Analyzing Online Sentiment Towards Marvel TV Shows

Rohan Mehta

Disney+ has now released seven Marvel TV shows. The episodic format of these most recent additions to the MCU has enabled Marvel to pursue more experimental styles, evidenced by series such as WandaVision, Moon Knight, and Ms. Marvel. As such, it is not uncommon for these shows to be met with a broad range of opinions. However with the release of Ms. Marvel, it seems that these viewpoints have become more polarized than ever, eliciting widespread reports of review-bombing [1, 2, 3, 4, 5, 6, 7, 8, 9]. Are these claims true? Does the distribution of user ratings for Ms. Marvel differ significantly from those of other shows? Does it differ significantly from the distribution of ratings according to critics? And what information can we glean about why it received the ratings it did – from both critics and users – by studying the language that features most prominently in those reviews which rated it favorably compared to those which did not?

Downloading and cleaning the data

I surveyed many sites for reviewing movies and TV shows including IMDB, Rotten Tomatoes, and Metacritic. Of all of these Metacritic appeared to be the easiest to scrape, since it had already split up reviews from critics and users into separate pages, and had a consistent structure for dividing reviews. After manually recording the URLs for the user and critic reviews of the seven Marvel TV shows released so far, the data was then imported, cleaned, and formatted. Although “users” is a somewhat uncomfortable term to use in this setting, we will just take it to mean all people who are not critics.

Note that if you actually looked at the Metacritic pages for these shows you would see that they have supposedly been "scored" two or three hundred times each by users. However, even after scraping all available pages, we only get around half the purported number of user reviews. I'm not sure why this is the case (perhaps you can score a show without leaving a review?), but if anyone has an inkling as to the reason for this inconsistency, please be sure to let me know!

In[1]:=
```
marvelShows = {…};
```

In[2]:=
```
reviewLinks = {…};
```

In[3]:=
```
distTypes = {…};
```

In[4]:=
```
criticReviews = StringSplit[#, "Critic score Publication By date"] & /@
    Import /@ (# <> "/critic-reviews" & /@ reviewLinks);
```

In[5]:=
```
userReviews = Map[StringSplit[#, "User score By date Most helpful"] &, Map[Import,
    (Table[# <> "/user-reviews" <> x, {x, distTypes}] & /@ reviewLinks), {2}], {2}];
```

In[6]:=
```
allReviews =
  AssociationThread[marvelShows → MapThread[List, {criticReviews, userReviews}]];
```

In[7]:=
```
allReviews = {Drop[StringSplit[#[[1]][[2]], "Read full review"], -1], Join @@
      ((x ↦ Drop[x, -1]) /@ (x ↦ Last@StringSplit[x, "All this user's reviews"]) /@
         #[[2]])} & /@ allReviews;
```

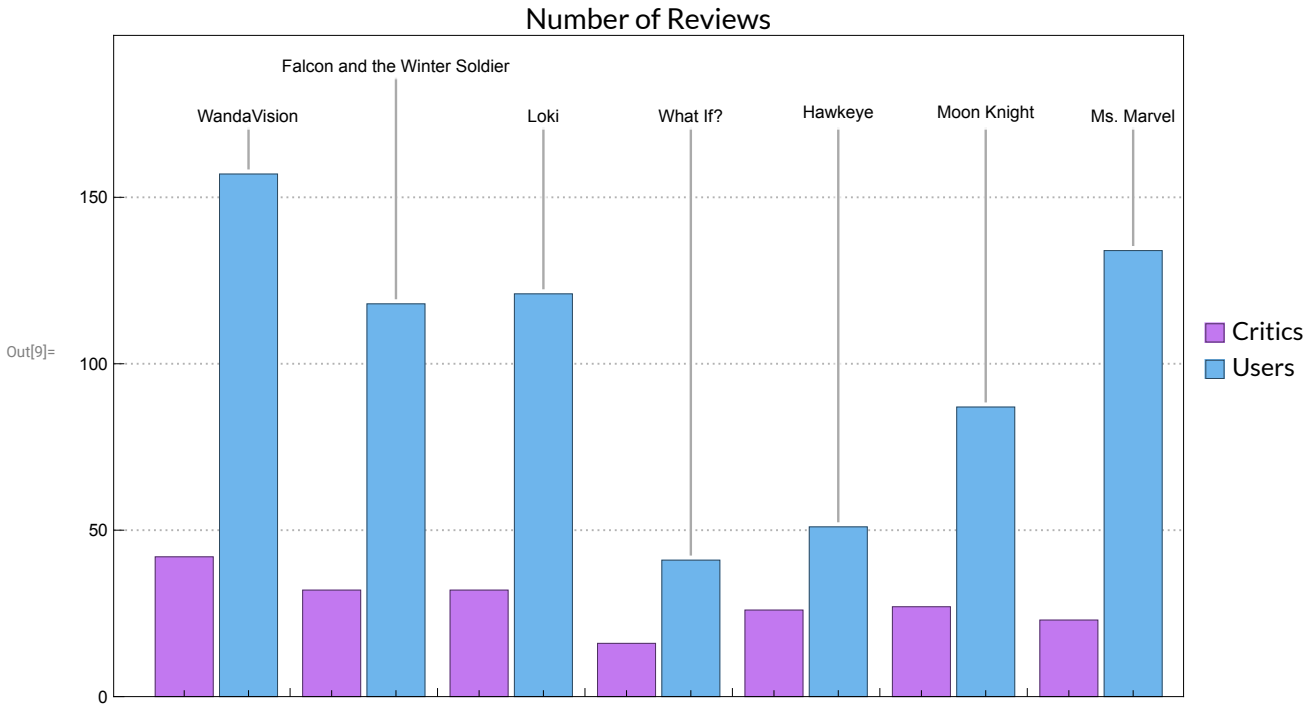
How many reviews did each show receive?

One of the easiest metrics to compute is how many reviews exist for each show. As may be expected, the oldest shows (all released early last year) tend to have the greatest number of user reviews (~120), since they have had the most time to accrue users. *WandaVision*, the first ever Marvel TV release, has the most user reviews at almost 160. Older shows also have a slightly greater number of critic reviews, though this difference is not nearly as great as the former (which also makes sense, as critics review all shows almost immediately, since it is their job).

In[8]:=
```
reviewCount = Map[Length, allReviews, {2}];
```

In[9]:=
```
BarChart[reviewCount, ···]
```

Out[9]=



We can also determine what percentage of each show's total reviews came from users versus critics. This metric does not turn out to be particularly interesting, as the breakdown is quite similar for all shows. However, we may consider that it gives us a rough heuristic for how strong an experience the show evoked in users. All users who posted their review for a given show had some motivation to do so,

likely due to a strong reaction they had while watching the show. So the show whose percentage of reviews due to users is the greatest could be interpreted as the most provocative.

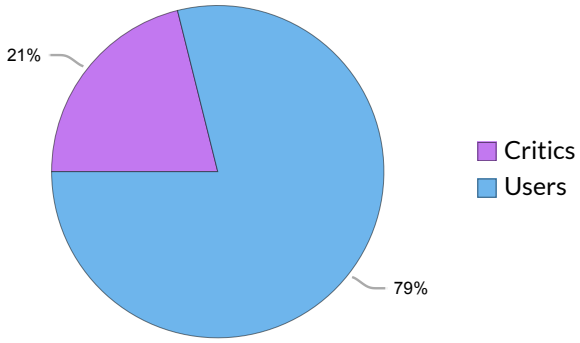
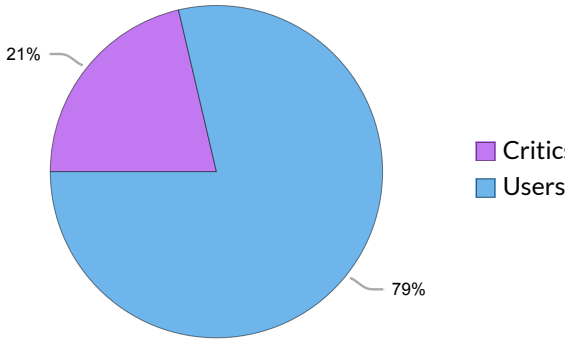
With 85% of its total reviews coming for users, *Ms. Marvel* technically earns this title, although *WandaVision*, *Loki*, and *Falcon and the Winter Soldier* are right on its heels at 79% and nearly every other show is only a few percentage points removed (except Hawkeye, which dips below 70% for this metric). All in all, this difference isn't large enough for us to make any useful distinctions save between *Hawkeye* and the other TV shows.

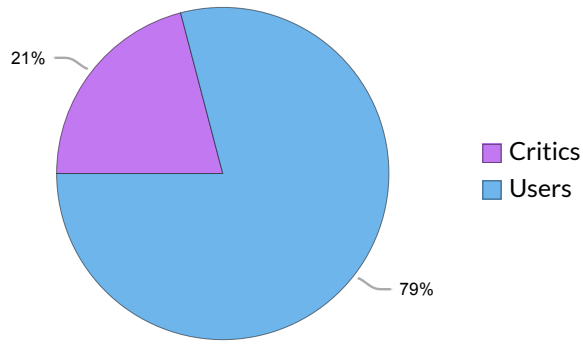
In[10]:=
```
reviewBreakdown[reviewCount_] := Row@{Spacer[50], PieChart[reviewCount, ...]}
```

In[11]:=
```
{titles, charts} =
  {Function[...] /@ Keys@reviewCount, reviewBreakdown /@ Values@reviewCount};
```

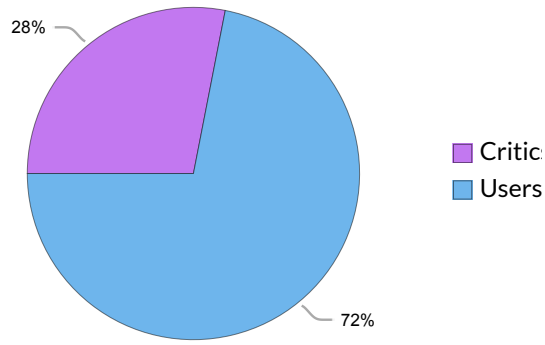
In[12]:=
```
Grid[Flatten[Transpose[Partition[#, UpTo[2]] & /@ {titles, charts}], 1], ...]
```

Out[12]=

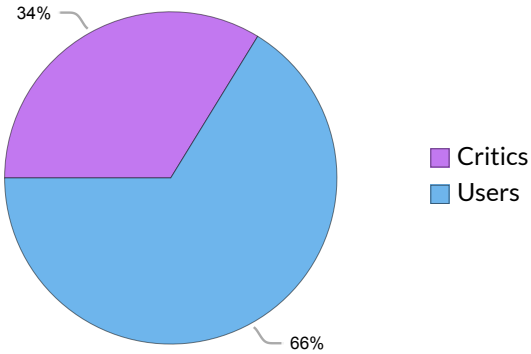
WandaVision	Falcon and the Winter Soldier
	
Loki	What If?



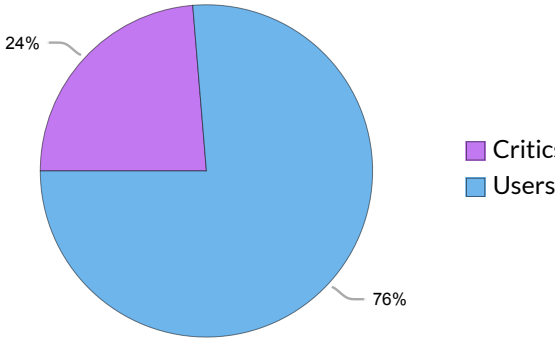
Hawkeye

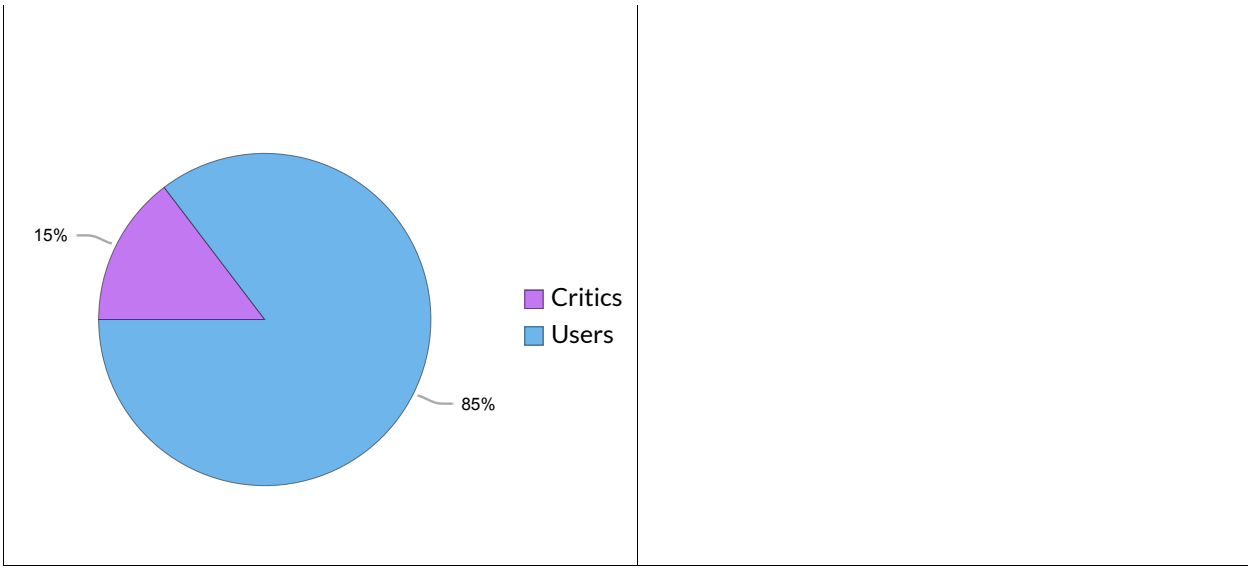


Moon Knight



Ms. Marvel





Finally, by looking up the release dates for all seven shows, we can also plot the quantity of user and critic reviews for each show in chronological order. This corroborates the trend we saw earlier in the bar graph, although it does reveal much more clearly that the number of critic reviews peaked for *WandaVision*, Marvel's first TV show release, and then declined going forward. It also demonstrates that the user interest generated by *Ms. Marvel* – as measured by the number of user reviews – seems to have returned to and even surpassed where it was for Marvel's second and third TV releases, after suffering a steep decline with *What If?*

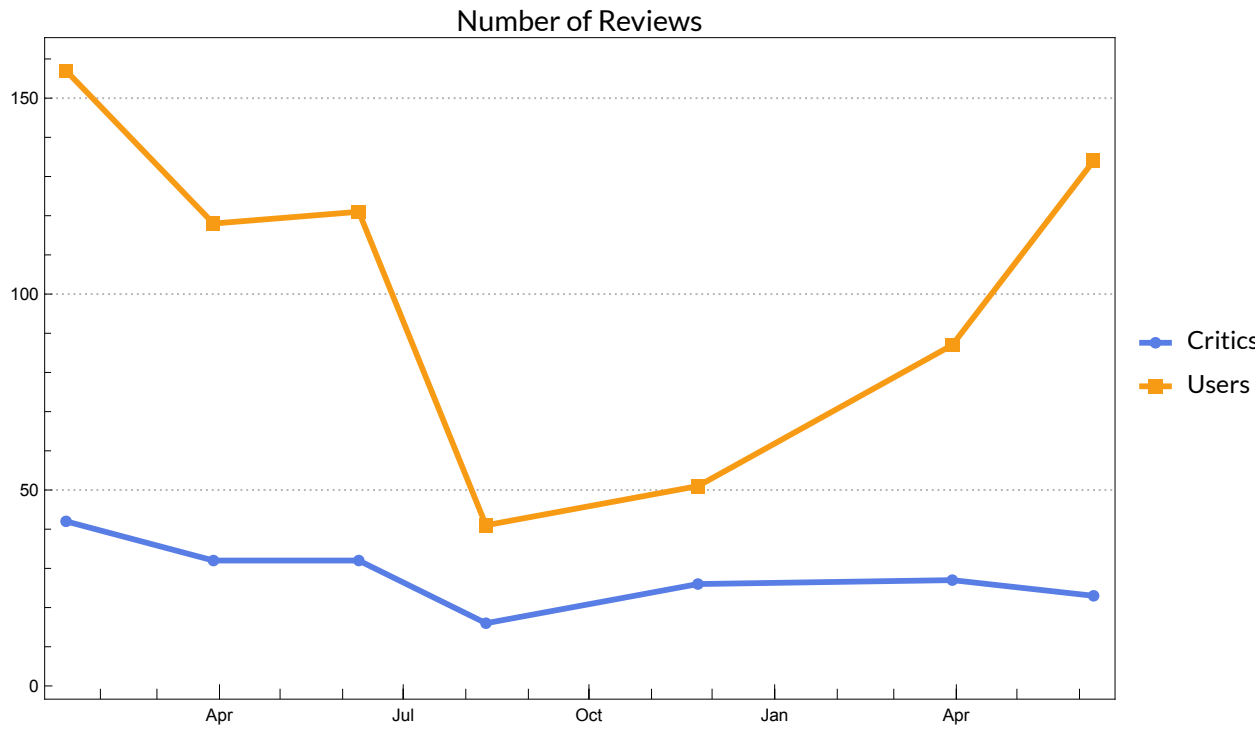
In[13]:=
```
releaseDates = <| ··· |>;
```

In[14]:=
```
userCount = {releaseDates[#], reviewCount[#][[1]]} & /@ Keys@reviewCount;
```

In[15]:=
```
criticCount = {releaseDates[#], reviewCount[#][[2]]} & /@ Keys@reviewCount;
```

In[16]:=
```
DateListPlot[{userCount, criticCount}, ···]
```

Out[16]=



What was the score distribution for both critics and users?

Now we can actually look at the scores for each review! To do this, we need to extract all the numbers that appear in a given review, and then take the first one (since each review always begins with its score). It should be noted that critics' scores range from 0-100 while users' scores range from 0-10. When we visualize the scores from critics, we see that *Ms. Marvel* is rated the highest out of all Marvel shows. This is in accordance with reports that it is the highest rated Marvel motion picture (i.e., movie or TV show) ever released [2, 3, 10]. *WandaVision* takes second place at just a tenth of a percentage point below *Ms. Marvel*, and *Loki* takes third place clocking in at a few percentage points below that. As stated before, *Hawkeye* and *What If?* rank last (in that order), in the high 60s. Overall though, everything is pretty close here, with the top five shows all within ten percentage points of each other.

