Supplementary Materials for **Inferential Tasks as an Evaluation Technique for Visualization**.

The contents of this document provide supplementary material for our validation experiment presented in Section 3 of the paper.

Procedure: Our experiment followed a procedure approved by Tufts University's IRB, and was posted as a HIT on Amazon's Mechanical Turk. After accepting the HIT, participants provided informed consent, then were required to watch a 5 minute tutorial video of the tool they would use. After the video, participants were given a training task using the *cars* dataset, which had to be answered correctly before proceeding. In addition to submitting answers for each inferential task, participants were asked to describe the relationship they observed in each generated visualization (i.e., their candidate observations) and to report confidence in their answer ranging from 1 (very low) to 7 (very high). Participants were compensated $3 for the experiment and were given a $2 bonus for completing both tasks and passing an attention check. We recruited 145 participants, and each completed both inferential tasks using either Polestar or Voyager 2. Both interfaces were embedded as iFrames for the experiment. Before analysis, we filtered data from participants who either did not finish both tasks ($N = 29$), or did not pass an attention check during task completion (any participants who completed the tasks in $t < 30$ seconds). For Voyager 2, we filtered 31 and 42 responses from Tasks 1 and 2, respectively. For Polestar, we filtered 26 and 33 responses from Tasks 1 and 2, respectively. An example of the experimental interface is shown in Figures 1 and 2. Demographics are provided in Table 1.

Ground Truth: Ground truth for both tasks was determined by identifying data attributes that, when plotted, displayed a clearly similar relationship to the original visualization described in each task. For Task 1 (Figure 3a), a correct answer was identified by selecting a data attribute that, when plotted with IMDB Rating, showed a similarly logarithmic relationship to what is seen in the visualization of IMDB Rating and IMDB Votes (Figure 3b). The possible correct answers were *US DVD Sales* (Figure 3c) or *US Gross* (Figure 3d). For Task 2 (Figure 4a), a correct answer was identified by choosing two different attributes that, when plotted, displayed a positive linear relationship as US Gross and Worldwide Gross (Figure 4b). The only correct answer was the combination of Rotten Tomatoes and IMDB Rating (Figure 4c).

Examples of incorrect answers are seen in Figure 5. We note that Figures 5a and 5b are visually similar to the visual target for Task 1 (Figure 3b), but we do not distinguish their relation as logarithmic. Therefore, we considered both responses to be incorrect answers for Task 1 prior to analysis. In general, a researcher is responsible for deciding if any given proposed solution (or possible answer) is valid for the evaluation. Correct and incorrect answers should be supplied in the study write-up, with a justification for what is considered correct or not. The original Voyager 2 tool can be used to assess all possible answers for our described experiment using: `https://vega.github.io/voyager2/`.

Reported Observations: After completing each inferential task during our Voyager 2 vs. Polestar validation study, participants were asked to optionally *describe the relationship* in a free-form text box as seen in the visualization (i.e., their candidate observation, as described in Section 2 of the paper). We list out provided responses (as they appear) to both tasks for each tool here:

Voyager 2, Task 1: "*Plot IMDB Votes and IMDB Rating. Observe the relationship of these 2 variables. Find another variable that shows a similar relationship with IMDB Rating. Describe the relationship.*"

1. They both show ratings in an upward trend.
2. The higher the rating the better the score.
3. US DVD Sales is similar in relation to IMDB Votes because they both increase identically as IMDB Ratings increase.
4. It seems that Worldwide Gross comes closest to sharing a similar relationship with IMDB Rating, as least as far as I can read the charts.
5. they both have a high density in the lower range of x with fewer points as x increases.

