IMDb Movie Success: An Analysis of Movie Release Factors and Box Revenue

The Four-mula: Hellen Han, Lila Rogers, Amy Duan, Eva Aggarwal

2025-03-20

Introduction:

This project aims to examine key factors influencing IMDb movie revenues to understand what drives a movie's box office success. Revenue, a widely used measure of success in the film industry, reflects audience demand and commercial viability. Revenue is defined as the total amount of money a given movie generates from all sources related to the film. Prior research has highlighted various influences, such as star power, genre, and marketing, but the relative importance of these factors remains debated.

One central research question our project aims to answer is: **What production and release factors have the greatest impact on a movie's total revenue?** By addressing this question, we aim to not only deepen the understanding of revenue drivers and gain valuable insights into IMDb movie success, but also improve predictive models for movies' box office performance in the future. We hypothesize that IMDb movies with higher budgets, higher average ratings, and higher vote counts tend to have higher revenues.

We obtained our data set from Kaggle, an online data science platform with a collection of community-developed open data sets. This data was collected by **Anand Shaw** from the **IMDb website** using various IMDb sites, and converted into a .csv file. The data was updated on a daily basis until 2 months ago.

Exploratory Data Analysis:

To get more indepth into our data set, the data set we used collects information available on the IMDb website for different movies, such that each observation describes characteristics of a specific movie. In general, the characters being measured follow basic information about the movie, various classifications of the movie's popularity and rating, and the monetary values associated with the movie. It measures the following **15 characteristics** per movie:

title: The name of the movie.

vote_average: The average rating the movie has received from users (on a scale, typically from 0 to 10).

vote_count: The total number of votes or ratings submitted for the movie.

status: The current state of the movie (e.g., “Released,” “Post-Production”).

release_date: The date when the movie was officially released.

revenue: The total earnings the movie made (usually in USD).

runtime: The duration of the movie in minutes.

adult: Indicates whether the movie is classified as adult content (e.g., “True” or “False”).

budget: The total cost of producing the movie (usually in USD).

imdb_id: The unique identifier for the movie on IMDb (Internet Movie Database).

original_language: The language in which the movie was originally produced (e.g., “en” for English).

popularity: A metric indicating how popular the movie is (typically based on views, searches, or ratings).

genres: The categories or genres the movie belongs to (e.g., Action, Comedy, Drama).

production_countries: The countries where the movie was produced.

spoken_languages: The languages spoken in the movie.

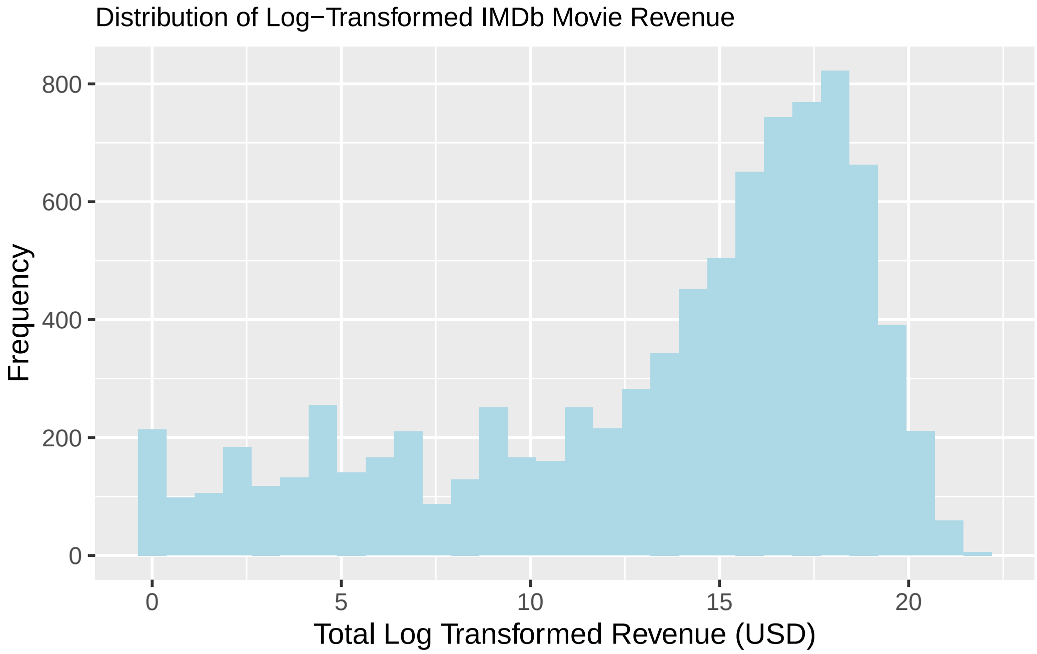
Due to the large size of the original data set and for the purpose of uploading this to our repository, we intentionally drop some characteristics, including id, original_title, tagline, production_companies, overview, keyword from the data set due to their redundancy, irrelevance to the research question and presence of a significant number of null values.