In[17]:=
```
extractScore[text_] := StringCases[text, "\n" ~~ x : DigitCharacter .. :> x];
```

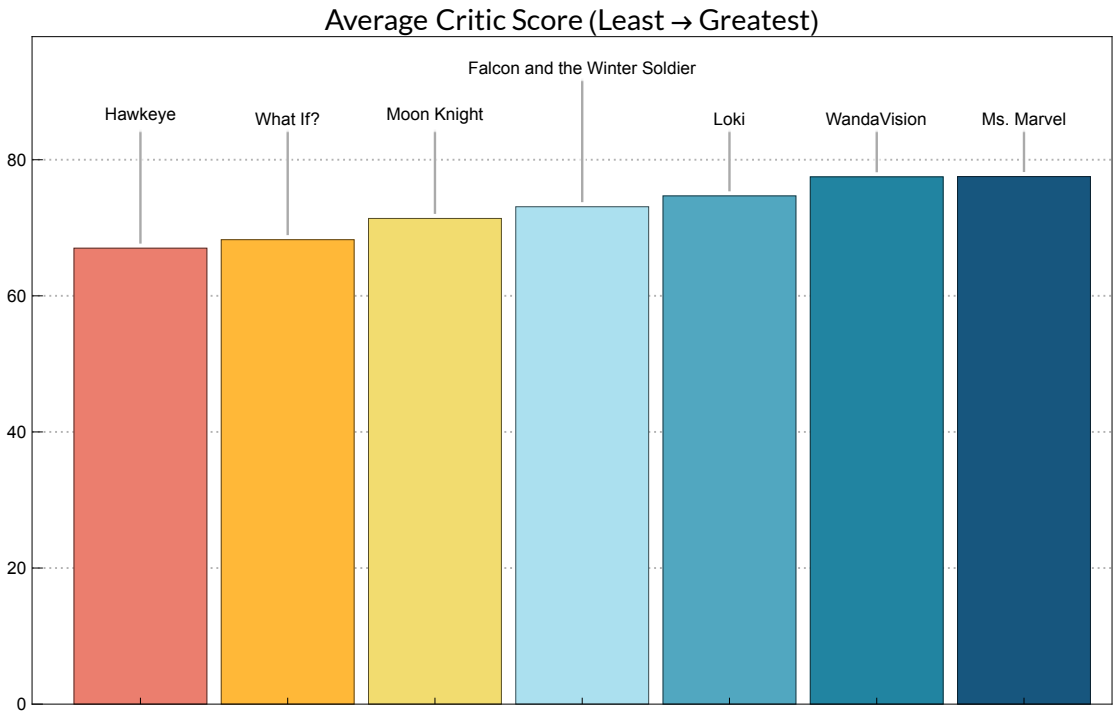
In[18]:=
```
scores = AssociationThread[{"Critics", "Users"} → #] & /@
   Map[ToExpression@*Flatten, Map[extractScore, allReviews, {3}], {2}];
```

In[19]:=
```
scores = <|"Critics" → Select[#[[1]], LessEqualThan[100]],
     "Users" → Select[#[[2]], LessEqualThan[10]]|> & /@ scores;
```

In[20]:=
```
avgScores = {N@Mean@#["Critics"], N@Mean@ #["Users"]} & /@ scores;
```

In[21]:=
```
BarChart[#[[1]] & /@ SortBy[avgScores, First], ···]
```

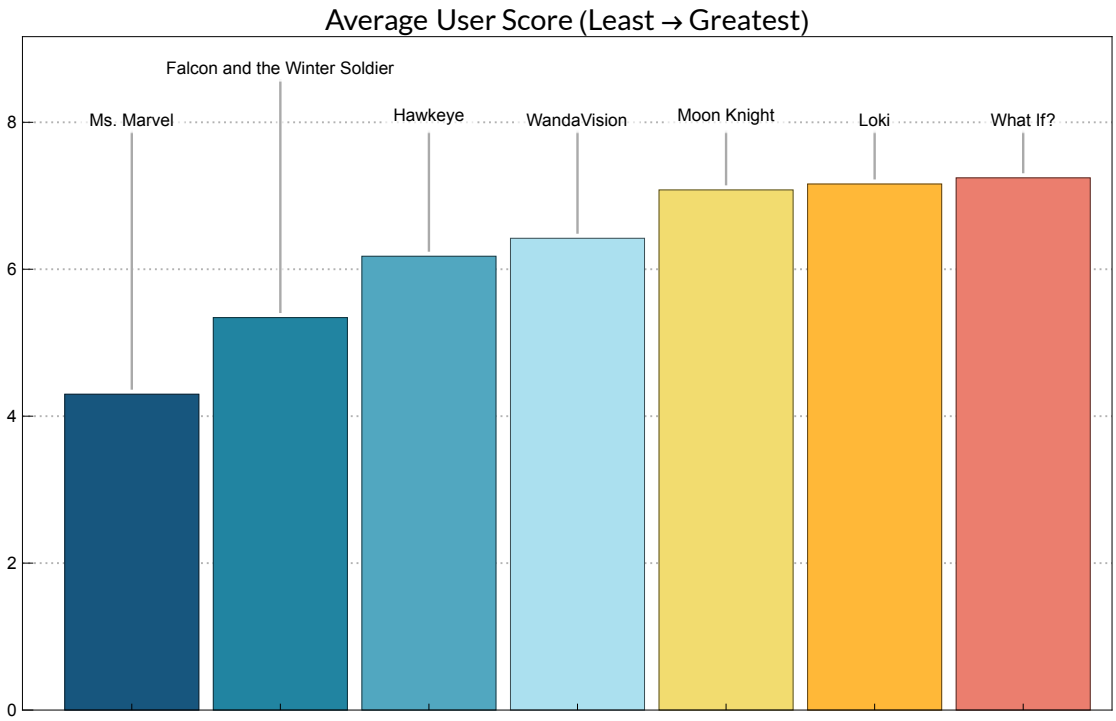
Out[21]=



However, when we look at user ratings *Ms. Marvel* is neither first, second, or third! In fact, it is dead last (and funnily enough, has dropped even further during my most recent web scrape). Before we try and investigate *why* this might be the case, you have to admit, it's pretty surprising. Usually we expect critics to be more particular and less forgiving than users, as fans place less of a premium on the artistry of a motion picture as long as it keeps them engaged. In fact, we can see how this works in the favor of shows like *What If?* which moved all the way up to first place (from second to last) when comparing user score versus critic score. *Ms. Marvel* appears to be suffering from the reverse effect, where fans are judging it more harshly than critics – punishingly harsh, I would say. So we do have some initial evidence that review-bombing could be happening.

In[22]:=
```
BarChart[#[[2]] & /@ SortBy[avgScores, Last], ...]
```

Out[22]=



If we actually plot the distribution of each show's user and critic scores as a histogram (and try to reconstruct a PDF from which this distribution might be generated, which we'll call a "smooth histogram"), this discrepancy becomes even more clear. First, let's note that the distribution of user scores for *Ms. Marvel* is extremely bimodal – more so than that of any other show – with 24 users giving it a perfect score, while 40 rank it somewhere in the 0-2 range. So it seems like our hypothesis that *Ms. Marvel* is the most polarizing Marvel TV show released to date is in fact correct. Some users really love the show, but for some reason, more hate it. This is much more convincing evidence of review-bombing. A good portion of users agree with critics' favorable estimation of the show, but another group is completely bucking this trend and reacting in the exact opposite way. Plus, the fact that this group is around twice the size of the group generating positive reviews is also notable. Review-bombers will leave as many reviews as they possibly can since their goal is making the show look bad.

Taking a look at the smooth histograms clears some things up as well. For nearly every other show, the major peaks of these curves more-or-less match up between the user and critic score distributions (*Falcon and the Winter Soldier* is somewhat of an exception), indicating that on average critics and users basically agree on how good the show was overall, even if their specific scores differ on a case-by-case basis. On the other hand, the smooth histogram for *Ms. Marvel* as generated by users' scores doesn't match up with the one generated by critics' scores at all. Its critic distribution is heavily centered around 80%, whereas its user distribution is almost centered at 0%! Moreover, whereas the critic distributions of most other shows observe a slow, sloping descent towards zero after peaking (indicating that a good portion of reviewers gave it a medium-ish score), *Ms. Marvel's* critic distribution exhibits an almost vertical, 90-degree drop (meaning that reviews with scores far below its peak at 80% are

essentially nonexistent). This provides some more quantitative backing to our claim that it really is the best reviewed Marvel motion picture ever.

Note that the smooth histograms often stray below 0, or above 10 or 100. This is because they are doing their best to estimate the underlying PDF for the data we're providing them with, and have no idea that reviews are arbitrarily capped at certain numbers. That said, they are still very useful for visualizing the overall shape of each show's score distributions.

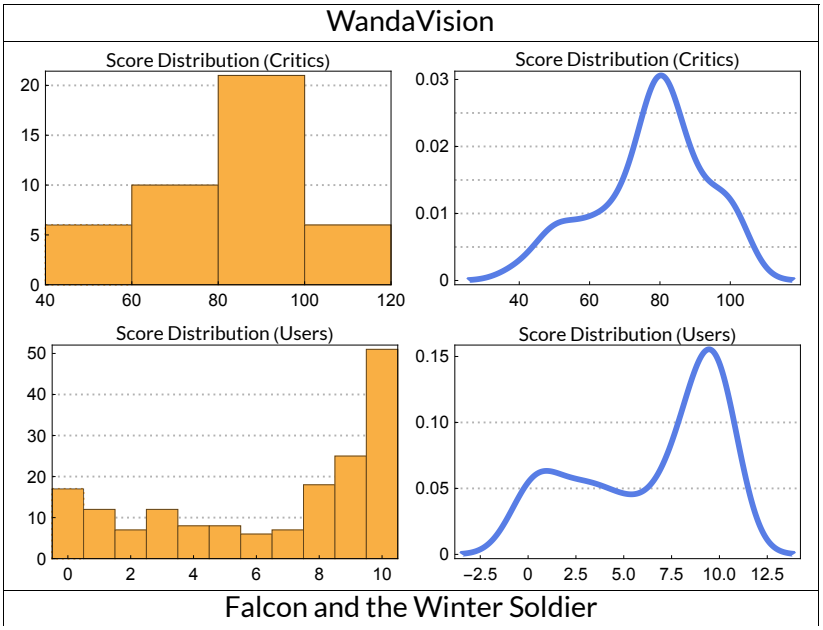
In[23]:=
```
getPlots[plotFn_] := Module[{newScores, newAssoc},
  newScores = AssociationThread[({Style[...] + } & /@ {"Critics", "Users"}) → #] & /@
    (KeyValueMap[(Append[#2, Style["Score Distribution (" <> ToString@#1 <> ")",
          Black, FontFamily → "Lato", FontSize → 10]] &), #] & /@ scores);
  Values /@ Map[plotFn[Drop[#, -1], ... + ] &, newScores, {2}]]
```

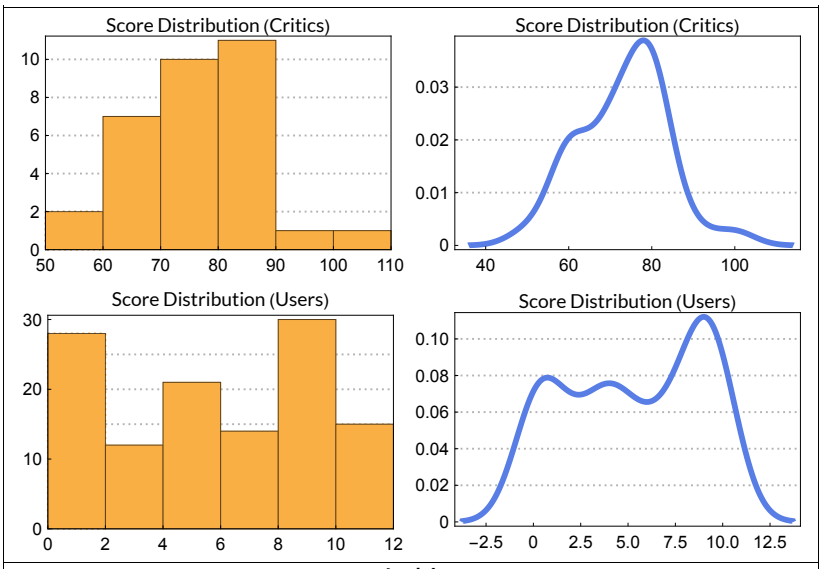
In[24]:=
```
{hard, smooth} = getPlots /@ {Histogram, SmoothHistogram};
```

In[25]:=
```
combined = Grid[{{#[[1]][[1]], #[[2]][[1]]}, {#[[1]][[2]], #[[2]][[2]]}}] & /@
   Merge[{hard, smooth}, Identity];
```

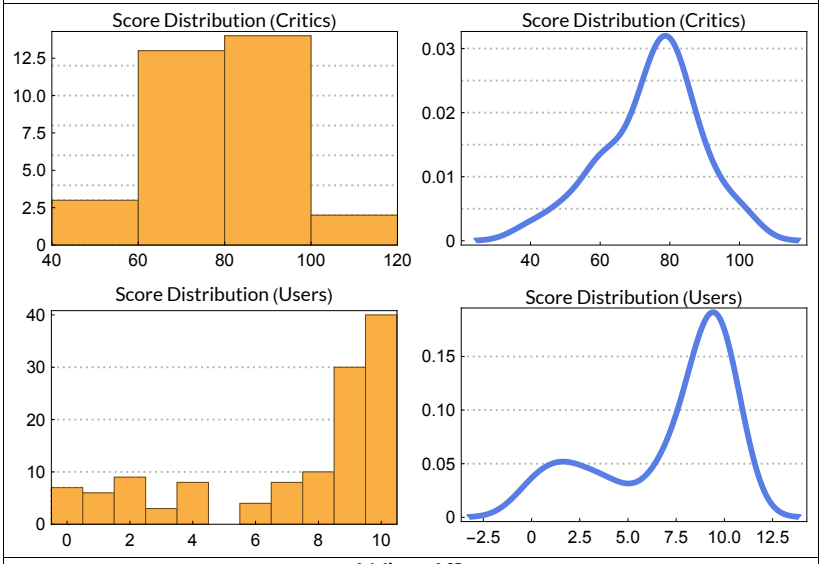
In[26]:=
```
Grid[Flatten[MapThread[{{#2}, {#1}} &, {Values@combined,
    (Style[#, Black, FontFamily → "Lato", FontSize → 15] & /@ Keys@combined)}], 1],
 Frame → All, FrameStyle → Directive[Black, Thin]]
```

Out[26]=

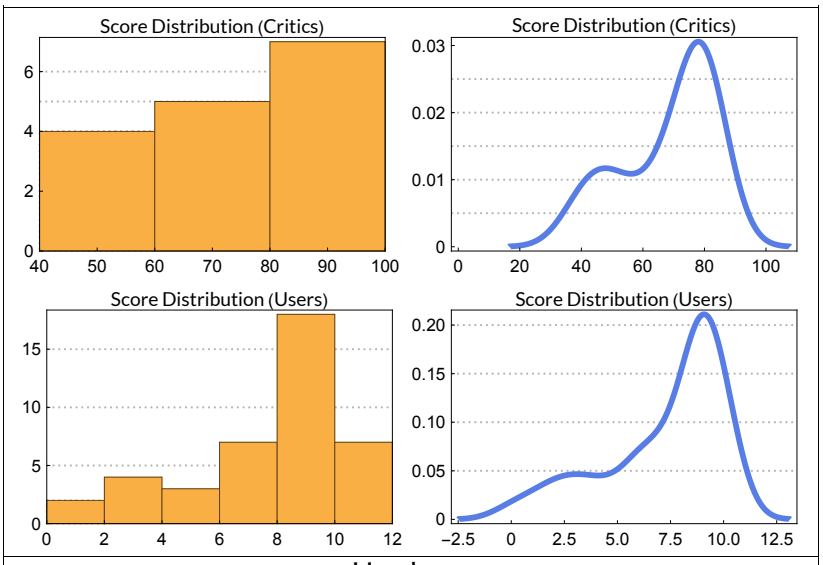



Score Distribution (Critics)
Score Distribution (Critics)
Score Distribution (Users)
Score Distribution (Users)

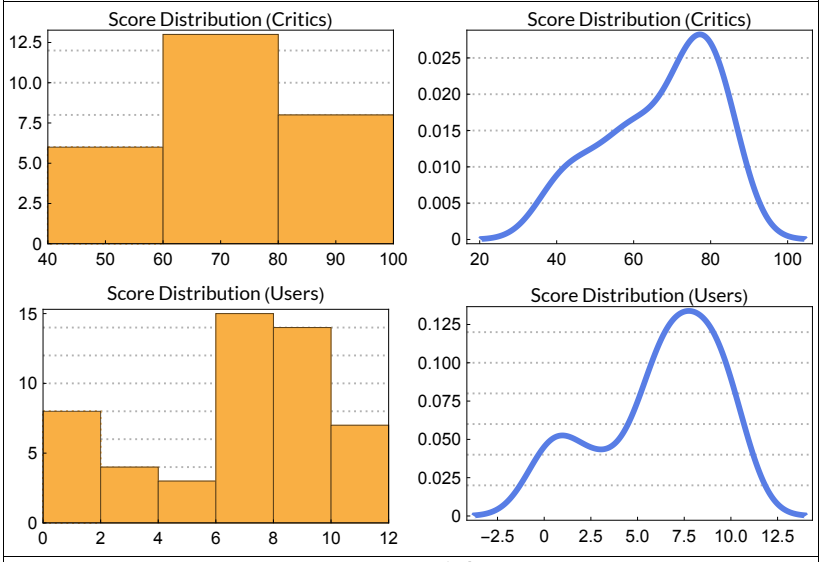
Loki


Score Distribution (Critics)
Score Distribution (Critics)
Score Distribution (Users)
Score Distribution (Users)

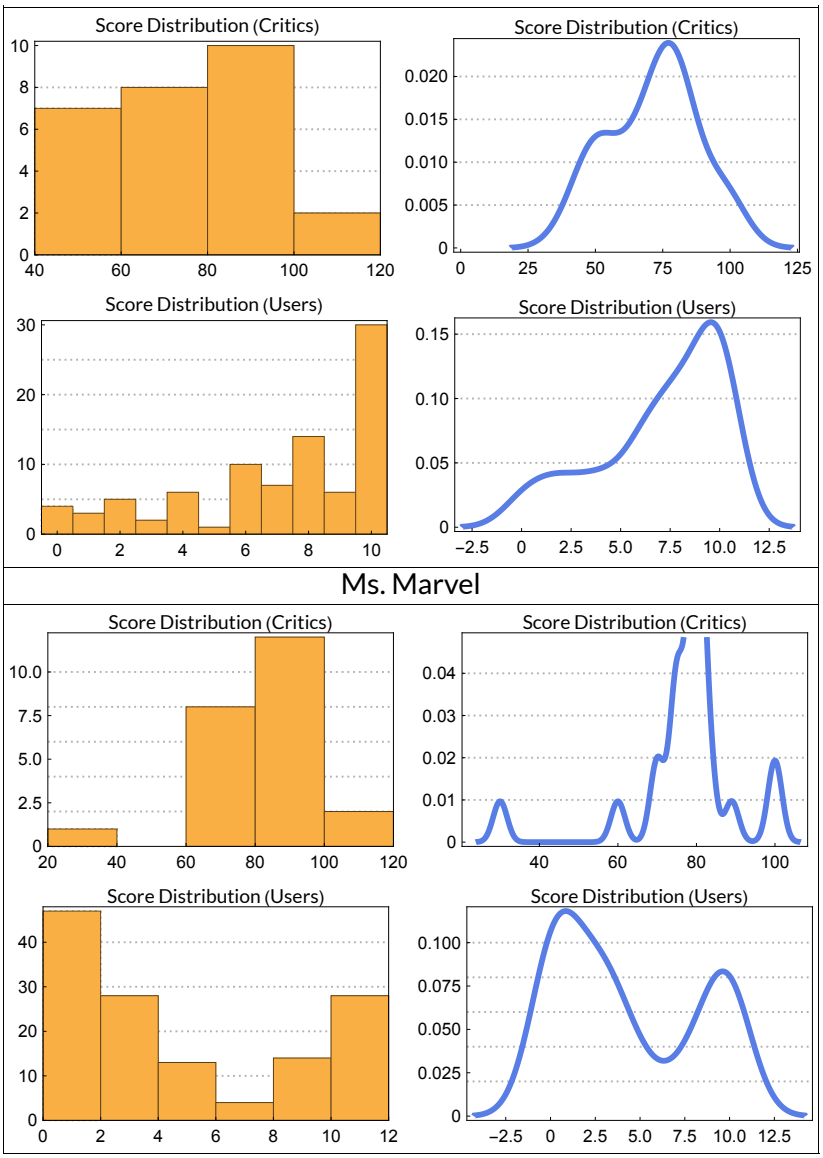
What If?