6. The votes remain low for the most part, but the larger the rating, the larger the total votes.
7. The ratings are higher with the larger budgets.
8. As with votes, US and worldwide gross increased the higher the rating.
9. A positive quadratic relationship
10. The greater the worldwide gross, the more likely the movie will have a higher rating, which is consistent with the number of votes corresponding to higher ratings.
11. Generally speaking, movies with a higher US Gross have a higher IMDB Rating.
12. There appears to be a similar correlation between receiving a higher IMDB Rating, leading to a higher US Gross, as both appear to rise alongside each the other.
13. As the IMDB rating increases, the number of votes and the US gross tends to increase as well, with a dramatic increase as the rating goes over 6.
14. Rotten tomatoes is similar. As the number of votes goes up, the rating is higher.
15. As the IMDB rating increase, the IMDB votes increases.
16. It seems that IMDB ratings are most closely related to Rotten Tomatoes ratings in that the higher the vote cote the higher the rating.
17. To me it looks like the Rotten Tomato rating correlates fairly well with the IMDB rating. When one is high, the other is high. I think that makes sense since they rate the same thing (movies) and each sample size is large enough that there won't be a huge variation in rating.
18. high votes and rating
19. There are alot of movies with few votes and very frew with alot of votes. This seems to agree with the worldwide gross as there are only a few movies with large gross but many with small amounts of gross.
20. IMDB Votes, IMDBRating and Production budget are same. they have similar value.
21. They are both similar in shape, where they both narrow. The others are drastically different.
22. The more IMDB votes, the more specific the IMDB or Rotten Tomatoes rating. The less IMDB votes, the more scattered the ratings.
23. Its get high
24. The Rotten Tomatoes rating is a good indicator of who has got the most votes.

Polestar, Task 1: "*Plot IMDB Votes and IMDB Rating. Observe the relationship of these 2 variables. Find another variable that shows a similar relationship with IMDB Rating. Describe the relationship.*"

1. among the two variables, i think the gross total increases
2. Many of the ratings vary a lot despite having a low Us Gross. As the US gross increases, the IMDB rating increases.
3. Most of the votes are between 0k and 100k with not much dispersing after 100k
4. lower IMDB number of votes have greater number of Rotten Tomatoe reviews
5. The Product budget reduces with IMDB Votes.
6. there is a cluster in the middle with few clusters at the extreme top and bottom for rating
7. IMDB Ratings and Rotten Tomatoes ratings are similar; generally, the more votes, the higher the rating
8. rotten tomatoes ratings and imdb ratings have a similar distribution per IMDB votes
9. Higher the IMDB Rating, higher the IMDB Votes and US Gross.
10. The more votes there are, the higher the mean ratings get for that number of votes.
11. The two components, Votes and US Gross, generate a similar chart when compared with the rating

12. The gross shows a similar outcome
13. This shows the relationship between the movie's cost and ratings.
14. Generally The more votes a movie has the better its rating.
15. the higher the votes, the higher the rating from imbd and rotten tomatoes
16. They both go up together.
17. The more reviews it seems like the more likely the film is to be highly rated.
18. US Gross has the majority of values falling between 5.5 and 7.5, just like IMDB Votes and also has outliers on the x axis that seem to occur near a 7 on the IMDB Rating.
19. There is large cluster at the lower end and hten it tapers out
20. High IMDB rating usually means higher US dvd sales?
21. most similar on the top end of the graph but none of them are close to identical
22. It clusters at the start but then goes up.

Voyager 2, Task 2: "*Plot US Gross and Worldwide Gross. Observe the relationship of these 2 variables. Find two different variables that show a similar relationship. Describe the relationship*"

1. as the rating increases so does the dvd sales in the us
2. As Rotten Tomatos Rating increases, IMDB rating also tends to increase.
3. These both have a gradual increase as one thing increases, the other increases.
4. As production budget increase, DVD sales increase
5. The charts are simplified but they show the same relation. They are both showing the more it grossed the higher it was ranked.
6. The ratings for movies using both of these provides similar findings.
7. they have similar relationship like rating and voting from viewers
8. A positive linear relationship
9. The higher the Production Budget usually means a higher rating and better overall gross.
10. the higher the budget the higher the ratings
11. its high
12. only a few movies surpass 200 million dvd sales
13. high rating
14. They are both growing is what I'd say and how it's similar to the first two variables.
15. These both have a gradual increase as one thing increases, the other increases.