Additional data cleaning was performed for the scope of this project. To obtain informative interpretations of the distribution of the revenue response variable in terms of recency, we analyze IMDb movies released from 2000-present. Missing values and unnecessary fields were removed, and variables names and units were standardized. Furthermore, we log transform revenue and budget, as the revenue generated and budget allocated across the IMDb movies are heavily skewed. A categorical vote_category variable was created from the vote_average variable to reflect the relative average rating category for the movie, creating levels “Low” for ratings 0-3, “Medium” for ratings 3-7, and “High” ” for ratings 7-10. Furthermore, a release_years_since_2000 variable was created that calculates the number of years of a movie’s release since 2000, excluding all movies that were released pre-2000.

We chose to examine movie revenue as our response variable.

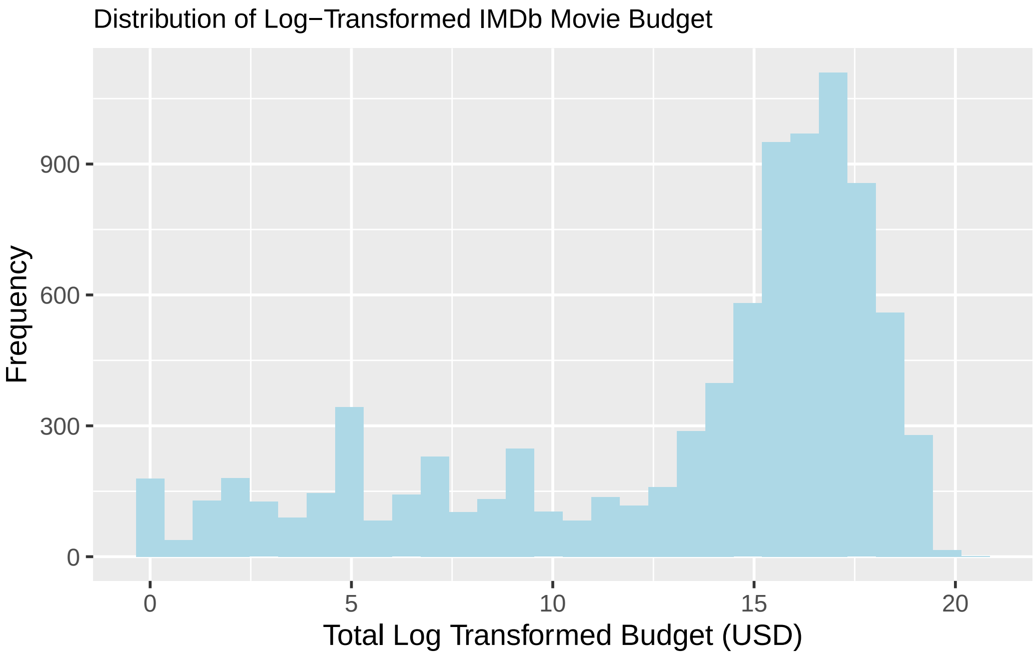
Univariate EDA

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   9.933  15.323  13.387  17.646  21.822 
```



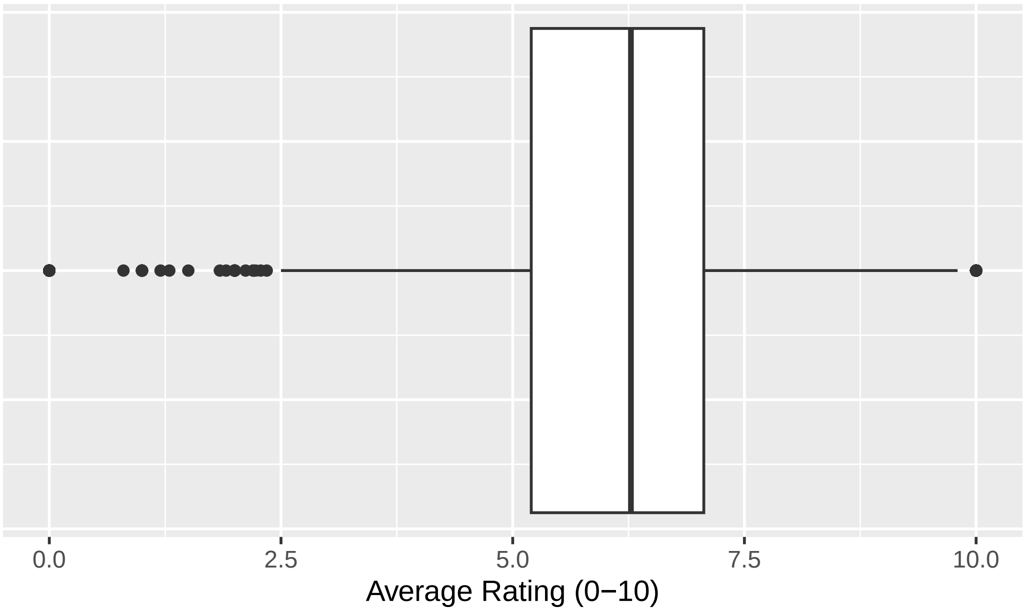