Hawkeye



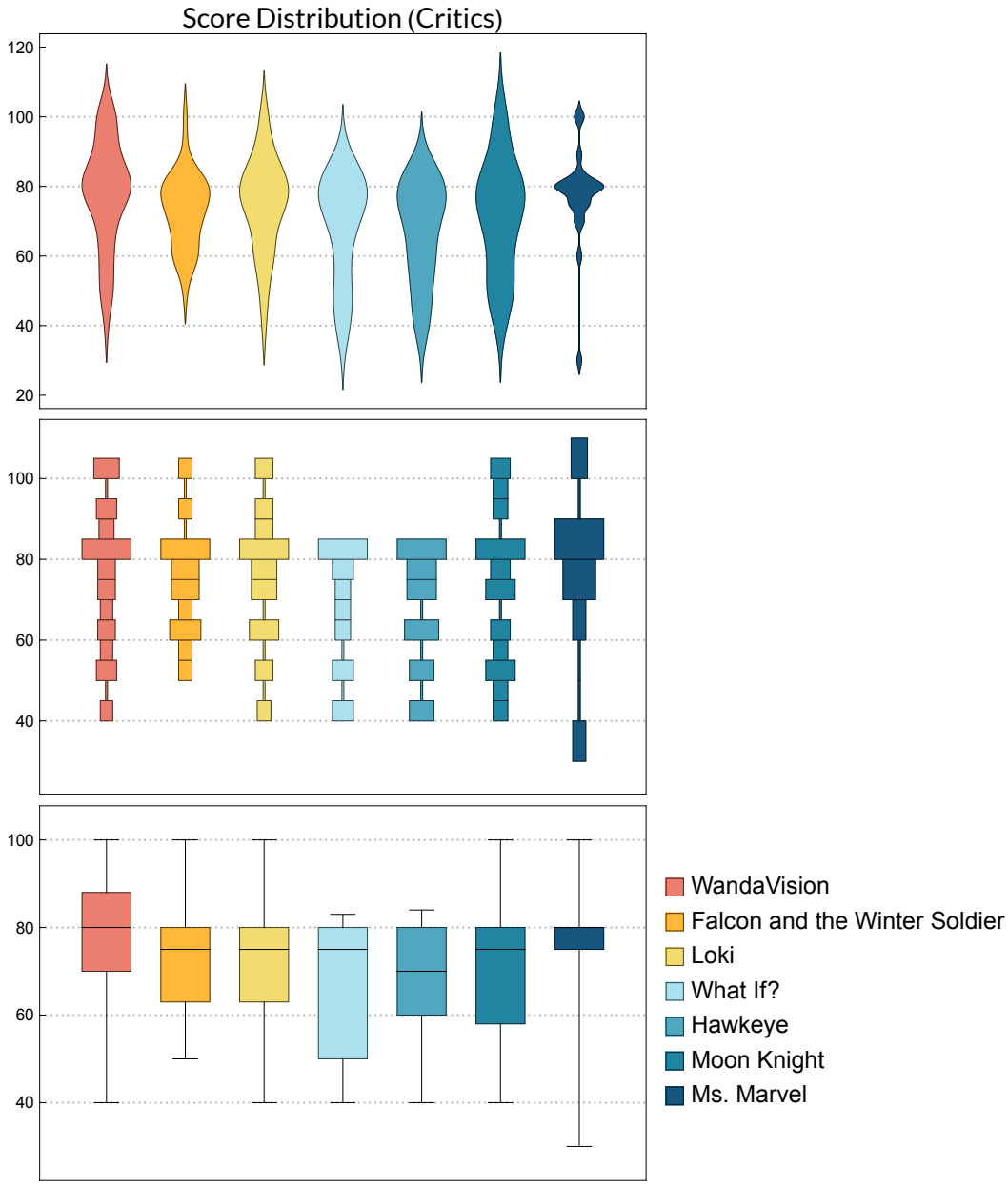
Moon Knight



We can also look at density plots of each show's score distributions. Starting off with the critic distribution, notice just how unusual *Ms. Marvel* looks. Like we said before, it has essentially no reviews below its peak, corresponding to an extremely, instantly thin tail (as opposed to the fatter, gradually thinning tails of the other shows' distributions). In terms of actual median score, it ties *WandaVision* for first place at 80%. If we re-render the plot by boxing together scores in a similar range, we see that *Ms. Marvel* is being weighed down by a single, extremely low critic score, whereas *WandaVision* is suffering the net effect of several, more mildly negative reviews. Rendering as a box-and-whisker chart reveals that the score weighing *Ms. Marvel* down – one that is more than 30 percentage points lower than its next lowest score – is a 30%, the lowest score given to any show by any critic. In fact, we can calculate that it (and every other "low" critic score *Ms. Marvel* receives) is a statistical outlier by a longshot (since they are below Q1 - 1.5 x IQR).

In[27]:=
```
Column[Table[DistributionChart[Values[(#1[[1]] &) /@ scores], ···],
  {style, {Automatic, "HistogramDensity", "BoxWhisker"}}]]
```

Out[27]=

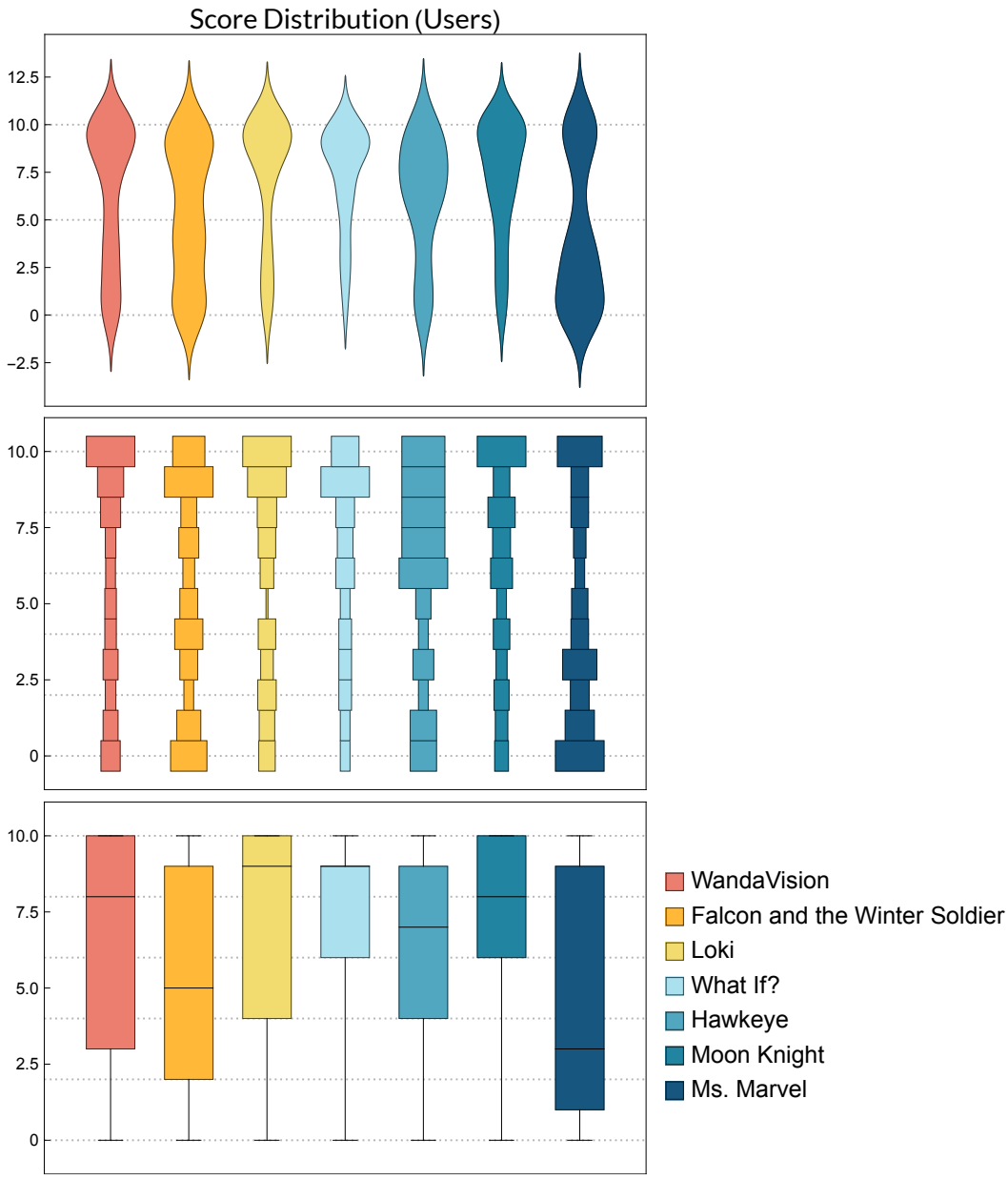


The density plots for the user distributions mostly approximate their corresponding critic distributions (just as we saw with the smooth histograms), save for *Ms. Marvel*. In fact, as we noted earlier, most shows actually fare slightly better under users. We can determine this since the "bulges" of their density plots (where most of the scores are concentrated) move a little upwards when comparing the user versus critic distributions. *Ms. Marvel's* "bulge", on the other hand, suffers a steep fall, earning it a dismal median score of 30%. By now, we can confidently say that the distribution of user scores for Ms. Marvel is being subjected to a tug-of-war between two extreme opinions, with the negative side mostly

winning out. We should also note that the user distribution for *Falcon and the Winter Soldier* mirrors many of the same trends, although to a far lesser extent.

In[28]:=
```
Column[Table[DistributionChart[Values@(#[[2]] & /@ scores), ···],
  {style, {Automatic, "HistogramDensity", "BoxWhisker"}}]]
```

Out[28]=



On the topic of polarization, here's another interesting question: How many perfect scores (100/100 or 10/10) and failing scores (0/100 or 0/10) did each show receive? *WandaVision* crushes all competition with a total of 51 perfect scores assigned by users and 6 assigned by critics, the highest of any show. *Loki*, *Moon Knight*, and *Ms. Marvel* all follow up with 2 perfect scores assigned by critics, and 40, 30, and 28 perfect scores assigned by users, respectively.

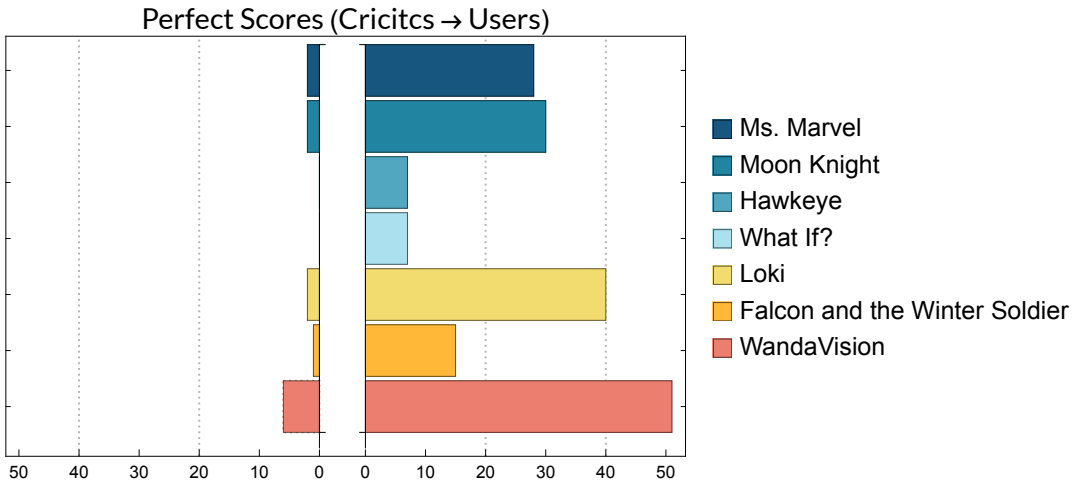
In[29]:=
```
tabulatedScores = Map[{Counts@#[["Critics"]], Counts@#[["Users"]]} &, scores, {1}];
```

In[30]:=
```
perfect = # /. Missing[__] → 0 & /@ {#[[1]][100], #[[2]][10]} & /@ tabulatedScores;
```

In[31]:=
```
fails = # /. Missing[__] → 0 & /@ Map[Lookup[0], tabulatedScores, {2}];
```

In[32]:=
```
PairedBarChart[First /@ perfect, Last /@ perfect, ···]
```

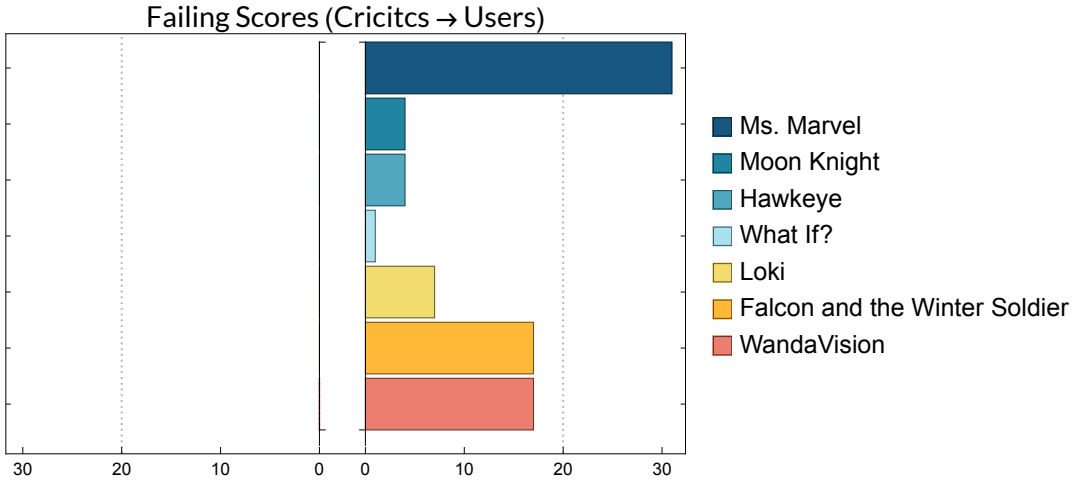
Out[32]=



Now let's take a look at failing scores. First, we should note that critics never gave a single show a 0. This is probably due to some combination of the greater control they can exercise when assigning a score (100 options versus 10) as well as the fact that this is their profession, and so they are less likely to be impulsive and biased in their ratings. Users, on the other hand, had no such qualms. Every show received at least one failing score. *Ms. Marvel* leads the pack with a whopping 31, almost twice the next highest of 17, held by *Falcon and the Winter Soldier* and *WandaVision*. And even this is a bit extreme, compared to the 1, 4, or 7 failing scores assigned to all the other shows.

In[33]:=
```
PairedBarChart[First /@ fails, Last /@ fails, ···]
```

Out[33]=



Ok, I know I'm belaboring the point by now, but I want to explore one final question. How did the amount of positive, negative, and neutral reviews for each show change over time, after each show's

release? In accordance with Metacritic's classification scheme, we define positive reviews as those with a score greater than or equal to 7 or 70, negative reviews as those with a score less than or equal to 3 or 30, and neutral reviews as everything else. Since all critic reviews are essentially released at the same time, we're not likely to get anything interesting out of examining them in this context. So we'll limit the scope to only user reviews for the moment.

When visualizing the data in this way, some interesting trends begin to emerge. Every show except *Ms. Marvel* starts out being led almost entirely by positive reviews, with the relative proportions of negative and neutral reviews growing over time. That said, the neutral review category is the least populated for every show, and as expected, is most absent in *Ms. Marvel*, taking up only ~7% of all user reviews. Most shows take anywhere from 4-6 months for the percent composition of their user reviews to settle down around definite values in a sort of "equilibrium" state. Interestingly, *Ms. Marvel* has comparatively few bumps or perturbations in the graph for its percent composition over time, and seems to have reached equilibrium very quickly. While *WandaVision* exhibits this same phenomenon, it quickly settled into a state dominated by positive reviews, as opposed to *Ms. Marvel's* state of being dominated by negative reviews. Given how quickly and smoothly this came about, and the diametrically opposed opinions of many critics and others users, this does suggest a level of coordination and intent among certain users that we would expect to see if review-bombing was happening.

By plotting the individual data points we can also see the density of reviews in each category over time. As may be expected, most of the oldest shows – all older than a year – have had only one or two reviews in the last 4-6 months. Surprisingly *Loki*, Marvel's third TV release, has actually been reviewed almost five times within the past two months, making somewhat of a comeback. We can also draw a dotted vertical line when each show's last episode was released to make sense of this data a bit more clearly. Many shows seem to reach a period of stagnation very soon after their last episode was released, before becoming active a little while later, and then settling into a longer, more permanent period of stagnation. Not all shows immediately stagnate after their last episode airs though, *WandaVision*, *Loki*, *Hawkeye*, and *Ms. Marvel* among them.

