Superhero Movie Performance in China: How Profitable are They?

Alison Yao (yy2564)

Superhero movies are one of the highest-grossing movie franchises in the world. China, with a population of almost 14 billion, has generated extraordinary box office performance and huge gross for these superhero movies. However, in mainland China, all entertainment content such as TV and films are censored by National Radio and Television Administration (NRTA) and China Film Administration (CFA). China mainland allows merely 34 foreign films into the Chinese market every year, so both the movie production companies (such as Marvel Studios and DC Film), and the CFA make an effort to introduce superhero movies thanks to their profitability for both parties.

As a Shanghainese and devoted fan who insists on cinematic experience, I will take this opportunity to investigate the relationship between the production budget of these superhero movies and the total gross in China. I gathered a sample of 35 superhero movies from *The Numbers* (www.the-numbers.com) and *Box Office Mojo* (www.boxofficemojo.com), listing first release year (Year), movie title (Movie Title), production budget (Budget), and Chinese market gross (Chinese Gross). Therefore, the regression model should of the form:

$$Chinese\ Gross = \beta_0 + \beta_1 \cdot Budget + random\ error$$

The units of Budget and Chinese Gross from the online sources are both in US dollars. An example of the dataset is shown below:

Year	Movie Title	Budget ($)	Chinese Gross ($)
2002	Spider-Man	139000000	4983142
2004	Spider-Man 2	200000000	6102882
2007	Spider-Man 3	258000000	18991487
2008	Iron Man	140000000	15274332
2008	The Incredible Hulk	150000000	9336251

Not all Marvel and DC movies are superhero movies. For example, Deadpool and Venom are anti-heroes. Therefore, I first made a list of only superheroes who are active on the big screen. I looked up each movie title on *The Numbers*, and copy-pasted the budget under "Title Summary" and the Lifetime Gross in China under "Performance". For example, *Spider-Man* (2002) data is here: https://www.boxofficemojo.com/title/tt0145487/?ref_=bo_se_r_1. Please note that lifetime gross is the summation of gross for several releases. Sometimes *The Numbers* lists "Gross" instead of "Lifetime Gross", but they mean the same if a movie is only released

once. It is possible that *The Numbers* does not provide the gross in China because a movie might not have made the 34-foreign-film list and was not released in China mainland at all. Occasionally, *The Numbers* does not provide the budget information, so I turned to this list from *Box Office Mojo* (https://www.the-numbers.com/movies/franchise/Marvel-Cinematic-Universe#tab=summary) for budget information. The two websites complemented each other and eventually, I acquired a sample of 35 superhero movies with both budget (Budget) and total Chinese market gross (Chinese Gross) information. Due to COVID-19 and the closing of cinemas, I only selected movies released before 2020. The raw data in the Budget and the Chinese Gross columns are very large, so I scaled them from dollars to million dollars, but this does not change the form of the regression model or any plots or intervals, or test statistics compared to the original unit. The first few observations of the adjusted dataset are shown below:

Year	Movie Title	Budget (Mil$)	Chinese Gross (Mil$)
2002	Spider-Man	139	4.983142
2004	Spider-Man 2	200	6.102882
2007	Spider-Man 3	258	18.991487
2008	Iron Man	140	15.274332
2008	The Incredible Hulk	150	9.336251

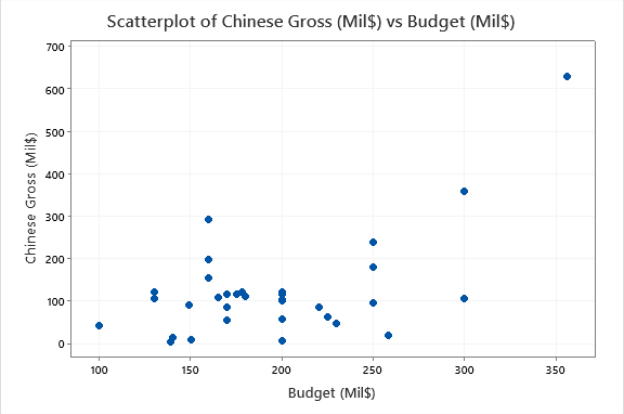
Some basic statistics are showed below:

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Chinese Gross (Mil$)	35	0	118.6	20.0	118.2	5.0	55.3	105.1	121.2
Budget (Mil$)	35	0	195.29	9.15	54.15	100.00	160.00	180.00	225.00

Variable	Maximum
Chinese Gross (Mil$)	629.1
Budget (Mil$)	356.00

Using Minitab, we can visualize the data in a scatter plot.