The histogram shows the distribution of the log transformed IMDb movie revenues response variable, which is left-skewed and unimodal, indicating that on average, movies listed on IMDb tend to generate higher revenues. Furthermore, the distribution of the log transformed IMDb movie revenues has a center of approximately 1,800,000 US dollars, described by the median, and a spread of approximately 18,600,000 US dollars, described by the interquartile range. The majority of IMDb movies generate between approximately 24,100 and 18,600,000 US dollars.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   9.629  15.425  13.222  17.034  20.500 
```



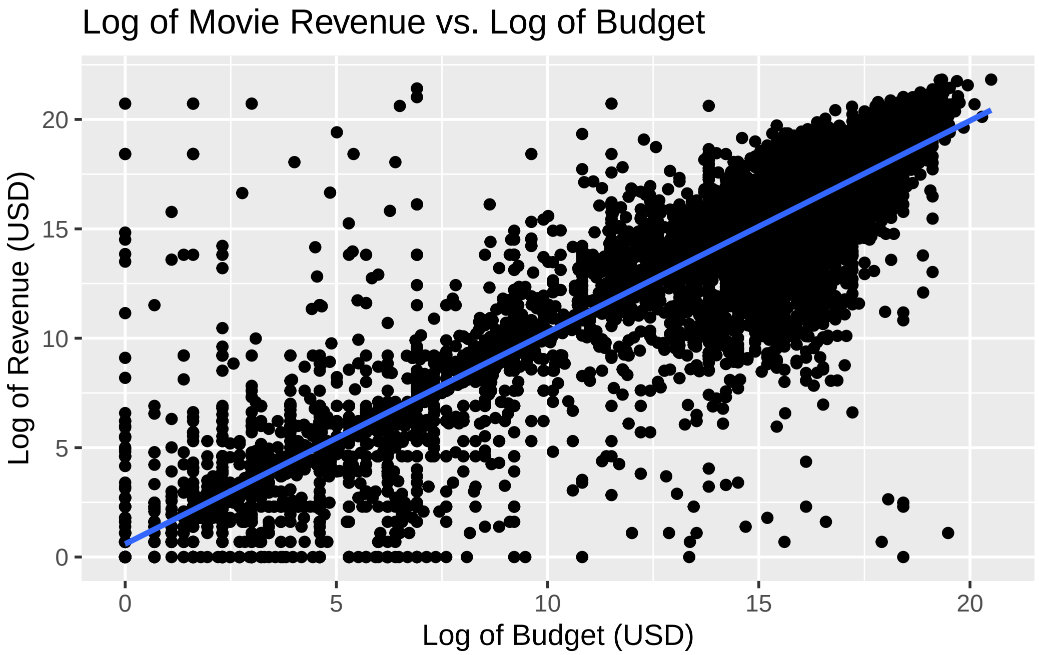
The histogram shows the distribution of the log transformed IMDb movie budget predictor variable, which is left-skewed and unimodal, indicting that although IMDb movie budget varies greatly, on average, movies listed on IMDb tend to have higher budgets. Furthermore, the distribution of the log transformed IMDb movie budget has a center of approximately 5,000,000 US dollars, described by the median, and a spread of approximately 25,000,000 US dollars, described by the interquartile range.