In[34]:=
```
months = {…};
```

In[35]:=
```
lastEpisode = <|…|>;
```

In[36]:=
```
scoresToReview = Map[Merge[#, Identity] &,
   Map[Association[First@*extractScore@# → #] &, allReviews, {3}], {2}];
```

In[37]:=
```
scoresToDate = Map[(StringJoin[Take[StringSplit[#, " "], -3], " "] // DateObject) &,
   Map[Select[#, (Not@*SameAs[Nothing])] &,
    Map[# /. {EndOfFile → Nothing, {} → Nothing, {EndOfFile} → Nothing} &,
     Map[Find[StringToStream@#, months] &, scoresToReview, {4}], {3}], {2}], {4}];
```

In[38]:=
```
getGoodMedBad[scores_] := Module[{intermediate, dummy, counted},
  intermediate = If[ToExpression@# ≥ 7,
      "Good", If[ToExpression@# ≥ 4, "Med", "Bad"]] & /@ scores;
  dummy = <|"Good" → 0, "Med" → 0, "Bad" → 0|>;
  counted = Merge[{Counts@intermediate, dummy}, Total];
  KeySortBy[counted, # /. {"Good" → 1, "Med" → 2, "Bad" → 3} &]]
```

In[39]:=
```
datesToType = (KeySortBy[AbsoluteTime] /@
    Map[getGoodMedBad, GroupBy[#, Last → First] & /@ (Flatten[#, 1] & /@
        KeyValueMap[Table[{#1, x}, {x, #2}] &] /@ (Last /@ scoresToDate)), {2}]);
```

In[40]:=
```
scoresOverTime =
  Drop[#, 1] & /@ (FoldList[{x, y} ↦ {Merge[{x⟦1⟧, y}, Total], (Total@y) + x⟦2⟧},
      {<|"Good" → 0, "Med" → 0, "Bad" → 0|>, 0} , #] & /@ Values /@ datesToType);
```

In[41]:=
```
scoresOverTime = Transpose /@
   AssociationThread[marvelShows → Table[{{Values[Keys /@ datesToType]⟦x⟧⟦y⟧, #},
       {Values[Keys /@ datesToType]⟦x⟧⟦y⟧, # / scoresOverTime⟦x⟧⟦y⟧⟦2⟧}} & /@
     Values[scoresOverTime⟦x⟧⟦y⟧⟦1⟧], {x, Length@marvelShows},
    {y, Length@scoresOverTime⟦x⟧}]];
```

In[43]:=
```
firstVisuals = AssociationThread[
   marvelShows → KeyValueMap[Show[(StackedDateListPlot[#2, ⋯])],
     DateListPlot[{{Callout[{lastEpisode⟦#1⟧, Ceiling[
           Max[Last /@ Flatten[#2, 1]] / 10]}, "Last Episode Released"]}},
      ⋯]] &]@(Map[First, scoresOverTime, {3}])];
```

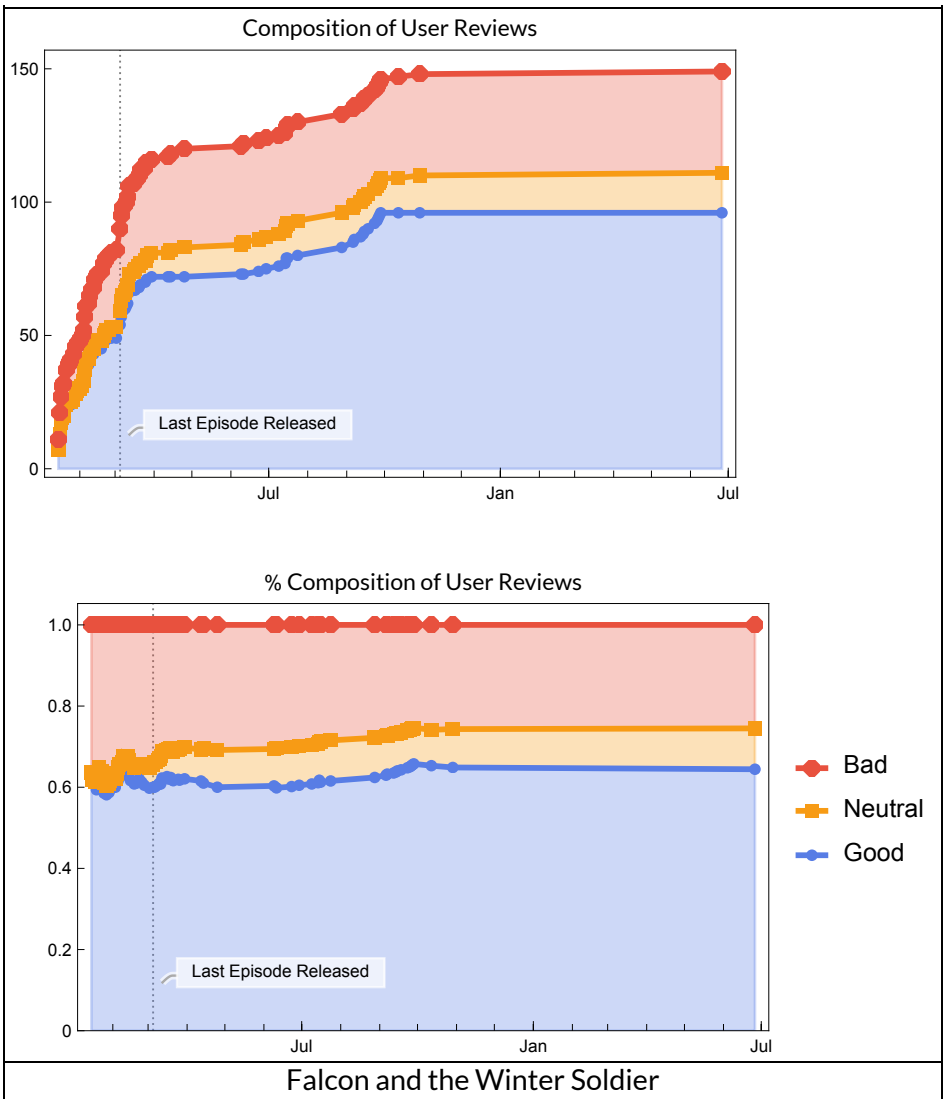
In[44]:=
```
lastVisuals = AssociationThread[
   marvelShows → KeyValueMap[Show[(StackedDateListPlot[#2, ⋯]), DateListPlot[
       {{Callout[{lastEpisode⟦#1⟧, 0.1}, "Last Episode Released"]}}, ⋯]] &]@
    (Map[Last, scoresOverTime, {3}])];
```

In[45]:=
```
visuals = Merge[{firstVisuals, lastVisuals}, Identity];
```

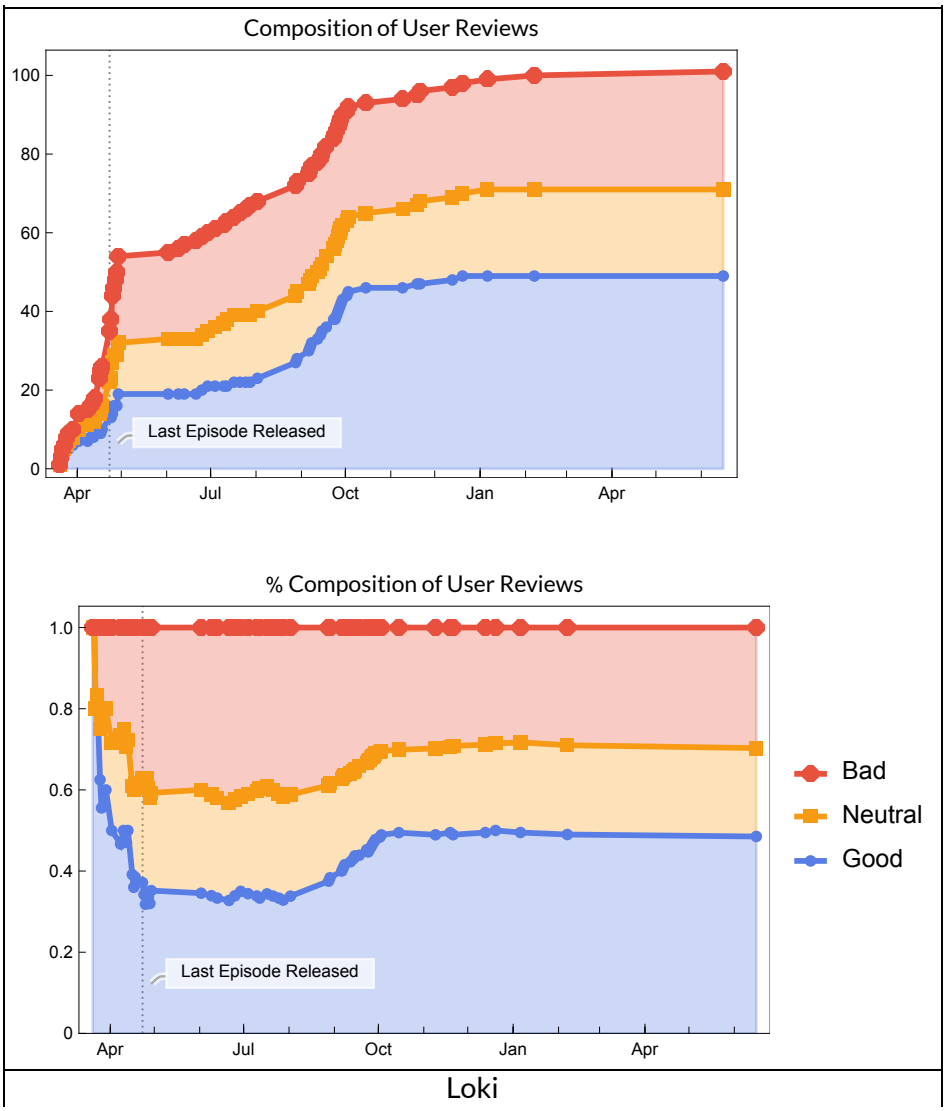
In[46]:=
```
Grid[Flatten[{{#⟦1⟧}, {Row[#⟦2⟧, Spacer[5]]}} & /@
   Transpose@{Style[⋯] & /@ Keys@visuals, Values@visuals}, 1], Frame → All]
```

Out[46]=

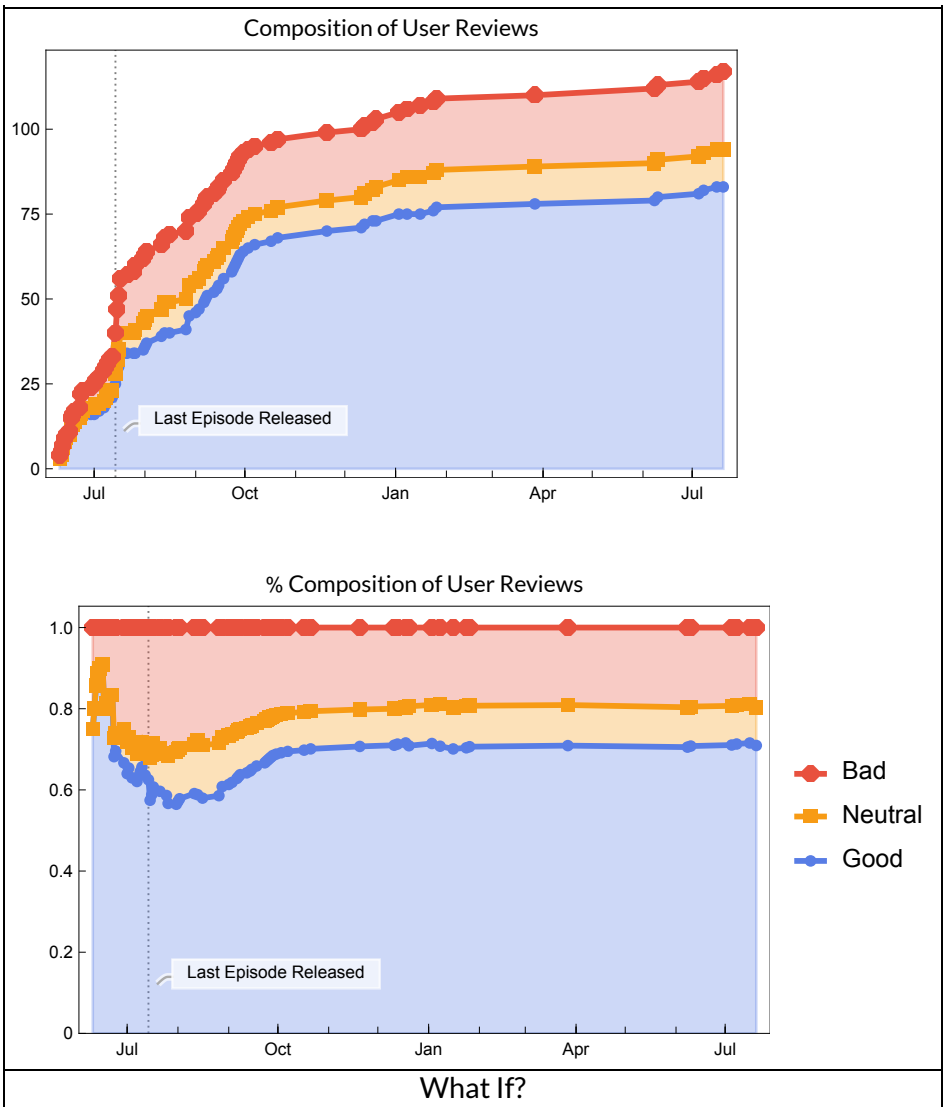
WandaVision

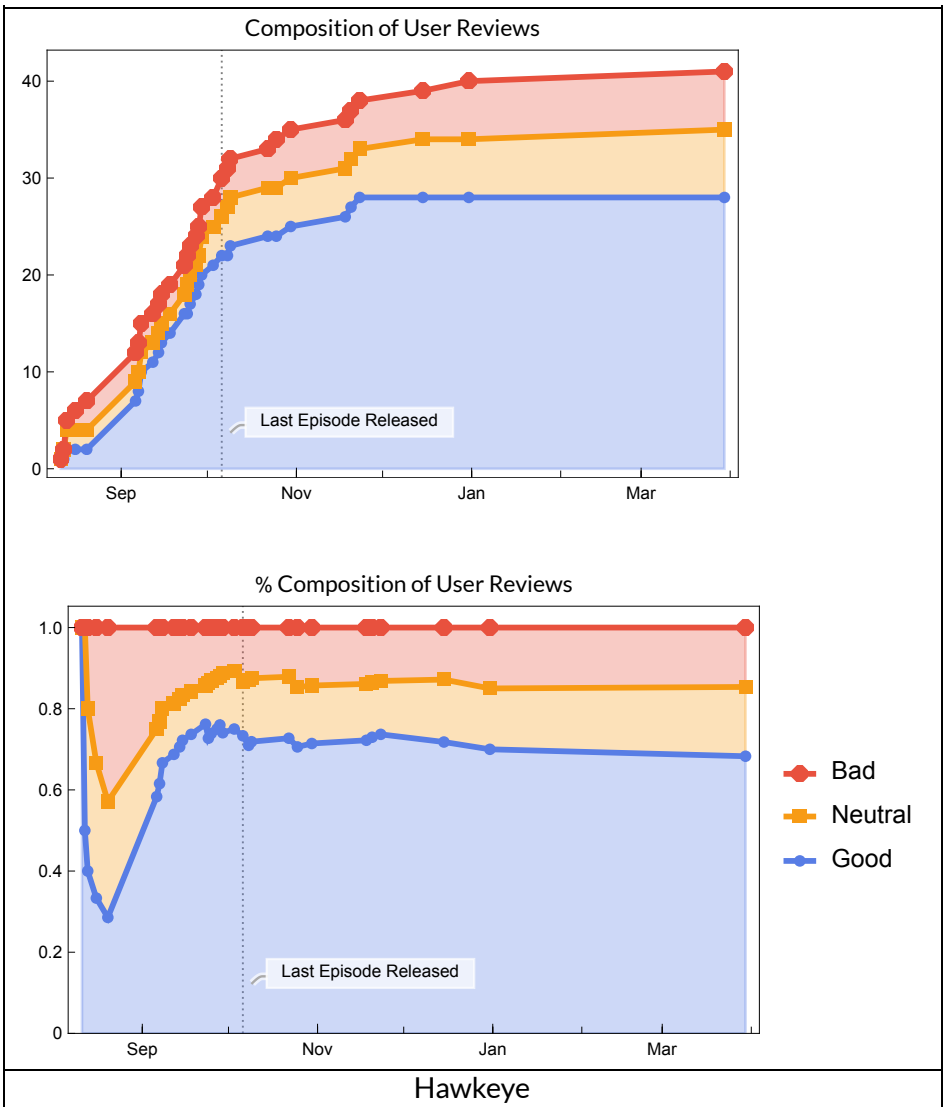


Falcon and the Winter Soldier

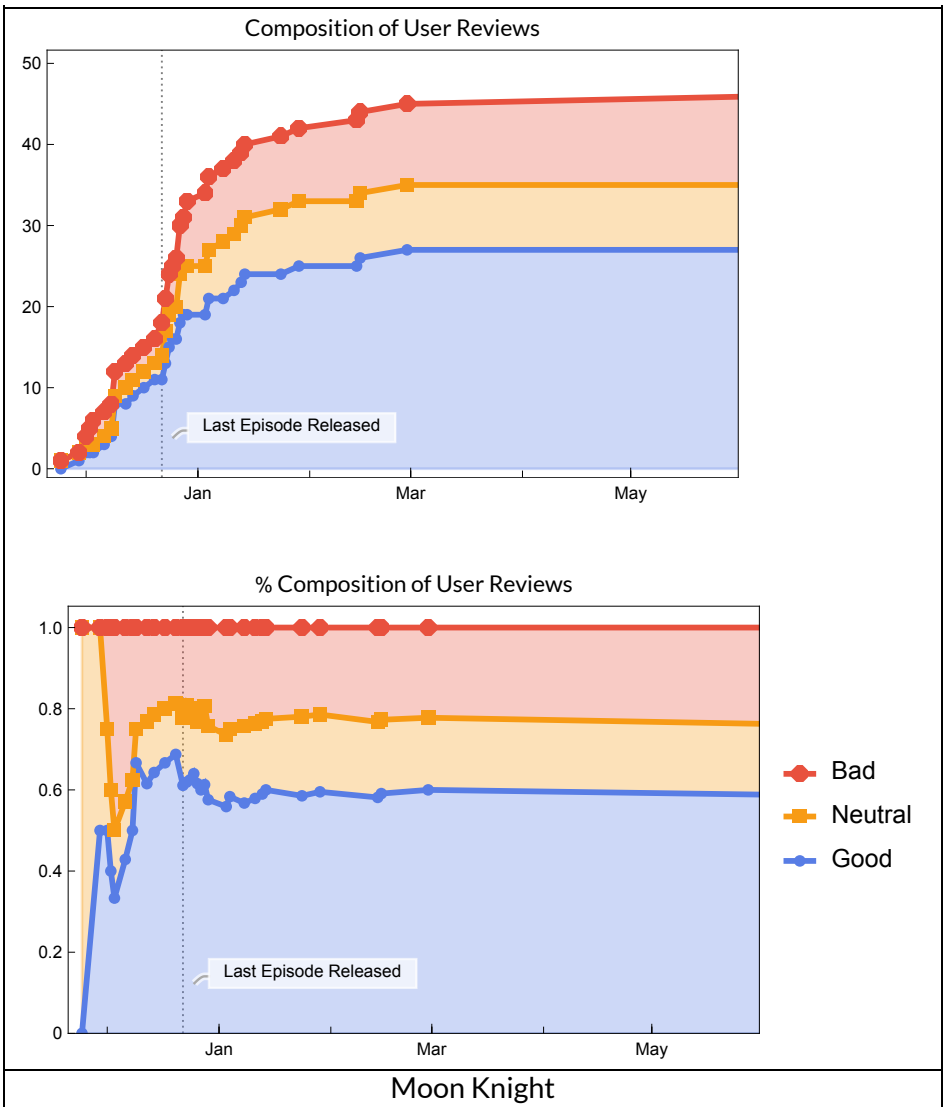

Composition of User Reviews
100
80
60
40
20
0
Last Episode Released
Apr
Jul
Oct
Jan
Apr
% Composition of User Reviews
1.0
0.8
0.6
0.4
0.2
0
Last Episode Released
Apr
Jul
Oct
Jan
Apr
Bad
Neutral
Good