There doesn't seem to have a strong linear relationship here. But we can see that most points are clustered within the budget of 100~300 million dollars and the Chinese Gross of 0~400 million dollars, with the exception of the one on the top right. *Avengers: Endgame* has a shocking budget of over 350 million dollars and a Chinese gross of over 600 million dollars. Let's not worry about it for now and try fitting the data to a least squares regression model with Chinese Gross being the predicting variable and Budget as the predicting variable.

Regression Equation

Chinese Gross (Mil$) = -121.6 + 1.230 Budget (Mil$)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-121.6	63.5	-1.91	0.064	
Budget (Mil$)	1.230	0.314	3.92	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
99.0825	31.77%	29.70%	5.24%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	150846	150846	15.37	0.000
Budget (Mil$)	1	150846	150846	15.37	0.000
Error	33	323973	9817		
Lack-of-Fit	18	253905	14106	3.02	0.018
Pure Error	15	70067	4671		
Total	34	474819			

Fits and Diagnostics for Unusual Observations

Obs	Chinese Gross (Mil$)	Fit	Resid	Std Resid		
30	291.8	75.2	216.6	2.23	R	
32	629.1	316.3	312.8	3.74	R	X

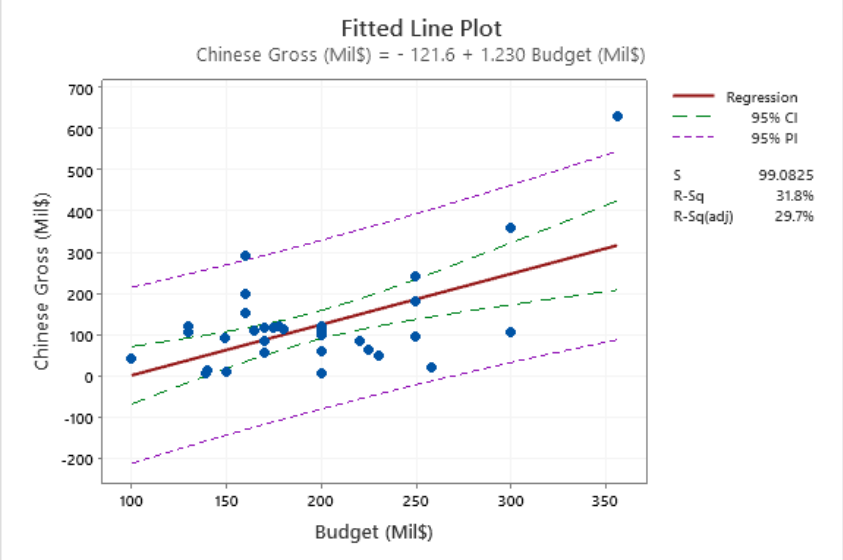
R Large residual
X Unusual X

The regression is relatively weak, with an R-squared of 31.77%, which means that Budget accounts for 31.77% of the observed variability in Chinese Gross. The adjusted R-squared of 29.70% is slightly lower than R-squared, but the predicted R-squared is only about one-sixth of the previous two, which means the model is not as good as making predictions on the population. The F-test is highly statistically significant, with a p-value of less than 0.001, so the relationship is very strong. The intercept means that given that a superhero movie has a Budget of 0 million dollars, the estimated expected Chinese Gross is -121.6 million dollars, but

this is not very meaningful here because the production budget of any movie has to be a positive number. The slope coefficient means that a 1 million dollar change in Budget is associated with an estimated expected 1.230 million dollar change in the Chinese market gross. The standard error of the estimate of 99.0825 means indicates that the model is useful in prediction to the extent that it can predict changes in Chinese market gross to within ± 198.165 million dollars, about 95% of the time.

In the section of Coefficients, the t-statistic given in the output of Budget (Mil$) tests whether or not to reject the null hypothesis of $\beta_1 = 0$. Its p-value of less than 0.001 is very small, so we strongly reject the null hypothesis. Therefore, there is a relationship between the changes in superhero movie budget and the changes in Chinese market gross. The t-statistic in the output of Constant tests whether or not to reject the null hypothesis of $\beta_0 = 0$. Its p-value of 0.064 is marginally statistically significant, so we have some marginal evidence against the null hypothesis.

The following scatter plot marks the confidence interval and the prediction interval.