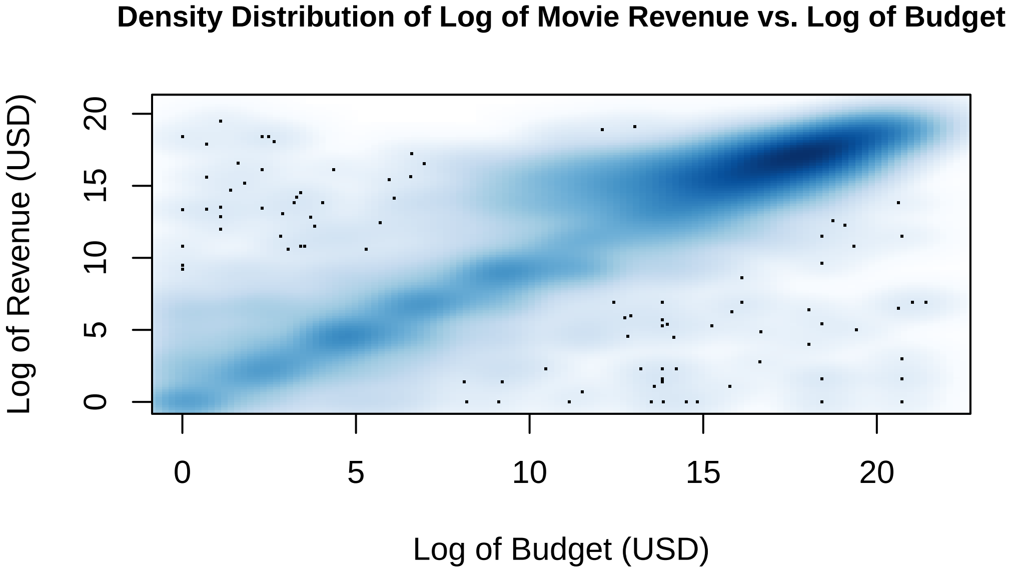
Distribution of IMDb Average Rating



The majority of IMDb movies' average ratings are concentrated between a rating of 4.0 to 8.0. A few IMDb movies have a low rating (average rating below 2.5), contributing to the left-skewness of the distribution of IMDb movies' average rating predictor variable. Furthermore, a few IMDb movies have an extremely high rating (average rating of 10.0).