Loki

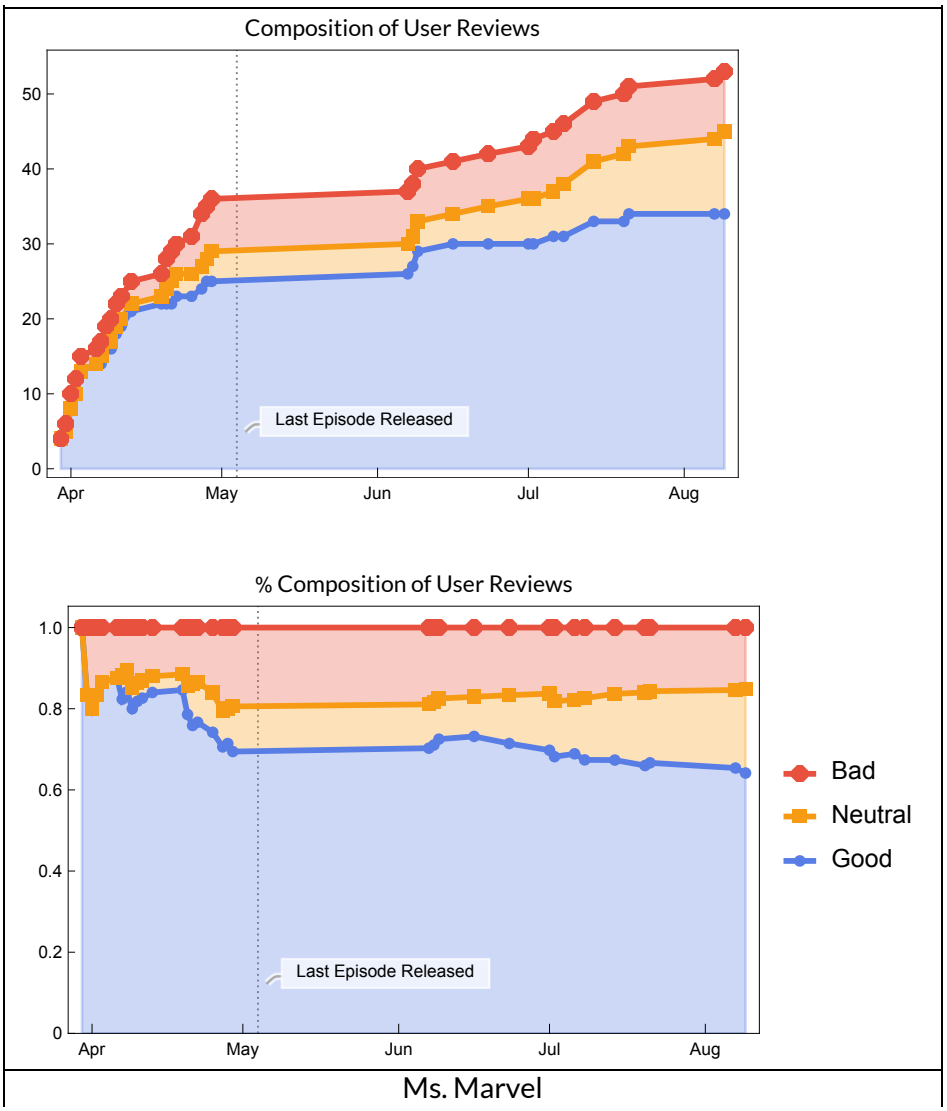

Composition of User Reviews
100
75
50
25
0
Last Episode Released
Jul
Oct
Jan
Apr
Jul
% Composition of User Reviews
1.0
0.8
0.6
0.4
0.2
0
Last Episode Released
Jul
Oct
Jan
Apr
Jul
Bad
Neutral
Good
What If?


Composition of User Reviews
40
30
20
10
0
Last Episode Released
Sep
Nov
Jan
Mar
% Composition of User Reviews
1.0
0.8
0.6
0.4
0.2
0
Last Episode Released
Sep
Nov
Jan
Mar
Bad
Neutral
Good

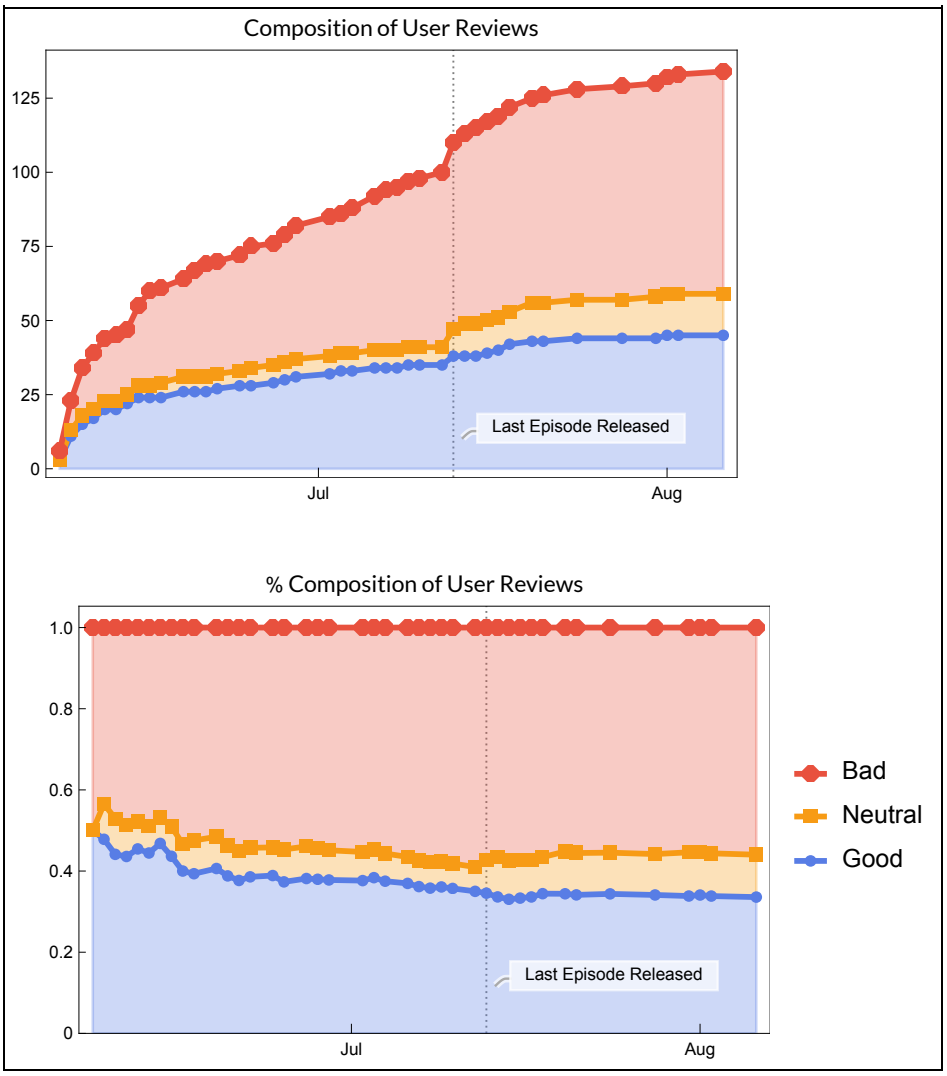
Hawkeye


Composition of User Reviews
50
40
30
20
10
0
Last Episode Released
Jan
Mar
May
% Composition of User Reviews
1.0
0.8
0.6
0.4
0.2
0
Last Episode Released
Jan
Mar
May
Bad
Neutral
Good

Moon Knight


Composition of User Reviews
50
40
30
20
10
0
Last Episode Released
Apr
May
Jun
Jul
Aug
% Composition of User Reviews
1.0
0.8
0.6
0.4
0.2
0
Last Episode Released
Apr
May
Jun
Jul
Aug
Bad
Neutral
Good

Ms. Marvel



So what can we say so far? The distribution of critic scores assigned to *Ms. Marvel* is very strange. It indicates that, almost unanimously, every critic enjoyed the show. So it seems to have hit a real jackpot in terms of what critics are looking for. The distribution of user scores assigned to Ms. Marvel is also very strange, if not for the same reason. It is being pulled in opposite directions by a group of users who, in accordance with critics, are rating the show very highly, and a larger group of users who are rating it terribly. Moreover, there tends to be little to no in-between area here. Taken together, this data does paint a very suggestive portrait of review-bombing. The show is high-quality and so liked by a good number who come across it. However for an even larger portion of people, something about the show drives them to purposefully give it an unreasonably low score. It is the easiest and most realistic explanation of the data. Of course, actually "proving" this is pretty difficult, if not impossible. After all, users can always just claim that they didn't like show.

We can, however, try and provide an even more convincing statistical argument. Let's think about what review-bombing really means for a minute: coordinating with other users to leave lots of bad reviews. We did see some possibility of coordination in the plots of review composition over time, but can we find some more explicit evidence? One thing we might want to take a look at is the length of reviews. It

doesn't seem unreasonable to expect a majority of *Ms. Marvel's* bad reviews to be pretty short, in an effort by review-bombers to churn out as many negative reviews as possible. Metacritic also allows users to rate the "helpfulness" of other users' reviews, which plays a role in determining on which page those reviews are shown. Review-bombers will likely want their negative reviews to be the first thing others see, so we might anticipate that negative reviews for *Ms. Marvel* will have a high helpfulness rating while positive reviews will not. Let's explore both of these things in turn.

How did the length of reviews vary?

We can plot histograms for the character length of positive and negative reviews for each show according to both critics and users. As critics almost never give reviews low enough to fall into the “negative” category (something we mentioned earlier), this categorization doesn’t really tell us anything about critics who didn’t like a given show. However, it does capture a substantial part of each show’s user population.

First, let's note that critic reviews never go above 800 characters, while we see some user reviews going into the low thousands. This is because critic reviews are usually just the first paragraph of a more in-depth review which is linked to at the end of the paragraph. Furthermore, the length distributions for critic reviews, whether good or bad (though the only show with a bad critic review is *Ms. Marvel*), tend to be roughly symmetrical. The length distributions for user reviews on the other hand exhibit much more skew. While *Ms. Marvel* does have a large right skew in the length distribution for negative user reviews (indicating a bias towards lots of shorter reviews), so do many other shows.

In[47]:=
```
scoresToLength =
  Map[KeyMap[ToExpression], Map[x ↦ StringLength /@ x, scoresToReview, {3}], {2}];
```

In[48]:=
```
criticLengthGood =
  First /@ Map[KeySelect[GreaterEqualThan[70]], scoresToLength, {2}];
```

In[49]:=
```
criticLengthBad = First /@ Map[KeySelect[LessEqualThan[30]], scoresToLength, {2}];
```

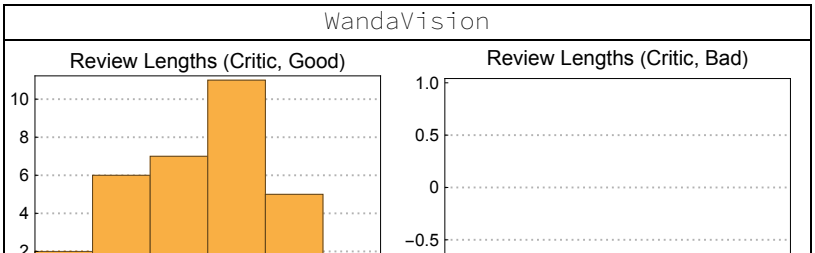
In[50]:=
```
userLengthGood = Last /@ Map[KeySelect[GreaterEqualThan[7]], scoresToLength, {2}];
```

In[51]:=
```
userLengthBad = Last /@ Map[KeySelect[LessEqualThan[3]], scoresToLength, {2}];
```

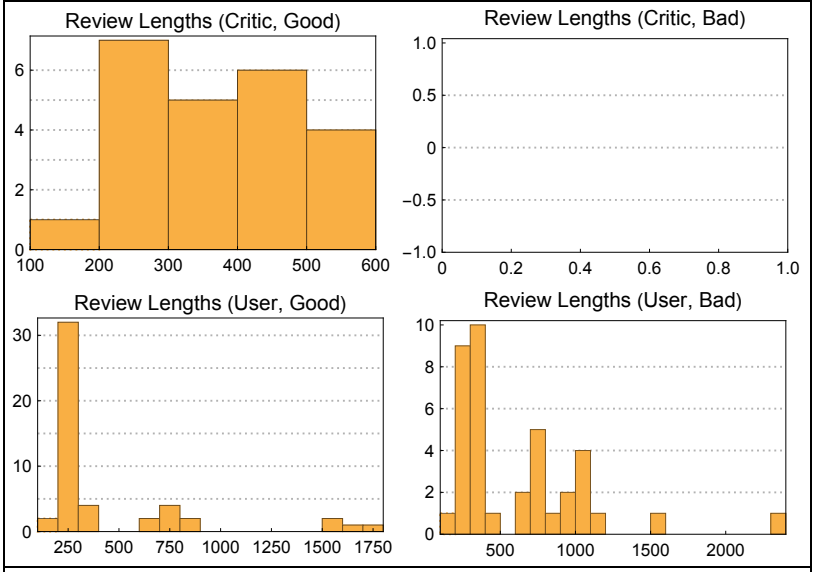
```
collected = (Partition[#, {2}] & /@
    Merge[Map[x ↦ Histogram[#, {100}, ···] & /@ Flatten /@ Values /@ x[[1]],
      {{criticLengthGood, Style["Review Lengths (Critic, Good)",
         Black, FontFamily → "Lato", FontSize → 12]}, {criticLengthBad,
       Style["Review Lengths (Critic, Bad)", Black, FontFamily →]} ,
      {userLengthGood, "Review Lengths (User, Good)"},
      {userLengthBad, "Review Lengths (User, Bad)"}}], Identity]);
```

In[53]:=
```
Grid[Flatten[MapThread[{{#1}, {#2}} &,
   {Keys@collected, Grid /@ Values@collected}], 1], Frame → All]
```

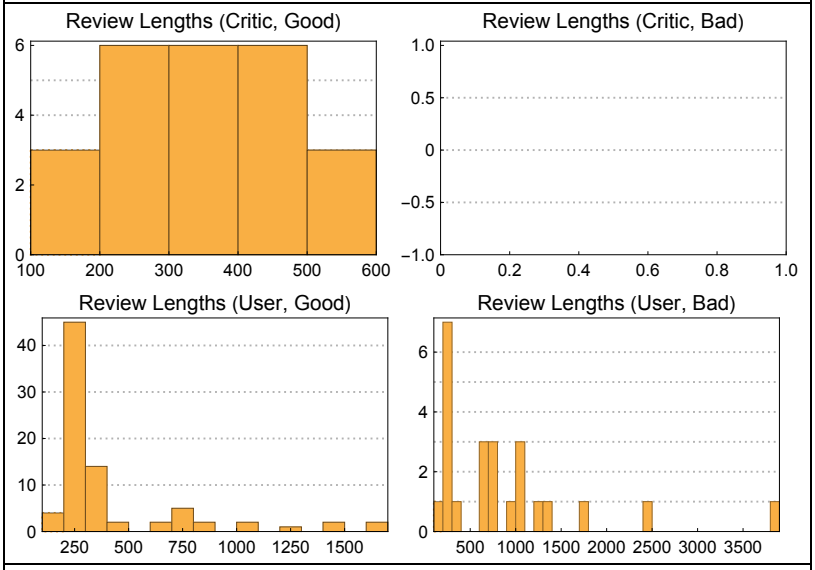
Out[53]=



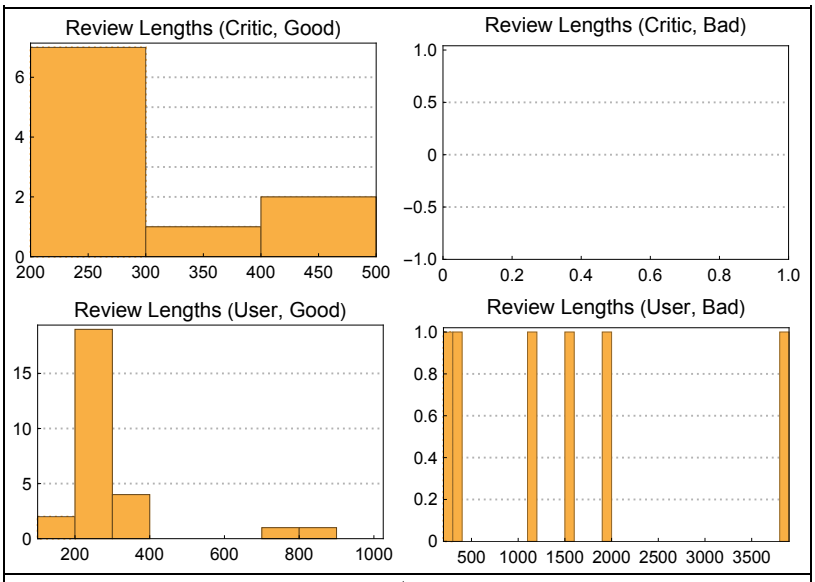
Falcon and the Winter Soldier



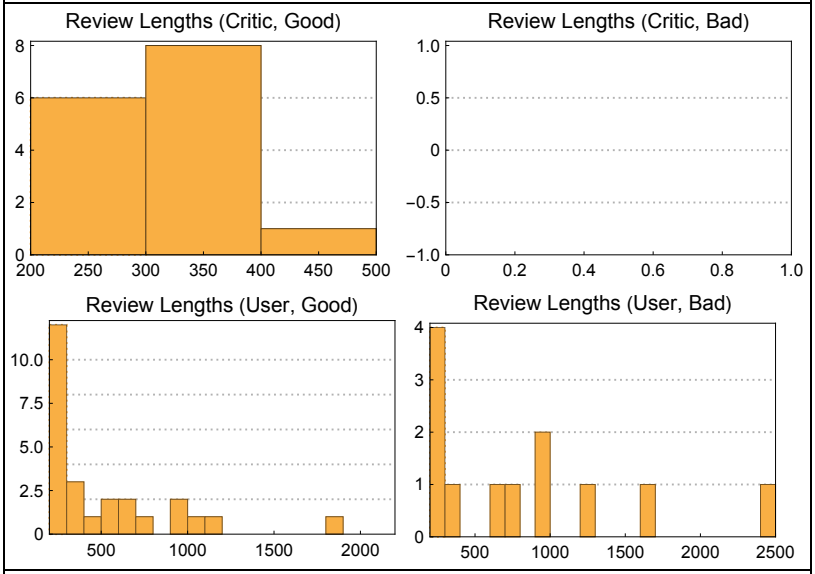
Loki



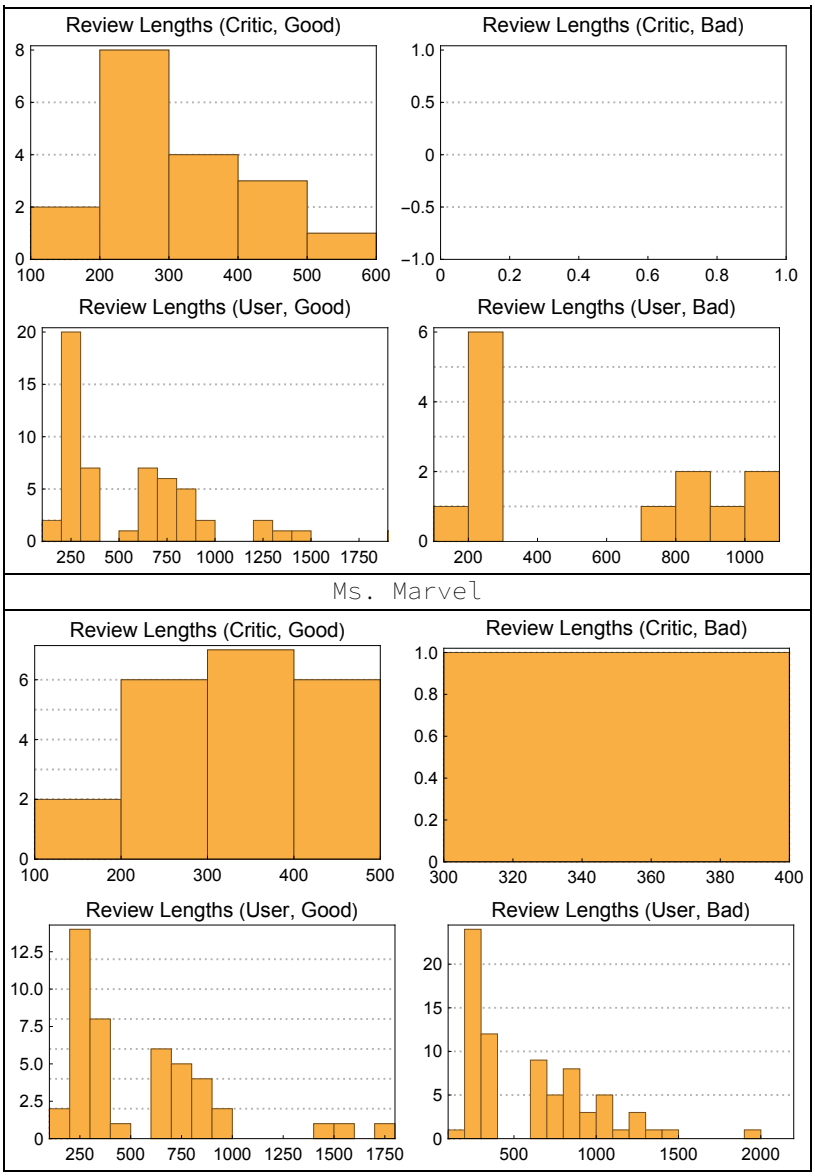
What If?