The confidence interval can be used to estimate the true average Chinese market gross for all superhero movies in the population that have a budget of a certain value; the prediction interval can be used to estimate the value of Chinese gross for a particular superhero movie that has the budget of a certain value. Let's take Marvel's *Black Widow* as an example. It was scheduled to release in May 2020 but has been put off for a year because of COVID-19. Although movies have an estimated budget to begin with, the true budget is not solidified until a little earlier than release because of unexpected modification and post-production. *Black Widow*

is a special case that is finished with production but is not released yet. And its budget is around 200 million dollars. Let's set 200 million dollars as "a certain value" mentioned above.

Regression Equation

Chinese Gross (Mil$) = -121.6 + 1.230 Budget (Mil$)

Settings

Variable	Setting
Budget (Mil$)	200

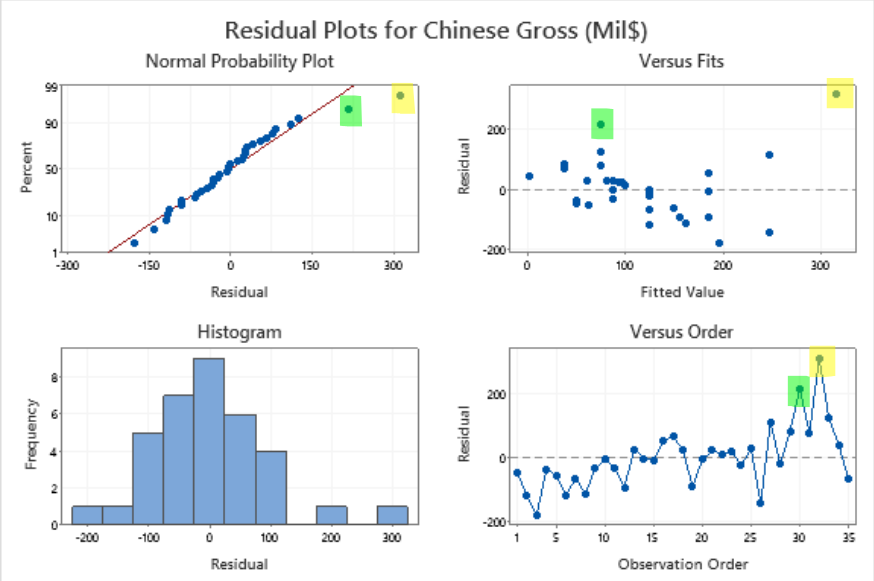
Prediction

Fit	SE Fit	95% CI	95% PI
124.424	16.8132	(90.2176, 158.631)	(-80.0423, 328.891)

The output from Minitab shows that the estimated average Chinese gross for all superhero movies in the population that have a budget of around 200 million dollars is in the range of 90.2176 million dollars and 158.631 million dollars. The output also shows that for the movie *Black Widow*, the estimated Chinese gross is within the range of –80.0423 million dollars to 328.891 million dollars. The lower bound is not very helpful, as the gross for any region cannot be negative. This regression is most definitely not sufficient to make a very precise prediction, but we can wait for two more months and see how the actual Chinese gross will turn out to be. *Black Widow* has officially made the 34-foreign-movie list in China on March 2nd, 2021.

As usual, the confidence interval is a subset of the prediction interval because the prediction interval takes into account the innate variability of Chinese Gross about the regression line in the population. Moreover, it is quite easy to spot that the confidence interval is narrower near the mean of budgets and wider to its left and right (more so to the right because the mean is on the left). Therefore, superhero movies with an average budget of 200 million dollars (class to the mean), as mentioned above, should yield a more accurate confidence interval, while the prediction of 100 million dollars or 400 million dollars should yield a less accurate confidence interval. Notably, besides *Avengers: Endgame*, the DC superhero movie *Aquaman* is also outside of the 95% prediction interval. The movie *Aquaman* seems to be especially popular in China.

Now, let's look at the "four in one plot" to check if the previous unusual points are really unusual and also check if our assumptions of the least squares regression hold true.



The plot of Residual versus Fits does not exhibit a lack of pattern, therefore indicating nonconstant variance. The normal plot of residuals has two points on the top right that deviate a lot from the straight line, displaying non-normality. As expected, there are indeed two unusual points in the plots. The ones highlighted in yellow represent *Avengers: Endgame* and the ones highlighted in green represent *Aquaman*.