Polestar, Task 2: "*Plot US Gross and Worldwide Gross. Observe the relationship of these 2 variables. Find two different variables that show a similar relationship. Describe the relationship*"

1. This is similar to the one given as the votes decreases as the production budget increases.
2. Generally clumped but the rating increases as the dvd sales increases.
3. they both seem to have an increasing pattern
4. As the production budget increases, US DVD sales increase as well
5. Generally, the higher the production budget, the higher the DVD sales
6. These two variables produce a similar chart to the previous one, but I don't see why having just more votes, despite of the rating, produces more DVD sales
7. increase running time receives high IMDB votes

8. both trend up and to the right
9. There are several values at the start of the graphic but the values grow more disperse as the values increase.
10. higher the IMDB Rating, higher the Rotten Tomatoes Rating
11. As production budget increases the number of sales increases
12. US gross and wordwise gross are two ways of measuring the same thing. Under that same vein of thought imbd and rotten tomatoes are both measuring the same thing
13. There is a tight cluster in the front and then they move on
14. amount of sales in the US versus the amount of votes on IMBD is somewhat similar
15. The higher the IMDB rating the more likely the Rotten tomato rating would also be high.
16. Movies with high IMDB Votes has sold less in millions

We note that the number of submitted candidate observations is less than the total number of participants from our experiment, particularly for Task 2 compared to Task 1 (perhaps due to task complexity or fatigue). For future experiments, and if necessary for the evaluation, we suggest requiring these responses during Stage Two of the task, especially if assessing the accuracy of the task after collecting responses or combatting HARKing. While we supply participants with freeform text boxes, it would also be possible to provide multiple choice answers or predefined candidate observations that can be selected by the participant.

Quantitative Results: We discuss our results with respect to the original Voyager 2 experimental goal, "Does Voyager 2 support breadth-oriented and depth-oriented analysis?"[WQM*17]. We find participants are significantly more accurate with Voyager 2 for Task 1 ($\chi^2(1,\ N = 59) = 5.52,\ p = 0.01$), but find no significant difference in accuracy for Task 2 ($\chi^2(1,\ N = 41) = 2.35,\ p = 0.125$). Completion time with Voyager 2 was also significantly faster for both tasks: (Task 1: ($W(25) = 0.83,\ p = 0.001$); Task 2: ($W(27) = 0.81,\ p = 0.001$)). Therefore, we find Voyager 2's ability to support depth-oriented analysis outperforms Polestar's with Task 1. We find no significant difference between Voyager 2 and Polestar's ability to support breadth-oriented analysis in Task 2. For this reason, we cannot quantitatively conclude that Voyager 2 supports breadth-oriented analysis better than Polestar. Finally, participants rated confidence in their answers on a Likert scale of 1 (very low) to 7 (very high). In general, Polestar participants reported higher confidence ($\mu = 5.06,\ \sigma = 0.08$) than Voyager 2 participants ($\mu = 4.7,\ \sigma = 0.13$), despite Voyager 2 participants generally outperforming Polestar participants.

The fact that participants are more accurate with Voyager 2 suggests that the recommended visualizations are of high quality and the design of the visualization successfully supports participants in exploring the recommendations. Furthermore, participants are significantly faster across both tasks using Voyager 2 compared to participants using Polestar, indicating that the recommended visualizations can aid participants in completing their tasks more efficiently than without recommendations. Interestingly though, Voyager 2 participants reported a lower confidence rating in their responses to both tasks when compared to Polestar participants. We theorize that Voyager 2 participants may have felt less sure of their answers after being exposed to many possible solutions via the system's recommendation engine. Due to the addition of optional free-response questions ("describe the relationship", listed above), we are additionally able to assess participants' insights generated while performing each task. We find that both tools are capable of supporting participants through exploratory and confirmatory analysis.