Bivariate EDA


Log of Movie Revenue vs. Log of Budget

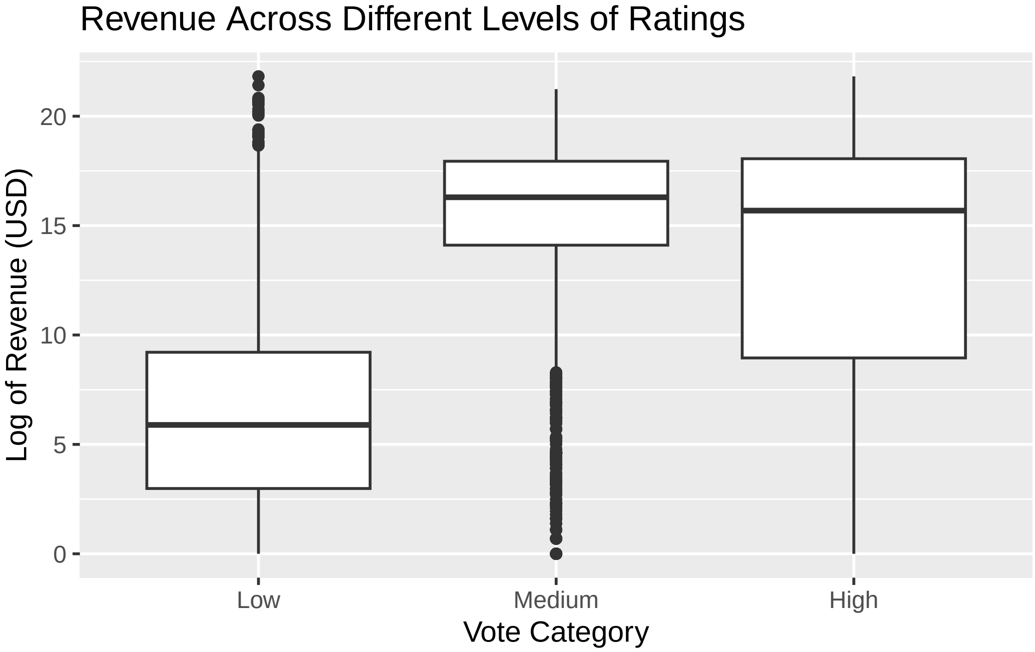

Density Distribution of Log of Movie Revenue vs. Log of Budget

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
```

```
1 (Intercept)    0.595   0.0649       9.16 6.08e-20
2 log_budget     0.967   0.00456    212.   0
```