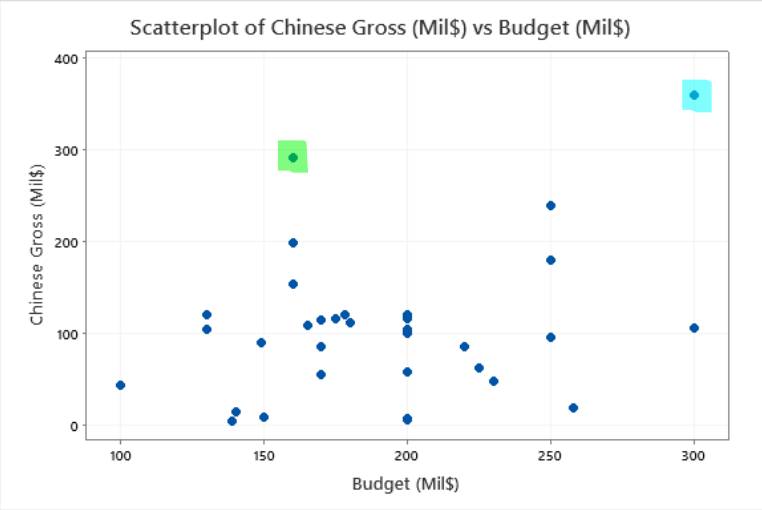
Avengers: Endgame (highlighted in yellow) is a leverage point and an outlier. Its predictor value and its target value are unusually large. As mentioned above, this superhero movie is the most commercially successful Marvel movie there is. It builds on its previous Avenger movie, *Avengers: Infinity War*, where half of the population in the universe is disintegrated, leaving the biggest cliffhanger in Marvel history and attracting more people to its sequel *Avengers: Endgame*. In addition, *Avengers: Endgame* assembles all superheroes in the Marvel Cinematic Universe, so the celebrity cast also attracts more people to the cinema. Most importantly, this movie marks the end of a decade-long Marvel era with the death of possibly their most popular superhero -- Iron Man, so it is the highlight of a decade-long game. Because of all these reasons, *Avengers: Endgame* was extremely popular worldwide and also in the Chinese market.

Aquaman (highlighted in green) is an outlier because its target value is larger when given a reasonable predictor value. This movie is oddly popular in China, even when most reviews are generally negative. After some research, it seems that apart from a cliche storyline, *Aquaman* is equipped with formulaic superhero movie settings, namely popular cast and epic CGI. The one

possible highlight of the movie is that the director James Wan is Asian with a Chinese name. Chinese moviegoers were excited to see one of our own in Hollywood.

The two unusual points can be disruptive to our regression model and many related statistics, but we cannot delete them. What we can do is to analyze the rest of the data without the two points and the regression model will not be applicable to them.

First, let's exclude *Avengers: Endgame* because it seems more extreme than *Aquaman* and probably has a bigger effect on the regression model and its statistics. There is a chance that *Aquaman* will no longer be unusual after excluding *Avengers: Endgame*. Here is the new scatter plot:



Aquaman (highlighted in green) still appears marginally unusual. Another superhero movie, *Avengers: Infinity War* (highlighted in blue), also appears unusual after excluding *Avengers: Endgame*. *Avengers: Infinity War* almost seems to be the new replacement of *Avengers: Endgame* in this new scatter plot, with a high budget and a high Chinese market gross. Here is the updated regression output:

Regression Equation

Chinese Gross (Mil$) = 4.1 + 0.522 Budget (Mil$)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	4.1	55.4	0.07	0.942
Budget (Mil$)	0.522	0.282	1.85	0.074