In addition to producing comparable results to the original Voyager 2 think-aloud study, our validation experiment was fast to run and scalable. Although the total cost of our experiment outweighed the original (Ours: \$3 - \$5 per person; Original: \$15 per person), we were able to increase the participant pool from 16[WQM*17] to 145 by remotely running the experiment on a crowdsourced platform. Furthermore, analysis of our quantitative results required no interviewing, as done in the original Voyager 2 study.

Measure	Counts
N	145
Age	18-25: 4.1%, 26-35: 35.2%, 36-45: 18.6%, 46-55: 4.3%, 56-65: 2.8%, Prefer not to specify: 34.5%
Gender	Female: 22.8%, Male: 42.1%, Prefer not to specify: 39.5%
Education	High School: 6.9%, Some college: 8.9%, Associate's: 5.5%, Bachelor's: 37.9%, Graduate: 8.9% Prefer not to specify: 31.7%

Table 1: Demographics for our Voyager 2 vs. Polestar study.

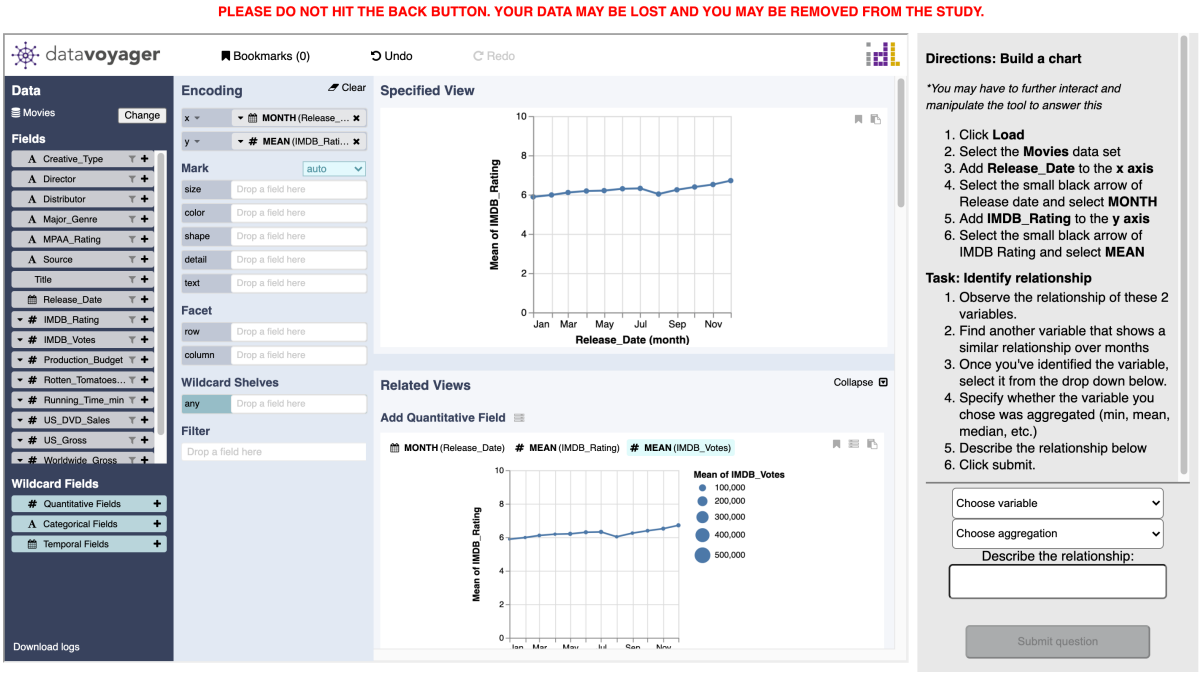


Figure 1: An example of the Voyager 2 interface used in our experiment, encoded as an iFrame.

