc., Massive Entertainment, Matrix Games, Mediatonic, Messhof, Mi'pu'mi Games, Midway, Mike Bithell, Mimimi Productions, Mode 7 Games, Mojang AB, Moon Studios, Mouldy Toof Studios, Mousechief, Mythic Entertainment, Nadeo, Namco Bandai Games, NetDevil, NetherRealm Studios, Neversoft Entertainment, New World Interactive, Nexon, Nicalis, Ninja Theory, Nippon Ichi Software, Northway Games, Omni Creative Group, Opus, Overhaul Games, Overkill Software, Owlchemy Labs, Panther Games, Papyrus, Paragon Studios, Passtech Games, People Can Fly, Planet Moon Studios, PLAYDEAD, Pocketwatch Games, PopCap, Press Play, Psyche Studios, Pulse Entertainment, Puny Human Games, Q-Games, Radical Entertainment, Ratbag, Ready at Dawn, Red Barrels, Red Hook Studios, Red Redemption, Red Thread Games, Related Designs, Relic, Relic Entertainment, Reptile, Respawn Entertainment, Rockstar North, Rockstar San Diego, Rocksteady Studios, roll7, Runic Games, S2 Games, Shockwave Productions, SkyGoblin, Slightly Mad Studios, Sonalysts, Spiderweb Software, Spike Chunsoft, Sports Interactive, SSI, Stainless Steel Studios, State of Play Games, Straandlooper, Strategic Studies Group, Supergiant Games, SUPERHOT, Swing Swing Submarine, Syntetic, Taldren, Taleworlds, Tequila Works, Terry Cavanagh, The Bitmap Brothers, THE BROTHERHOOD, The Chinese Room, The Game Kitchen, Thunder Lotus Games, Tomorrow Corporation, TopWare Interactive, Torn Banner Studios, Totally Games, Transhuman Design, Tribute Games, Triumph Studios, Troika Games, Twisted Pixel Games, Two Tribes, Ty Taylor and Mario Castaneda, U.S. Army, Ubisoft Paris, Ubisoft Reflections, Undead Labs, Valve Software, Vigil Games, Visceral Games, Visual Concepts, Vlambeer, VooFoo Studios, Wargaming.net, Westwood Studios, Yacht Club Games, Zachtronics Industries, Zeboyd Games, ZeniMax Media, Zoink!

Use of repository contents

1. crawl with the metacritic.py spider, being sure to output data (data_initial.csv)
 - a. `cd ./critic` #go to the scrapyproject critic
 - b. `scrapy crawl metacritic` #crawl with the metacritic spider
2. conduct phase 1 of inspection:
 - a. run inspectP1.py on data_initial.csv, receiving inspect.csv
 - b. manually address the issues in the *inspect* field of inspect_revised.csv & clear the *inspect* field
3. conduct phase 2 of inspection:
 - a. run inspectP2.py on inspect.csv, receiving inspect_revised.csv
 - b. manually address the issues in the *inspect* field of data_final.csv & remove the *inspect* field
 - c. the resultant is data_final.csv
4. explore your data_final.csv (in this case with a Tableau workbook)

Extras: test_cases.csv & its validation pipeline iterations are provided as an example and proof of function.

Behind the scenes of *inspect* - phase 1

- must be $[0, \infty)$ bounded integers:
 - *numCriticReviews*, *numUserReviews*
 - *meanCriticReview*, *meanUserReview*
 - *scrapyStatus*
- must be $[0, 1)$ bounded floats:
 - *fractionCriticPositive*, *fractionCriticMixed*, *fractionCriticNegative*
 - *fractionUserPositive*, *fractionUserMixed*, *fractionUserNegative*
- is a datetime:
 - *releaseDate*
- *scrapyStatus* = 200, else the crawler experienced an error at this url

Behind the scenes of *inspect* - phase 1

- no suspicious presence of non ASCII alphanumeric strings:
 - *title, developer, publisher*
 - exceptions made to characters observed to parse correctly:
.:_-,.+!'() ` / # & * ~ “ | ? % \$ @ [] { } and whitespace
 - *esrb* is exempted due to false positives on nulls; the set of observed values parses correctly

Behind the scenes of *inspect* - phase 2

- $\Sigma \text{fractionCriticPositive}, \text{fractionCriticMixed}, \text{fractionCriticNegative} = 1 \pm 0.001$
 - or all null
- $\Sigma \text{fractionUserPositive}, \text{fractionUserMixed}, \text{fractionUserNegative} = 1 \pm 0.001$
 - or all null