Model Summary

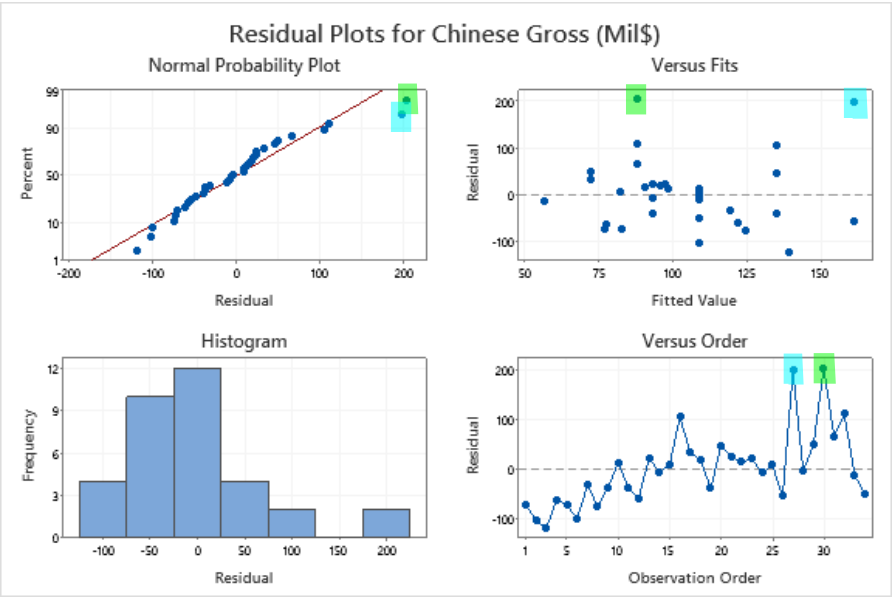
S	R-sq	R-sq(adj)
76.3686	9.65%	6.83%

Analysis of Variance

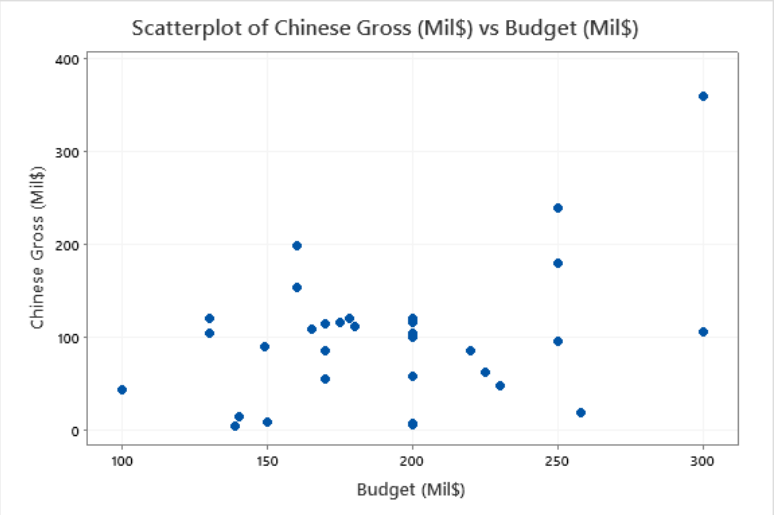
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	19941	19941	3.42	0.074
Error	32	186629	5832		
Total	33	206570			

The relationship becomes much weaker, with an R-squared of only 9.65%. *Avengers: Endgame*, as a leverage point and outlier, made the previous regression model look falsely stronger. The F-statistic is marginally significant. The intercept means that given that a superhero movie has a Budget of 0 million dollars, the estimated expected Chinese Gross is 4.1 million dollars, but the intercept is not very meaningful here because the budget has to be a positive number. The slope coefficient means that a 1 million dollar change in Budget is associated with an estimated expected 0.522 million dollar change in the Chinese market gross. The t-statistic for Budget marginally rejects the null hypothesis of $\beta_1 = 0$. Therefore, there is a marginal relationship between the changes in Budget and the difference in Chinese Gross. The t-statistic for Constant, which tests for the null hypothesis of $\beta_0 = 0$, is not statistically significant at all. Therefore, we fail to reject this null hypothesis.

As expected, *Aquaman* (highlighted in green) and *Avengers: Infinity War* (highlighted in blue) stand out as unusual in the "four in one plot":



Now, let's proceed with excluding *Aquaman* and fitting 33 observations to a simple regression model. The latest scatter plot is shown below:



Avengers: Infinity War still appears unusual. Here is the latest output:

Regression Equation

Chinese Gross (Mil$) = -19.2 + 0.611 Budget (Mil$)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	-19.2	49.9	-0.38	0.704
Budget (Mil$)	0.611	0.253	2.42	0.022