References

[WQM*17] Wongsuphasawat K., Qu Z., Moritz D., Chang R., Ouk F., Anand A., Mackinlay J., Howe B., Heer J.: Voyager 2: Augmenting visual analysis with partial view specifications. In *ACM Human Factors in Computing Systems (CHI)* (2017). URL: `http://idl.cs.washington.edu/papers/voyager2`.

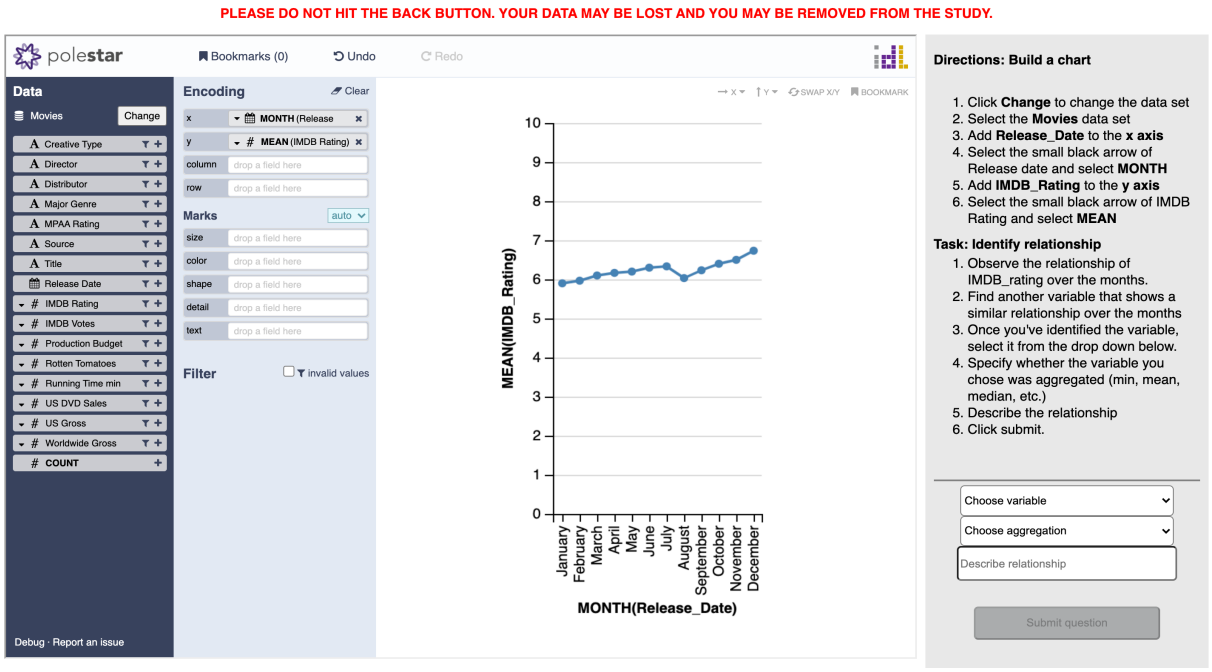
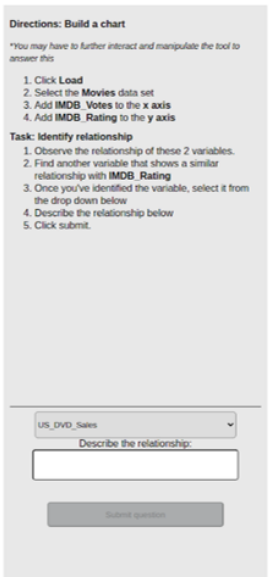
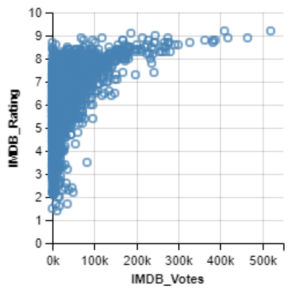


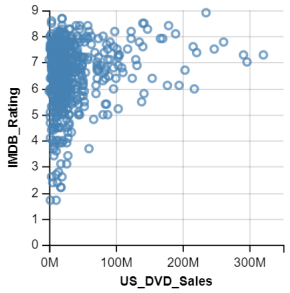
Figure 2: An example of the Polestar interface used in our experiment, encoded as an iFrame.