```
[1] 0.8365862
```

From our visualization plotting the log transformed movie revenue versus the log transformed movie budget, as well as our relatively high R^2 value associated with the linear regression model, we can identify that there is a strong linear correlation between the log transformed movie revenue and log transformed movie budget. The p-value is effectively 0, further exemplifying that there is a linear relationship between the log transformed movie revenue and the log transformed movie budget. This is expected, as high budget movies are often more anticipated and therefore, more people tend to purchase tickets. Because our data set is so large, this relationship is also displayed using a smoother scatter visualization, which illustrates point density seen in our scatter plot.



From our boxplot, we can see that the median revenue for Medium rated movies (approximately 8,880,000 dollars) is actually slightly higher than those with High ratings (approximately 5,400,000 dollars). The median revenue for Low rated movies is significantly lower (approximately 400 dollars). The IQR of movie revenues with High ratings is also much larger than for Medium and Low rated movies.

Potential Interaction Effects

To check interaction effects between some of our predictors to ensure we don't have issues with multicollinearity in our models, we checked interaction effects between some variables that are likely to have some collinearity. For 3 pairs of predictor variables, we fit an interaction model with revenue as the response variable to see if there is a statistically significant impact of the interaction term on the model.

Budget and Popularity

Budget and popularity are likely to have interaction effects because higher-budget movies often receive more marketing, leading to increased visibility and higher popularity. However, the p-value for the interaction term (budget:popularity) is 0.48, indicating that the effect is not statistically significant at the 0.05 level, meaning budget's impact on revenue does not significantly change based on popularity.

term	estimate	std.error	statistic	p.value
(Intercept)	-4711433.669	1415969.481	-3.327	0.001
budget	2.975	0.031	95.869	0.000
popularity	84395.635	28855.053	2.925	0.003
budget:popularity	0.000	0.000	0.706	0.480

Vote Average and Vote Count

Since movies with higher vote counts often also have higher vote averages, there could be a multiplicative effect on revenue. If an interaction is present, it means that vote count alone does not fully explain revenue — its impact depends on the vote average. And, the p-value for the interaction term (vote_average:vote_count) is 2.331652e-45, indicating that the effect is extremely significant at the 0.05 level, meaning there is very strong evidence that the relationship between vote count and revenue depends on the vote average. So, there is a very strong interaction effect between these two variables.

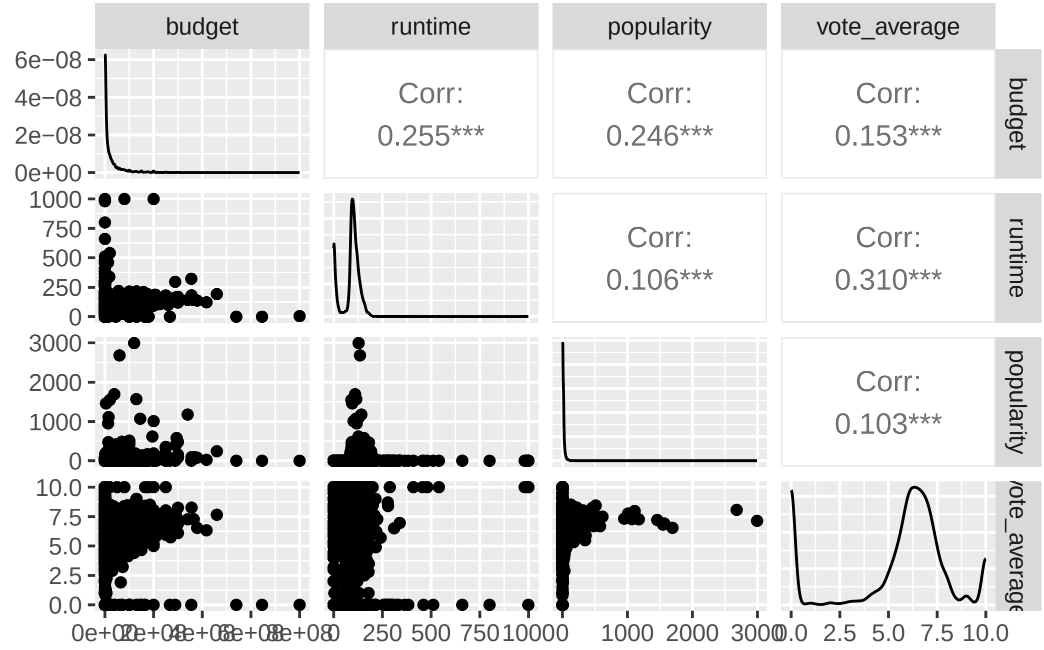
term	estimate	std.error	statistic	p.value
(Intercept)	10071337.816	2855919.960	3.526	0.000
vote_average	-1527973.810	467221.817	-3.270	0.001
vote_count	102244.575	4248.822	24.064	0.000
vote_average:vote_count	-8149.997	573.312	-14.216	0.000

Action Movies and Runtime

Action movies have longer runtimes typically compared to other genres due to storytelling, long fight sequences, and action scenes. We thought that there may be an interaction effect between being an action movie and runtime when predicting revenue. Longer runtimes could allow for larger audiences and increasing box office earnings. However, excessively long movies may also reduce viewers due to time constraints. However, from the p-value of the interaction term, which is 2.991935e-15, we can see that this term is quite significant at the 0.05 level. This implies that there is some interaction between a movie having the 'action' genre and longer runtimes, or the effect of runtime on revenue is different for action movies compared to non-action movies. This could suggest that longer runtimes contribute positively to revenue for action films, possibly because audiences expect more elaborate fight sequences and storytelling in the genre, making them more willing to engage with longer films.

term	estimate	std.error	statistic	p.value
(Intercept)	13916205.3	3864760.77	3.601	0.000
action_movie	-892455.3	9145331.04	-0.098	0.922
runtime	372326.6	38529.42	9.663	0.000
action_movie:runtime	630145.9	79709.91	7.905	0.000

We also explored the correlation between all possible combinations of the variables budget, runtime, popularity and vote_average, to see how they interact with each other. The results indicate that the correlations between these variables are relatively low, with none exceeding 0.310. This suggests that there is no strong linear relationship between them, reducing the likelihood of multicollinearity issues when including these variables in a predictive model. As a result, we can confidently incorporate them into our analysis.



Methodology

This section includes a brief description of your modeling process. Explain the reasoning for the type of model you're fitting, predictor variables considered for the model including any interactions. Additionally, show how you arrived at the final model by describing the model selection process, interactions considered, variable transformations (if needed), assessment of conditions and diagnostics, and any other relevant considerations that were part of the model fitting process.

Results

Describe the key results from the model. The goal is not to interpret every single variable in the model but rather to show that you are proficient in using the model output to address the research questions, using the interpretations to support your conclusions. Focus on the variables that help you answer the research question and that provide relevant context for the reader.

Discussion

In this section you'll include a summary of what you have learned about your research question along with statistical arguments supporting your conclusions. In addition, discuss the limitations of your analysis and provide suggestions on ways the analysis could be improved. Any potential issues pertaining to the reliability and validity of your data and appropriateness of the statistical analysis should also be discussed here. Lastly, this section will include ideas for future work.

Conclusion

> **!** Important
>
> Before you submit, make sure your code chunks are turned off with `echo: false` and there are no warnings or messages with `warning: false` and `message: false` in the YAML.