Hawkeye



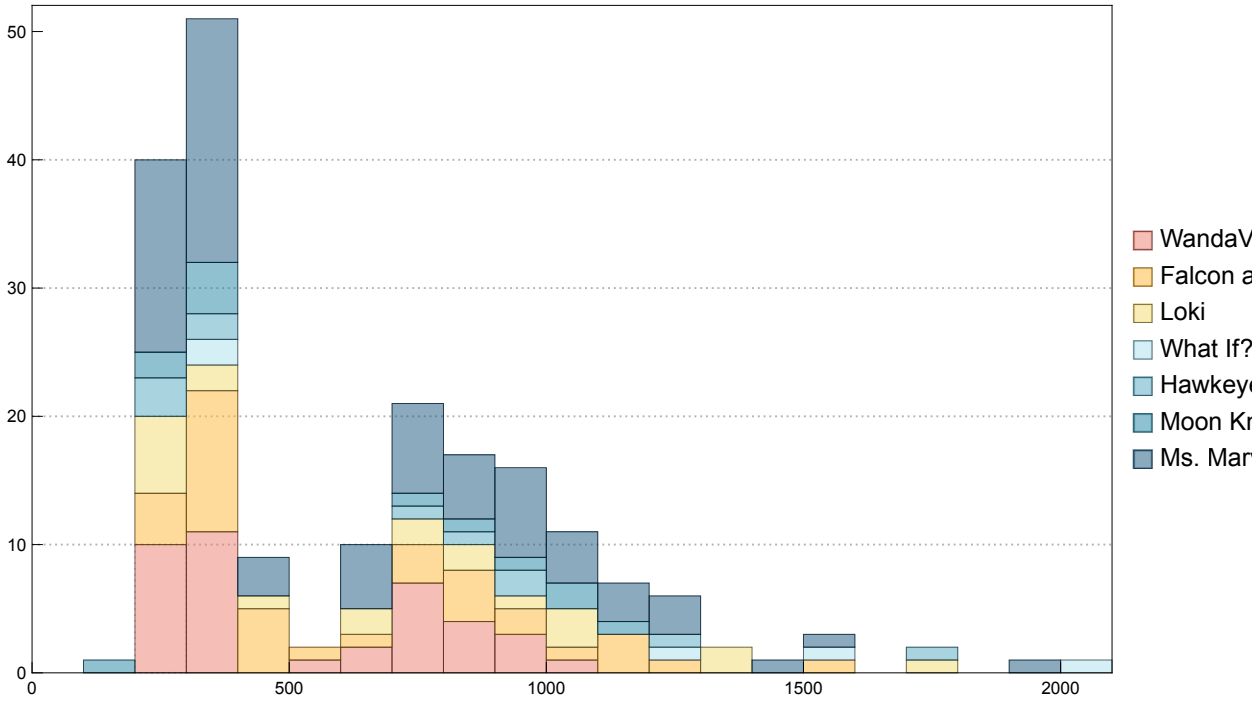
Moon Knight



If we re-render the data from all shows' negative user reviews as a stacked bar graph, we can see this more clearly. Though Ms. Marvel does have the most low-length negative reviews, this is largely due to its far greater number of negative reviews to begin with. The overall percentage split seems pretty much the same across the board.

In[54]:=
```
Histogram[Round[# / 100] * 100 & /@ (Flatten@*Values /@ userLengthBad), ···]
```

Out[54]=

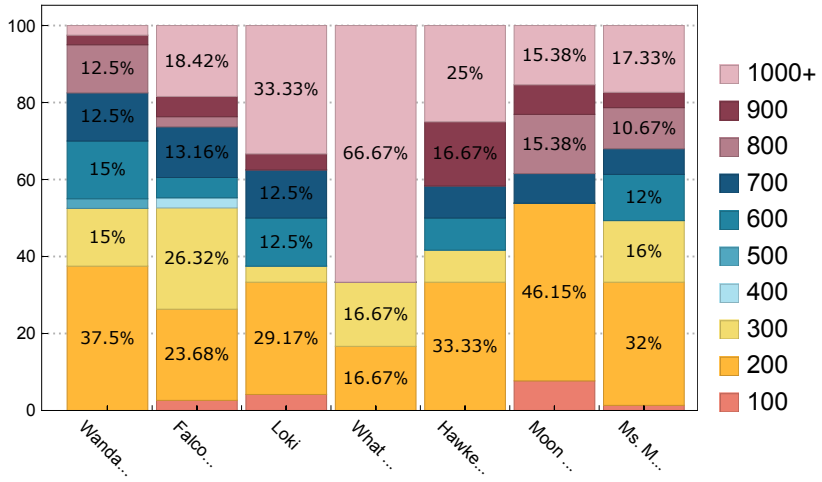


In[55]:=
```
splitBy = Values@(KeyMap[If[# * 100 == 1000, "1000+", ToString[# * 100]] &] /@ KeySort /@
      ((Merge[{#, AssociationThread[Range[10], Table[0, {x, 10}]]}, First] & /@
        ((Length /@ GroupBy[#, x ↦ If[x < 1000, Floor[x / 100], 10]] & /@
          (Flatten@*Values /@ userLengthBad)))));
```

If we really want to verify this, we can re-render the data once again, this time in the format of a stacked bar chart (where the y-axis represents the percent of reviews falling into each length category). We see that *Moon Knight* actually has the largest percentage of short reviews (100-200 characters) at around 50%, and that *What If?* has the largest percentage of long reviews (1000+ characters), at almost 70%! However, this actually only amounts to a measly four reviews since *What If?* received so few reviews (and thus negative reviews) to begin with. The same is true for *Moon Knight*, and so we can largely dismiss some of these more dramatic results as noise.

In[56]:=
```
BarChart[Values@splitBy, ···]
```

Out[56]=



So it seems like we don't have very strong evidence of *Ms. Marvel* being review-bombed when examining the length distributions of negative user reviews. But maybe that's the problem – so far we've constrained our analysis to just negative user reviews, without comparing them to positive and neutral user reviews. What happens if we do this?

Well, let's find out. All we need to do is plot the scores of all user reviews for each show against their corresponding review lengths and then calculate their LSRLs. Unfortunately, the results aren't terribly interesting, or even that different from what we saw before. Because there is so much variability in the data, most of the regression lines end up with a slope of about zero, and R^2 values on the order of 10^{-3} (indicating that the data doesn't following anything remotely close to a linear relationship). All this means is that there is no pattern, for any show, in terms of review length and score. Once again, the only exception here is *What If?*, which exhibits a negative linear trend and an R^2 value of around 20%. However, its population of user reviews is simply to small for us to have much confidence in this result, especially when compared to other shows. If anything, it might indicate that users who scored the show poorly *sometimes* wrote longer reviews, perhaps because the "multiversal" themes of the show were ripe with inconsistencies that comic book fans simply felt compelled to point out. When actually reading thorough the negative reviews for *What If?*, this does seem to be somewhat – but not largely – the case.

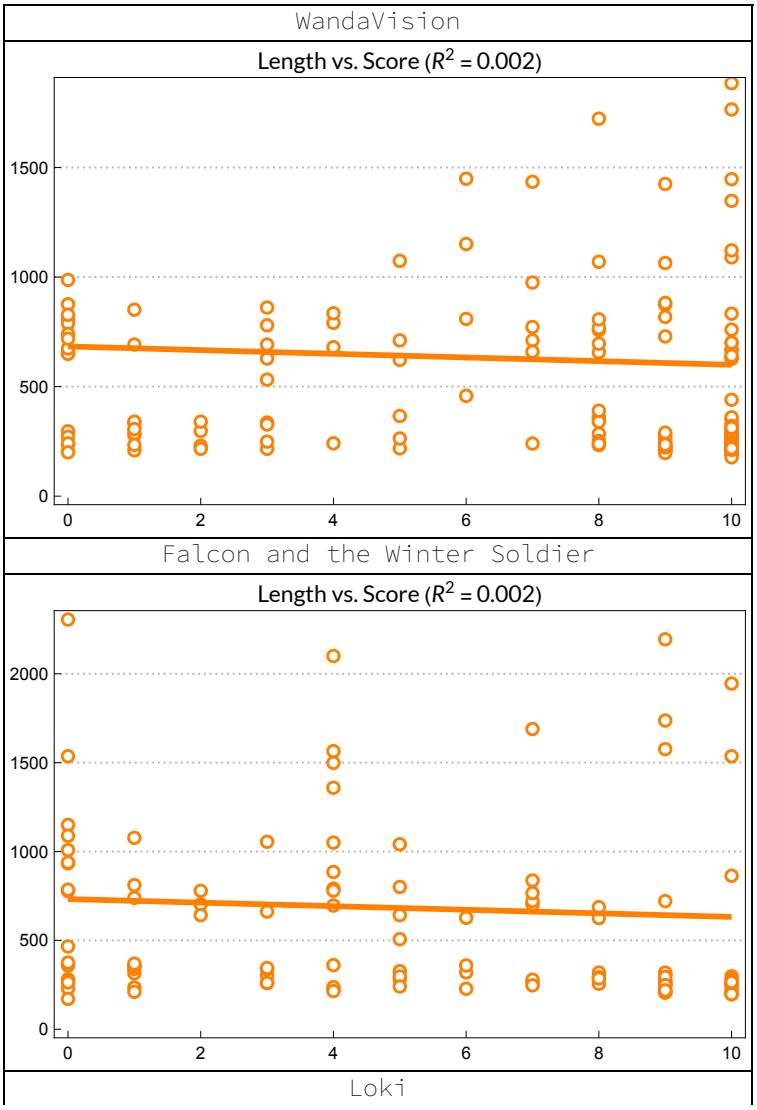
In[59]:=
```
lengthScoreCorrelations = Map[
   (Flatten[#, 1] &)@*KeyValueMap[Table[{#1, x}, {x, #2}] &], scoresToLength, {2}];
```

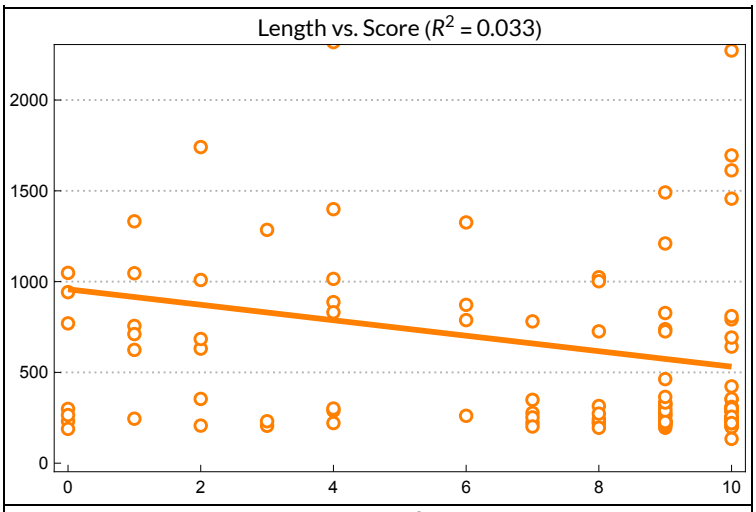
In[60]:=
```
{lengthModels, lengthPlots} =
  {Map[(LinearModelFit[#, x , x] &), lengthScoreCorrelations, {2}],
   {ListPlot[First@#, ···] & /@ lengthScoreCorrelations, Association@
     Table[marvelShows[[x]] → ListPlot[Last@lengthScoreCorrelations[[x]], ···],
      {x, Length@lengthScoreCorrelations}]}};
```

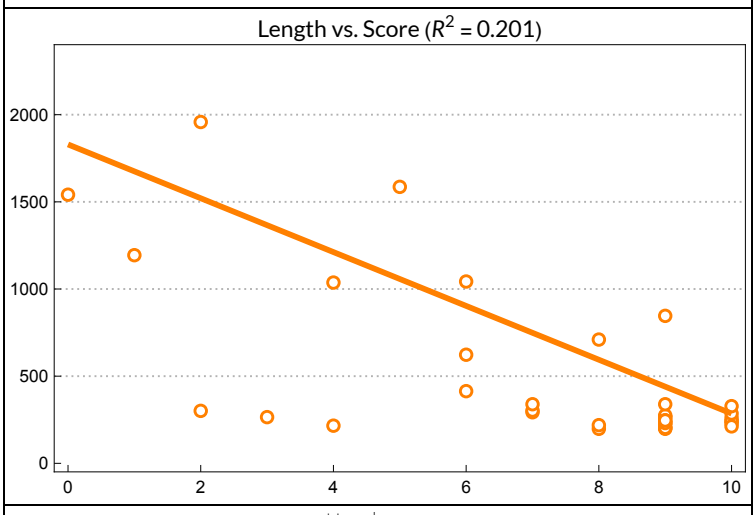
In[61]:=
```
newLengthModels = {Plot[#[[1]]["BestFit"], {x, 0, 100}, ...] & /@ lengthModels,
   Plot[#[[2]]["BestFit"], {x, 0, 10}, ...] & /@ lengthModels};
```

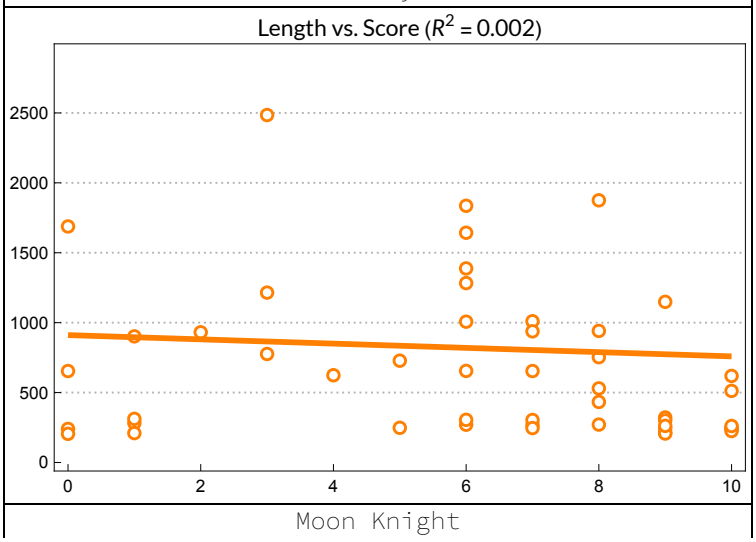
In[62]:=
```
Grid[Flatten[KeyValueMap[{{#1}, {Last@#2}} &] @ (Merge[Map[Show,
      MapThread[(Merge[{KeyMap[ToString, #1], KeyMap[ToString, #2]}, Identity]) &,
       {lengthPlots, newLengthModels}], {2}], Identity]), 1], Frame → All]
```

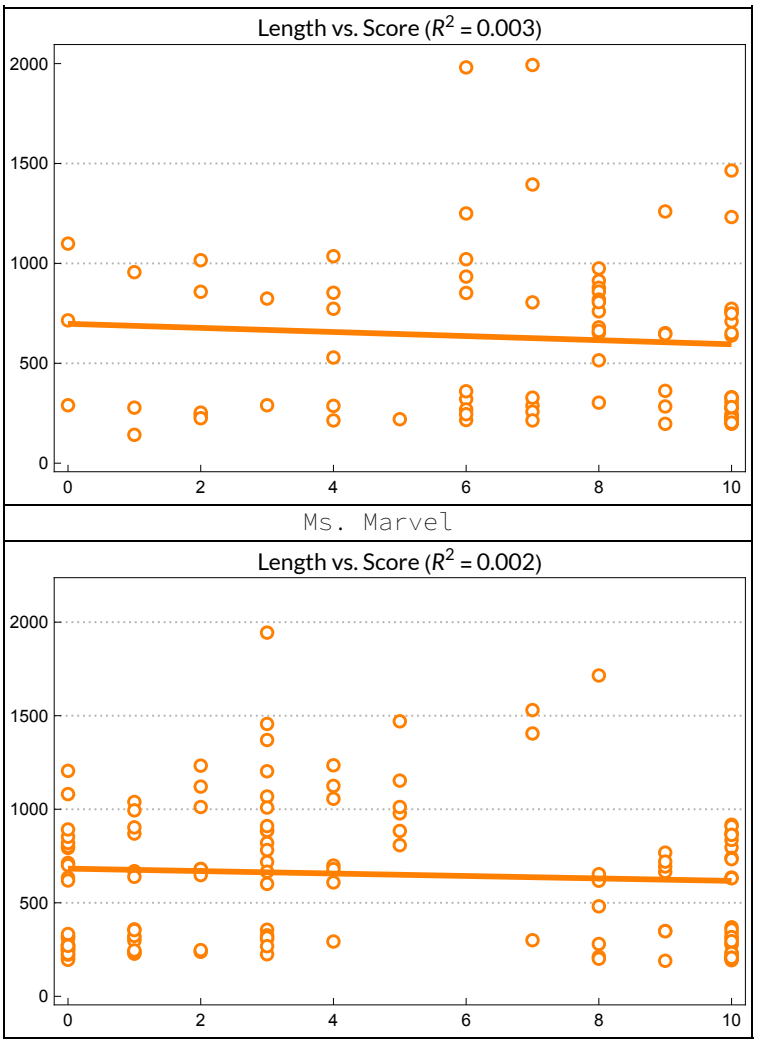
Out[62]=




Length vs. Score (R^2 = 0.033)
2000
1500
1000
500
0
0
2
4
6
8
10
What If?