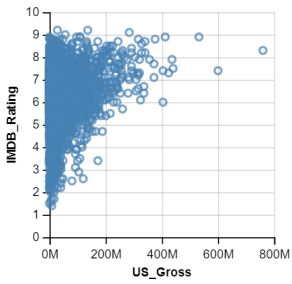
(a) Task 1 prompt



(b) (Visual target) IMDB Rating and IMDB Votes

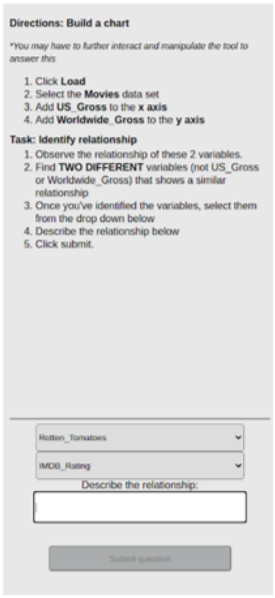


(c) (Correct) US DVD Sales

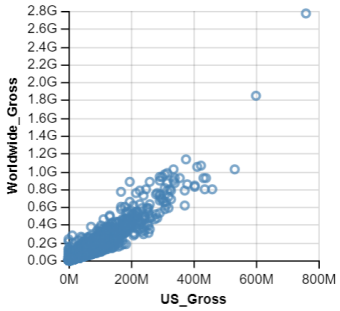


(d) (Correct) US Gross

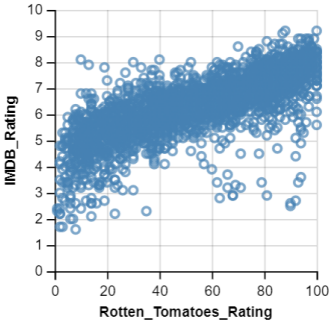
Figure 3: (a) Task 1's prompt, (b) the visualization that must be generated by participants (i.e., the *visual target*), (c-d) the possible correct answers to the task.



(a) Task 2 prompt

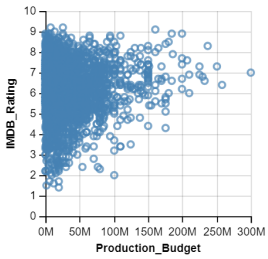


(b) (Visual target) US Gross and Worldwide Gross

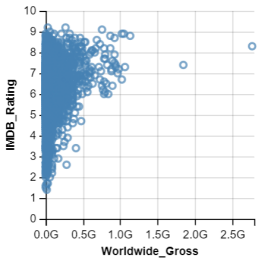


(c) Correct: Rotten Tomatoes and IMDB Rating

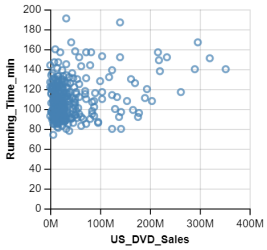
Figure 4: (a) Task 2's prompt, (b) the visualization that must be generated by participants (i.e., the *visual target*), (c) the only possible correct answer to the task.



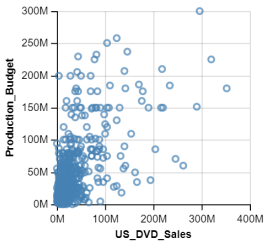
(a) Task 1 incorrect



(b) Task 1 incorrect



(c) Task 2 incorrect



(d) Task 2 incorrect

Figure 5: (a-b) Examples of incorrect answers for Task 1. Although both examples look visually similar to Figure 3b, we do not distinguish them as logarithmic. (c-d) Examples of incorrect answers for Task 2.