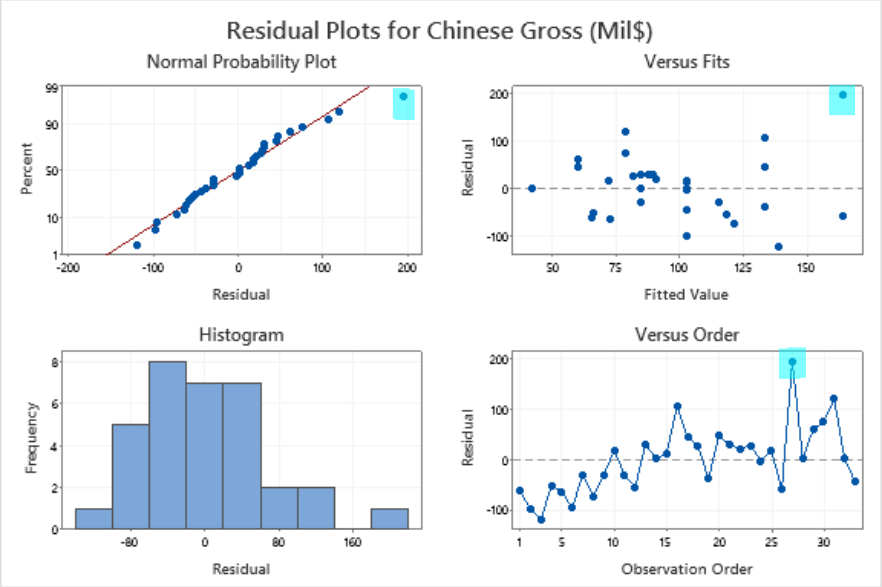
Model Summary

S	R-sq	R-sq(adj)
67.9463	15.85%	13.14%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	26965	26965	5.84	0.022
Error	31	143117	4617		
Total	32	170082			

The relationship becomes a little stronger than before, with an R-squared of 15.85%. The t-statistic for Budget rejects the null hypothesis of $\beta_1 = 0$. Therefore, there is a relationship between the changes in Budget and the difference in Chinese Gross. The t-statistic for Constant, which tests for the null hypothesis of $\beta_0 = 0$, is again not statistically significant at all. Therefore, we fail to reject this null hypothesis.

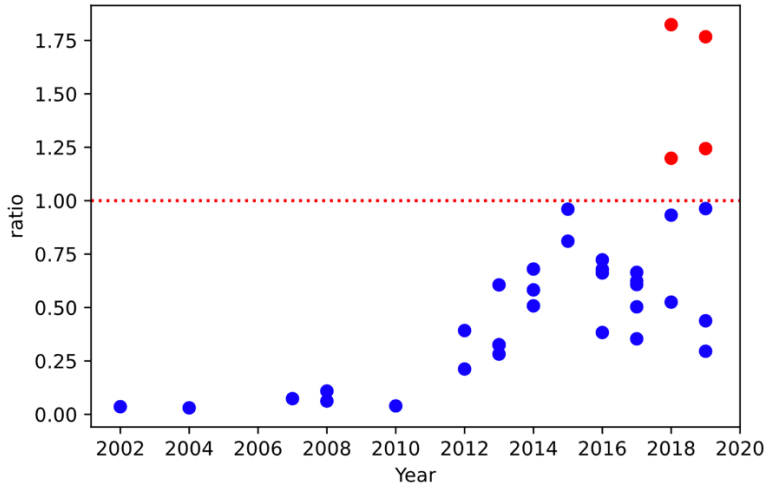


Again, *Avengers: Infinity War* (highlighted in blue) appears unusual in the "four in one plot". But I am not going to continue omitting it because omitting unusual points in analysis creates a vicious cycle.

In conclusion, there is at least a marginal relationship between the production budget of superhero movies and the total gross in the Chinese market, even though the regression models only account for less of one-third of the variability. The associated increase in Chinese market gross with the increase of 1 million dollars in superhero movie budget varies a lot depending on how *Avengers: Endgame* and *Aquaman* are handled. The simple regression model is oversimplifying the complicated relationship between superhero movie budgets and total gross in China. As we can see from the leverage point and outliers, storylines, plots, milestone significance, and even director ethnicity are all potentially relevant factors, perhaps a multiple regression will be a better model.

FURTHER ANALYSIS:

I calculated the ratio of Chinese Gross over Budget as an indicator of whether a superhero movie can profit by relying solely on the Chinese market. Then, I used Python to make a scatter plot and marked the points in different colors.



From the scatter plot of ratio versus Year, we can see that there are four observations above the y=1.0 horizontal line. That is to say, the four superhero movies, *Avengers: Infinity War, Aquaman, Avengers: Endgame,* and *Spider-Man: Far from Home*, if released only in the Chinese market, guarantee profitability regardless of how they perform in other parts of the world. This ties back to the very beginning of the report. China is too lucrative a market for superhero movies to fail the Chinese censorship. That is why studios like Marvel are already censoring their own superhero movies for the Chinese market. Marvel even went as far as co-producing *Iron Man 3* with a Chinese partner to make sure it is popular with the Chinese audience, although it did not seem to be the case judging from the data I collected.

Another trend we can observe is that the ratio of Chinese Gross over Budget is growing as the years pass by. Some points even exhibit an exponential-akin growing trend. With more young children growing up watching superhero movies in China, the trend will likely continue.