Length vs. Score (R^2 = 0.201)
2000
1500
1000
500
0
0
2
4
6
8
10
Hawkeye


Length vs. Score (R^2 = 0.002)
2500
2000
1500
1000
500
0
0
2
4
6
8
10
Moon Knight



To be fair, a lot of this analysis is unfounded, since the length of someone's review is extremely random, depending on many factors other than their opinion of the show (such as their native language, writing style, environment and amount of time available when writing, etc.). But I still wanted to go through all of these steps for completeness. In conclusion, we don't see evidence for review-bombing in the distribution of review lengths for *Ms. Marvel* when compared to other shows. Then again, I'm not sure how prevalent such a relationship is in other review-bombed motion pictures, but short, "low-effort" reviews do seem like something that would accompany review-bombing. Given the other compelling evidence (both statistical and non-statistical) that *Ms. Marvel* has been review-bombed, it is interesting that we don't see too much of this phenomenon, and we'll touch on some possible explanations for this a little later. In the meantime, let's move on to analyzing another characteristic of user reviews: their helpfulness rating.

How helpful were reviews rated?

If we extract the helpfulness ratings from each shows' user reviews and plot them against their corresponding scores, we see a very definitive pattern begin to emerge. Most shows have a moderately

strong, positive linear trend, with positive reviews being rated as very helpful, and negative reviews being rated as very unhelpful. It is not difficult to imagine why: many negative reviews will likely be rude or overly disparaging rather than providing considerate, constructive feedback.

Only three shows stand out from this trend with negative LSRLs: *Hawkeye*, *Moon Knight*, and *Ms. Marvel*. In other words, for these shows, negative reviews are rated as helpful, while positive reviews are not. Of these three, *Ms. Marvel's* LSRL has the highest R^2 value, indicating that the negative linear trend best explains its data. Moreover, we note that the LSRLs for the other two shows are actually underestimates, since a majority of their data points lie above their corresponding LSRLs. This means they are underestimating the trend as more negative than it really is. Meanwhile, most of the data points for *Ms. Marvel* fall below its LSRL, so its LSRL is overestimating the trend as less negative than it really is. Thus of all shows, *Ms. Marvel* experiences this effect most – even more so than our regression predicts. This is exactly the sort of behavior we might expect to accompany review bombing!

We should however note that the scale for measuring helpfulness is a little flawed here. Both a review with an upvote of 1 and 100 will achieve the same helpfulness rating of 100%, since 100% of users who have weighed in rated the review as helpful. However one rating is quantitatively more valuable than the other. Unfortunately taking this into account when determining our LSRL is not particularly easy to do.

In[63]:=
```
extractHelpful[text_] :=
 StringCases[text, ((x : DigitCharacter ..) ~~ " of " ~~ (y : DigitCharacter ..) ~~
     " users found this helpful") → {x, y}]
```

In[64]:=
```
helpfulScoreCorrelations = Map[DeleteCases[#, {}] /. <| |> → Nothing &,
   Map[(ToExpression /@ extractHelpful[#]) /. {{} → Nothing, {{0, 0}} → Nothing} &,
    scoresToReview, {4}], {2}];
```

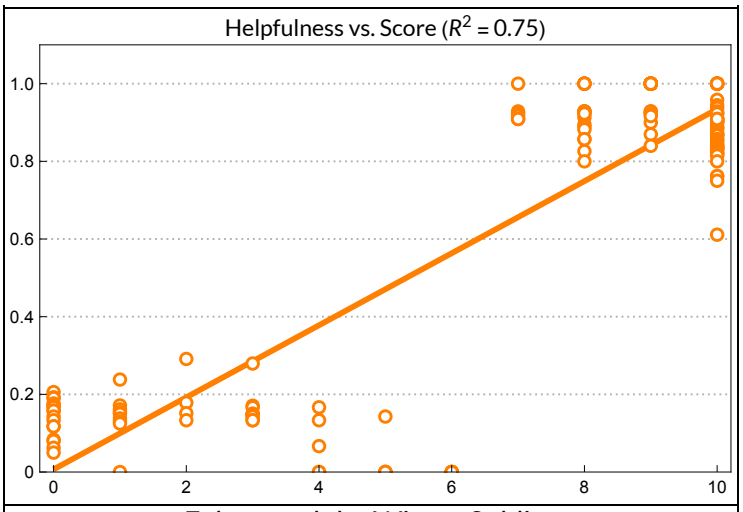
In[65]:=
```
helpfulScoreCorrelations =
  Flatten[#, 1] & /@ KeyValueMap[Table[{ToExpression@#1, x}, {x, #2}] &] /@
    Last /@ Map[(Divide @@ (Flatten@#)) &, helpfulScoreCorrelations, {4}];

{helpfulModels, helpfulPlots} =
  {LinearModelFit[#, x , x] & /@ helpfulScoreCorrelations,
   ListPlot[#, ⋯] & /@ helpfulScoreCorrelations};
```

In[67]:=
```
newPlots =
  Show @@ # & /@ Merge[{Plot[#["BestFit"], {x, 0, 10}, ⋯] & /@ helpfulModels,
     helpfulPlots}, Identity];
```

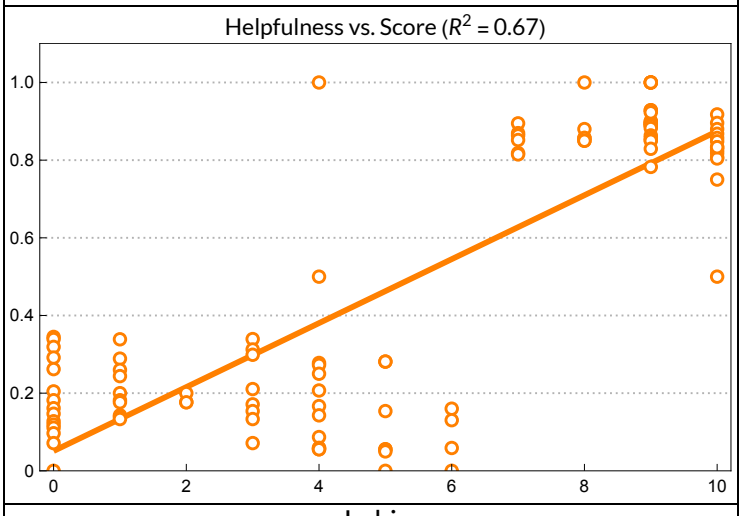
In[68]:=
```
Grid[Flatten[MapThread[{{#1}, {#2}} &,
   {Style[#, Black, FontFamily → "Lato", FontSize → 15] & /@ Keys@newPlots,
    Values@newPlots}], 1], Frame → All]
```

Out[68]=

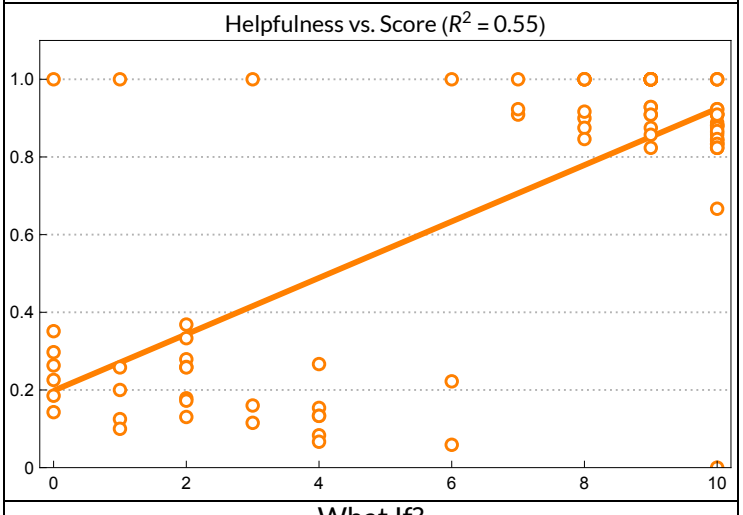
WandaVision



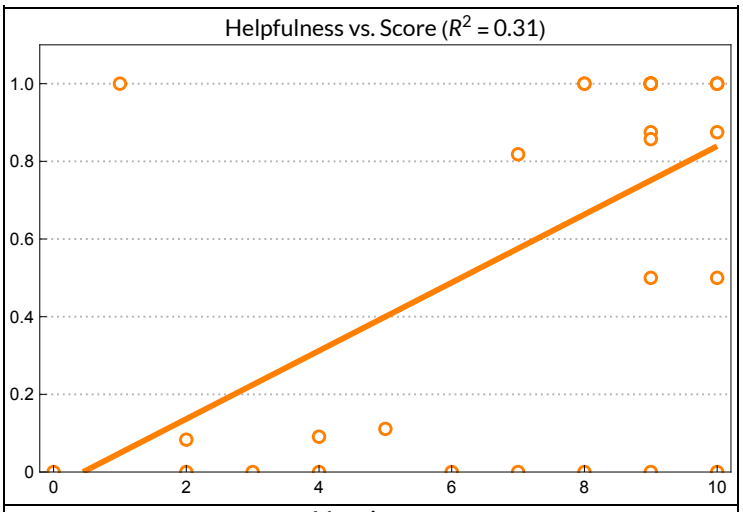
Falcon and the Winter Soldier



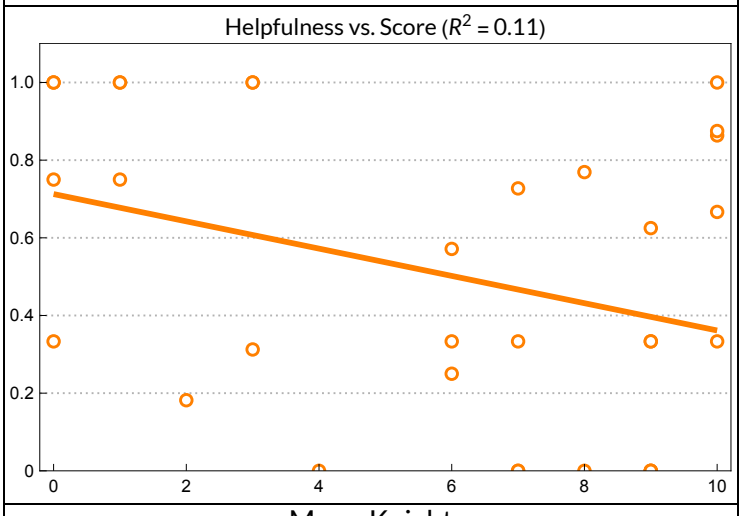
Loki



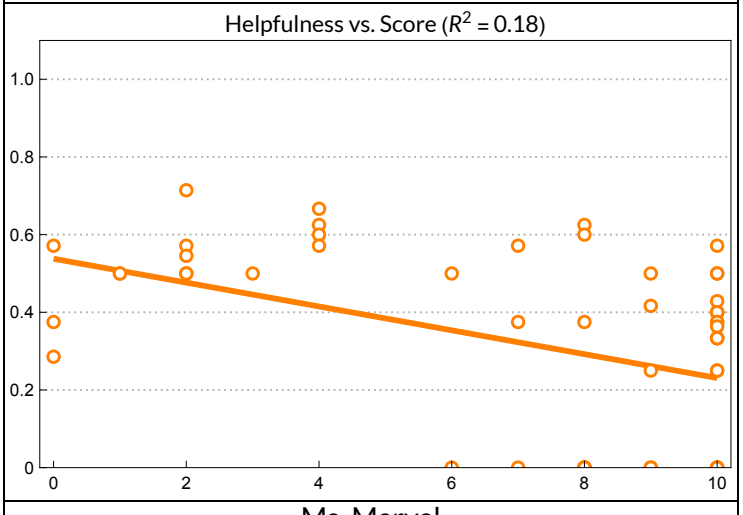
What If?



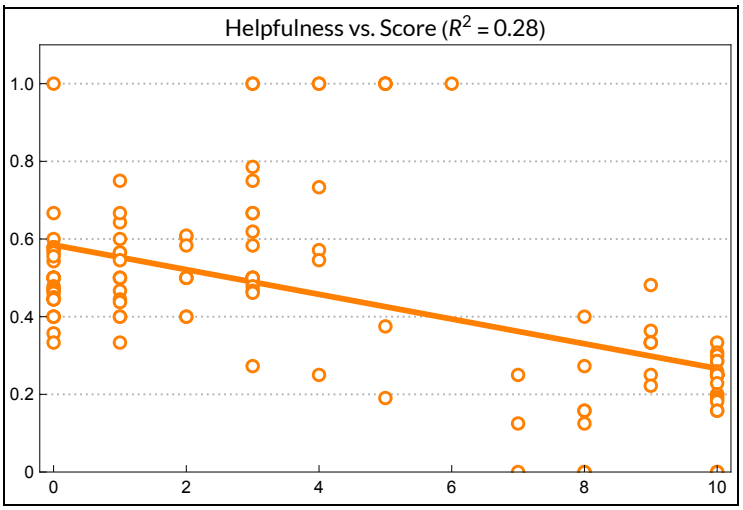
Hawkeye



Moon Knight



Ms. Marvel



All in all, we have obtained some pretty convincing evidence of review-bombing. Review-bombers are downvoting positive reviews for *Ms. Marvel* while at the same time upvoting negative ones. While a few shows seem to observe this trend, their correlations is much weaker than *Ms. Marvel's*, and are often underestimates. Interestingly, many users who have posted positive *Ms. Marvel* reviews have reported instances of this exact same phenomenon other platforms, such as YouTube [11].

We could keep digging, but I think that our statistical argument is sufficiently sturdy at this point, especially when taking into account the various non-statistical information we have in favor of review-bombing. At this point, I'm certainly convinced, and I expect most others will be to. So let's now take *Ms. Marvel's* review-bombing to be a proven fact. But answering our initial question only opens up an even more interesting one: *why is the show being review-bombed to begin with?*

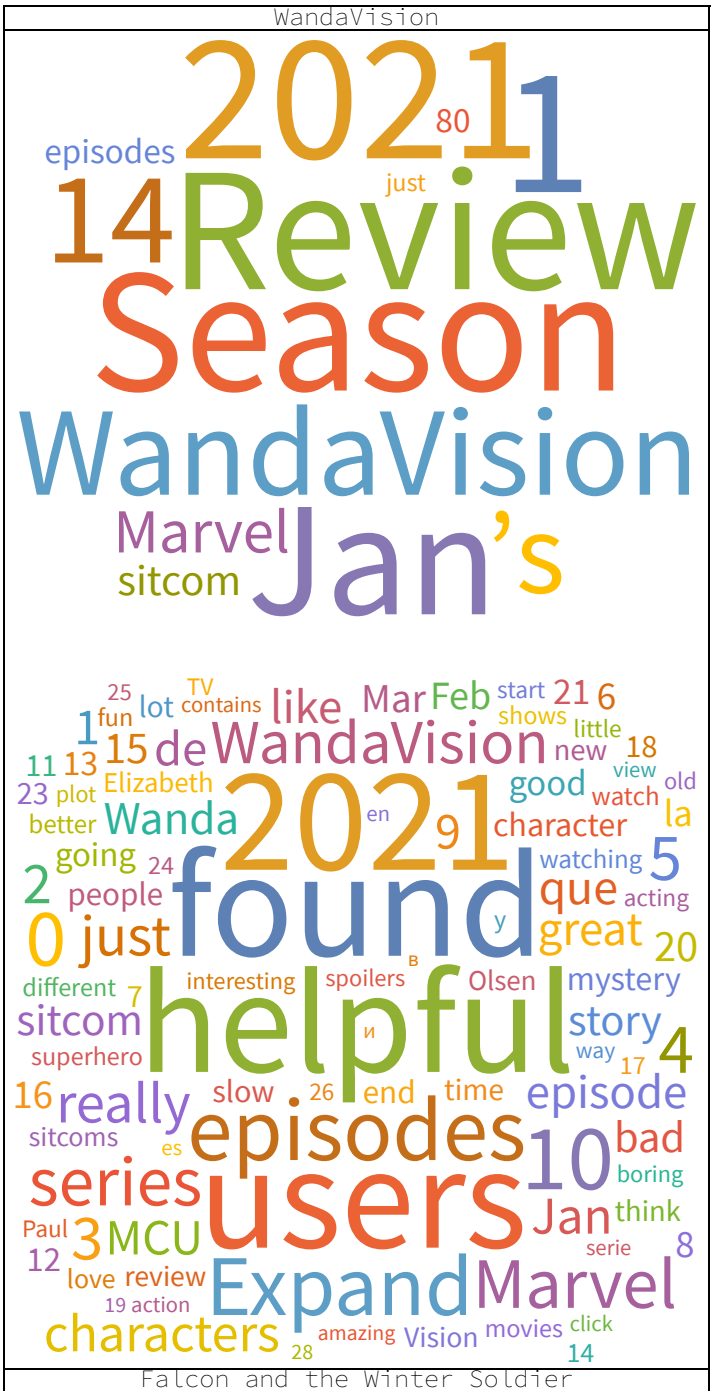
What types of language are prominent in reviews?

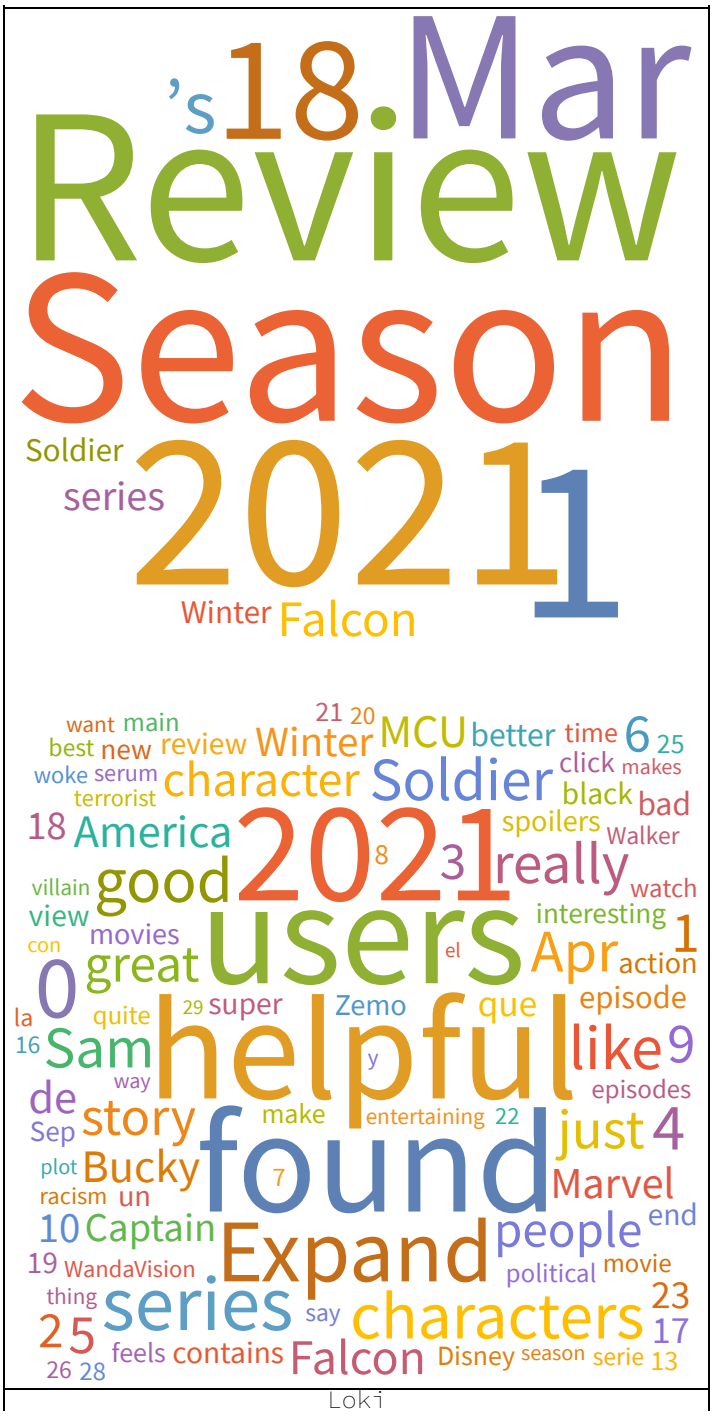
We can assume that users will at least try and explain the rationale behind their score within their review (though we can't always assume their full honesty). So let's look at the actual text that features in these reviews. If we count the occurrences of all words in the critic and user reviews for each show (disregarding "stopwords" – e.g., "this", "the", "is", etc. – as well as those that don't appear at least 10 times), we can generate wordclouds for each show according to both critics and users (note that in the visualization below, the critic wordclouds are on the top and the user wordclouds are on the bottom). The critic wordclouds reveal very little, since critics' reviews use very distinct and variegated language, and so the only words that appear frequently across all reviews are things such as the show title, the release date of the show, and the names(s) of main actor(s). User wordclouds, meanwhile, tend to be quite a bit more descriptive. Notably, since there are less possible scores for users to assign, the word-clouds also show the most common score(s).

In[69]:=
```
topWords = Map[Select[(# // StringJoin // DeleteStopwords // WordCounts),
    GreaterEqualThan[10]] &, allReviews, {2}];
```

In[70]:=
```
Grid[Flatten[Transpose[Partition[#, UpTo[1]] &/@
      ({Keys@#, Values@#} &@(Column[#, Center, Spacings → 2] &/@
          Map[WordCloud[#, ImageSize → Medium] &, topWords, {2}]))],
  1], Frame → All, Spacings → {Automatic, 0}]
```

Out[70]=




’s
18.
Mar
Review
Season
2021
1
Soldier
series
Winter
Falcon
want
main
21
20
MCU
better
time
6
25
best
new
review
Winter
click
makes
woke
serum
character
Soldier
black
bad
terrorist
18
America
2021
spoilers
Walker
8
3
really
villain
good
watch
view
movies
users
interesting
1
con
el
great
Apr
action
0
la
quite
29
super
Zemo
que
episode
16
Sam
helpful
like
9
y
way
episodes
de
story
make
entertaining
22
just
4
Sep
found
plot
Bucky
7
Marvel
racism
un
10
Captain
people
end
Expand
19
WandaVision
political
movie
thing
series
say
characters
23
17
25
feels
contains
Falcon
Disney
season
serie
13
26
28
Loki

2021
Review
Season
1
8
Loki
's
Marvel
Jun
series
23
lot
TVA
future
far
Martino
worth
contains
little
13
8
22
scenes
especially
Marvel
multiverse
view
season
Avengers
Disney
makes
Owen
amazing
Sophia
shows
spoilers
Expand
episodes
de
Hiddleston
15
bad
Sep
Aug
17
interesting
11
really
Loki
great
Di
Sylvie
thing
9
series
2
5
episode
14
female
's
7
watch
10
helpful
0
Jul
1
better
click
2022
action
story
new
way
good
make
plot
MCU
found
just
Tom
3
love
work
6
development
users
entertaining
12
Jun
watching
time
acting
characters
2021
character
like
Wilson
review
best
18
4
20
finale
know
end
TV
24
16
What If?

Review
Season
Aug1
2021
Marvel
2021
1
episodes
3
like
6
de
4
story
helpful
Sep
Good
Expand
10
la
feel
episode
really
found
0
9
interesting
7
8
series
users
en
just
stories
MCU
5
2
Hawkeye

Hawkeye
Review
Season
23
2021
's
1
Nov
shows
better
fun
2022
Hailee
series
moments
story
2021
Expand
Marvel
really
good
far
1
characters
9
4
Clint
helpful
episodes
3
episode
Bishop
's
le
like
Hawkeye
que
2
la
de
0
found
Dec
just
10
scenes
6
new
best
Disney
users
great
7
MCU
Jeremy
5
character
Kate
Steinfeld
action
Christmas
8
Moon Knight

Moon
Mar
Review
Season
's
1
Knight
29
2022
bad
love
Hawke
main
season
people
characters
MCU
10
like
just
character
Apr
best
Egyptian
great
far
6
Moon
2022
Marvel
8
2
performances
users
really
job
think
pretty
Ethan
7
0
helpful
story
3
9
Steven
de
que
good
found
episode
4
1
's
series
end
interesting
Knight
Oscar
5
performance
feel
amazing
Expand
episodes
acting
Disney
watching
Isaac
shows
start
watch
fun
action

Ms. Marvel

Let's take a closer look at *Ms. Marvel* in particular. We are interested in how the wordclouds differ between positive and negative reviews. We can generate four types of wordclouds: a wordcloud of the words that feature prominently in positive reviews, a wordcloud of the words that feature prominently in negative reviews, and wordclouds of words that are exclusive to positive and negative reviews up (so

we can get an even better sense of how they differ). Since we are now restricting ourself to a smaller set of reviews, we only require a word to be mentioned twice to be included in one of our wordclouds (this also means that our wordclouds for exclusivity are only sensitive up to a count of two).

In[71]:=
```
scoresToTopWords =
  Map[Select[WordCounts[DeleteStopwords@StringJoin[#]], GreaterEqualThan[2]] &,
   scoresToReview, {3}];
```

In[72]:=
```
scoresToTopWords = Map[KeyMap[ToExpression], scoresToTopWords, {2}];
```

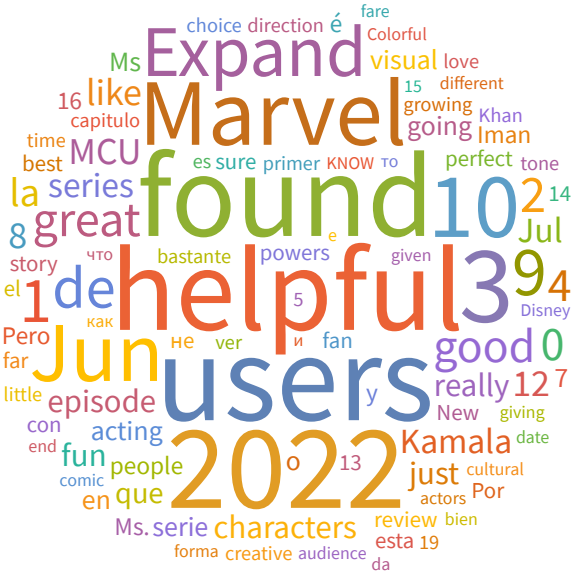
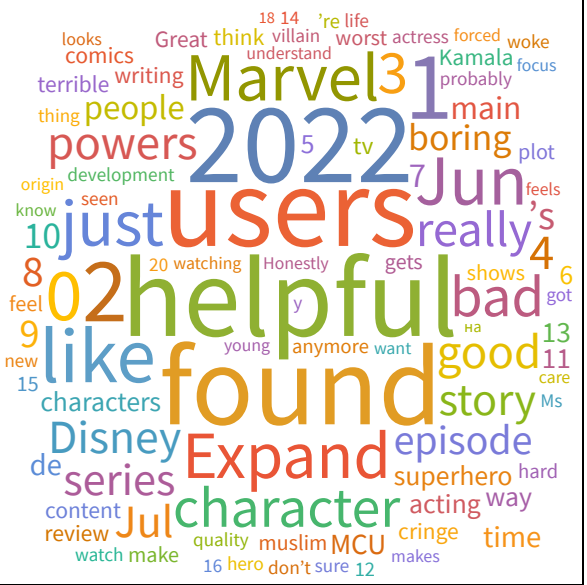
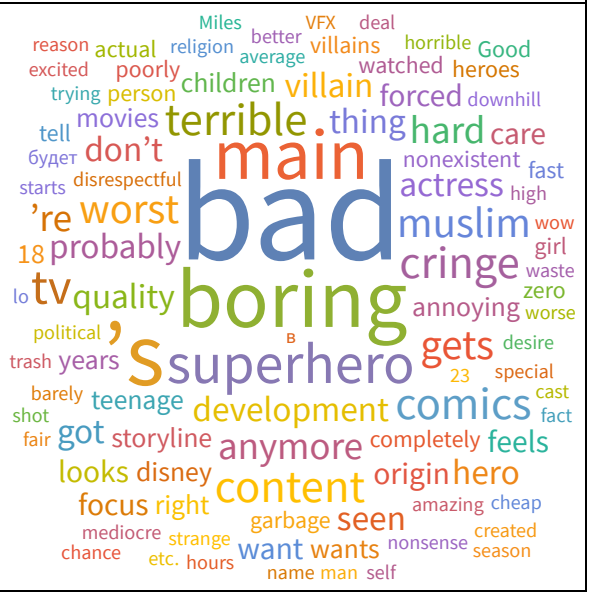
In[73]:=
```
{userGood, userBad} =
  {Last /@ Map[Merge[#, Total] &@* Values, Map[KeySelect[#, GreaterEqualThan[7]] &,
       scoresToTopWords, {2}], {2}], Last /@ Map[Merge[#, Total] &@* Values,
     Map[KeySelect[#, LessEqualThan[3]] &, scoresToTopWords, {2}], {2}]};
```

In[74]:=
```
goodVBad[show_] := Module[{good, bad},
  {good, bad} = {userGood[show], userBad[show]};
  Grid[{{"Good User Keywords (All)", "Bad User Keywords (All)"},
    {WordCloud[good, ImageSize → 300], WordCloud[bad, ImageSize → 300]},
    {"Good Users Keywords (Unique)", "Bad User Keywords (Unique)"},
    {WordCloud[Association@(# → good[#] & /@ Complement[Keys@good, Keys@bad]),
      ImageSize → 300],
     WordCloud[Association@(# → bad[#] & /@ Complement[Keys@bad, Keys@good]),
      ImageSize → 300]}}, Frame → All]]
```

Doing this for *Ms. Marvel* is very revealing. First, let's examine both the wordcloud of all good reviews, as well as the wordcloud containing words exclusive to good reviews. There is a lot of overlap here. Interestingly, we find lots of Spanish and Russian language in both of these categories. Upon further review of the data, there are quite a few reviews left by Spanish and Russian speakers (~5). Overall, we see the generic positive language we would expect, such as "good", "great", "creative", "super", "fun", "love", etc . There is also lots of mention the characters and cast (specifically, Iman Vellani, the show's lead actor) and descriptors such as a "charismatic", "colorful", and "relatable". If we look we at the wordcloud of all negative reviews, we see – in addition to generic negative language (e.g., "bad", "trash", "boring", "terrible", "cringe") – words like "Muslim","powers", "comics", "religious", "woke", "political", and "VFX", all of which are further amplified in the wordcloud containing language exclusive to negative reviews.

In[ ]:= **goodVBad["Ms. Marvel"]**

Out[ ]=

Good User Keywords (All)	Bad User Keywords (All)
	
Good Users Keywords (Unique)	**Bad User Keywords (Unique)**
	

So let's break this down. Firstly, a large number of negative reviews mention themes of race and religion. Actually reading some negative reviews makes this even more obvious. Many users invoke "superhero fatigue" as their rationale for rating the show badly, saying that the only thing special about Ms. Marvel is that she is a "Pakistani" or "Muslim" superhero and criticizing the show for "spending too much time around her roots" ("she's Indian" – someone wasn't paying attention! – "we get it"). Of course, this logic is quick to break down when we apply it to other heroes in the MCU...we have been adding white superhero after white superhero for quite a while now without any mention of superhero fatigue. And that's because we have never taken ethnicity to be synonymous with uniqueness. Instead

it is the perspective of a superhero as informed by their worldview, skills, and the people they surround themselves with that allows them to leave a distinct mark among the panoply of Marvel heroes. Ethnic background undoubtedly colors a part of this perspective, but it does not constitute it. In the case of Ms. Marvel, this perspective is that of a teenager growing up in a world of superheroes whom she endlessly admires and fantasizes about, until getting powers of her own forces her to rethink what being a superhero actually means. I don't mean to imply it's a revolutionary concept (really, it's a pretty time-tested comic book trope), but it's hardly devoid of purpose either. And casting it in a new cultural and religious context *does* promise that we won't be watching Spiderman 2.0, something that many review-bombers say they don't want, but perhaps secretly desire.

Secondly, some users point to the change in Ms. Marvel's power set from the comics ("power", "comics") as a reason for disliking the show. To be fair, I can't really relate to this as I have little familiarity with the source material. From some light research, it seems that in the comics Ms. Marvel's powers are more akin to shapeshifting, which better resonate with the themes of struggling to figure out her identity that penetrate both the comics and the show. I do agree this is definitely the most extensive reworking of powers (shapeshifting to energy manipulation) we've seen in the MCU so far, but it's interesting that we've never seen people get so upset over power changes before. I mean, the whole point of *WandaVision* was to retcon the origin and scale of Wanda's abilities, and correct the watered-down Scarlet Witch we had from *Age of Ultron* to *Avengers: Endgame*. Similarly, in *Black Widow*, Taskmaster's identity and capabilities underwent serious modifications for the sake of the story (for example, making her abilities due to an external suit, and removing the downside of memory loss that comes when copying new abilities). And don't even get me started on the Mandarin's "rings" in *Shang-Chi*. The list goes on and on. Of course, this is definitely is a valid criticism of the show, and I respect anyone who feels like they could not enjoy the show because of it. But it also seems like the sort of thing that would-be review-bombers could easily use as false justification for their scoring. The same goes for complaints about poor cinematography ("VFX"). And it is perhaps for this very reason that we don't see as many short reviews as we might expect for *Ms. Marvel*. Of course, that last part is just conjecture!

Finally, words like "woke" and "political" criticize the "agenda pushing" and "virtue signaling" that many assume to be implicit in any show with a diverse cast. Obviously, Hollywood is no stranger subliminal messaging (just think about *Jurassic World: Dominion's* very conspicuous decision to cast the film's "villain" as a Tim Cook lookalike), but just because a show has a diverse cast does not mean it is making any particular criticism of certain groups of people and their beliefs. Personally, I do not feel *Ms. Marvel* did.

Overall, it seems that the language profile of negative reviews focused mainly on the racial and religious identities of the show's main characters and its perceived political ideology, as well as, among a smaller portion of users, deviations from the comics. Positive reviews on the other hand focused more on the show's content, namely its story and actors. This is as close to proving review-bombing we can get: we have identified a large group of users who is giving the show terrible scores despite critics (and many other users) rating it extremely highly, a coordinated effort among this group of users to down-vote other positive reviews, as well as a bias that they share. We should also note that the only other

Marvel TV show with a largely non-white cast, *Falcon and the Winter Soldier*, exhibits many of the same statistical trends we have found to be associated with review-bombing in *Ms. Marvel*. However, they are not as extreme as they are in *Ms. Marvel*, and *Falcon and the Winter Soldier* does not have the same robust baseline in terms of critic score. Interestingly, it does not exhibit the same downvoting-of-positive-reviews effect *Ms. Marvel* has either.

What statistical conclusions can we ultimately draw?

- Critics liked the show a lot
- Some users liked the show a lot, but more hated it a lot
- *Ms. Marvel* is the most polarizing TV show Marvel has released to date
- A group of users banded together to purposefully downvote positive reviews and upvote negative ones, unlike most other shows where the opposite happens
- Themes of race, religion, and the show's politics / ideology are highly cited reasons for low scores
- Taken together, these factors paint a highly suggestive picture of review-bombing
- This review-bombing appears to be an attack on the various ways in which the show is more diverse and different than other Marvel motion pictures

Why I did this project

To be honest, the motivation I provided at the outset of this journey was a little disingenuous. After all, we went through a bunch of statistical overkill to "prove" something that could be divined with a little bit of common sense. I mean, Ms. Marvel had a one star rating on Rotten Tomatoes within three minutes of its release [9], not nearly enough time for anyone to have watched a substantial part of the 50 minute episode. And this reaction isn't new either – it mirrors the exact same one that happened several years ago when the comics were first released [7], before Ms. Marvel was able to win hearts and minds. Moreover, the point really wasn't to change anyone's opinion on the topic either (although if it did, that's great!). As you can tell, the people review-bombing the show are deeply entrenched in some pretty pernicious biases, and so I doubt any of the data-driven insights we've discovered here today will mean anything to them, or will inspire them to elevate their own discourse to a commensurate level of accuracy. So what was the point then? Well, it was really just a passion project of mine. It combined my interest in data science with my own strong reaction to the show, and the way I related to many of its themes as an American-born Indian. Plus, the statistician in me wanted to see if I could get any pretty plots out of this! So if anybody needs some evidence to back up their claims of review-bombing, you now have a good 50-or-so pages worth, with lots of colorful pictures!

Additionally, the framework used here is very general and could be a useful tool for analyzing reactions to Marvel movies and future Marvel shows as well. I encourage you to dig into the code, modify it to your own needs, and see what else you can discover! Make sure to let me know, or send any other thoughts, opinions, and advice my way! **The easiest way to send me your comments is probably through the corresponding blog post at:** https://mehta-rohan.com/projects/ms_marvel.html.

References

- [1]: https://www.yahoo.com/video/ms-marvel-praised-fans-flooded-140258354.html
- [2]: https://in.ign.com/ms-marvel/173635/news/ms-marvel-is-the-highest-rated-mcu-project-ever-on-rotten-tomatoes-despite-the-review-bombing
- [3]: https://www.looper.com/928487/ms-marvel-just-claimed-a-shocking-mcu-title/
- [4]: https://www.nme.com/news/tv/ms-marvel-directors-respond-to-review-bombing-3248481
- [5]: https://www.newsweek.com/ms-marvel-mcu-disneyplus-review-bombing-racist-white-nonsense-1714538
- [6]: https://screenrant.com/ms-marvel-show-review-bomb-backlash-iman-vellani/
- [7]: https://www.digitalspy.com/tv/ustv/a40563393/ms-marvels-iman-vellani-responds-disney-marvel-show/
- [8]: https://tribune.com.pk/story/2361695/ms-marvel-is-being-review-bombed-on-imdb-for-its-south-asian-muslim-superhero
- [9]: https://www.indiatimes.com/explainers/entertainment/explained-what-is-review-bombing-why-is-ms-marvel-at-the-receiving-end-of-this-practice-571956.html
- [10]: https://www.koimoi.com/hollywood-news/ms-marvel-is-now-mcus-highest-critically-rated-series-on-rotten-tomatoes-takes-over-agents-of-s-h-i-e-l-d-wandavision-others/
- [11]: https://twitter.com/angryjoeshow/status/1534239532719263747?lang=en