IMPORTANT: If the analysis is completed using software other than R, or not written up using R Markdown, the project should receive a 0 regardless of its content.

Part 1: Data (2 points)

- 1 pt for correct reasoning for generabizability – Answer should discuss whether random sampling was used. Learners might discuss any reservations, those should be well justified.
- 1 pt for correct reasoning for causality – Answer should discuss whether random assignment was used.

Part 2: Data manipulation (10 points)

- Create new variable based on `title_type` : New variable should be called `feature_film` with levels yes (movies that are feature films) and no (2 pt)
- Create new variable based on `genre` : New variable should be called `drama` with levels yes (movies that are dramas) and no (2 pt)
- Create new variable based on `mpaa_rating` : New variable should be called `mpaa_rating_R` with levels yes (movies that are R rated) and no (2 pt)
- Create two new variables based on `thtr_rel_month` :
    - New variable called `oscar_season` with levels yes (if movie is released in November, October, or December) and no (2 pt)
    - New variable called `summer_season` with levels yes (if movie is released in May, June, July, or August) and no (2 pt)

Part 3: EDA (9 points)

Conduct exploratory data analysis of the relationship between `audience_score` and the new variables constructed in the previous part

- 3 pts for plots
    - Plots should address the research questions (1 pt)
    - Plots should be constructed correctly (1 pt)
    - Plots should be formatted well – size not too large, not too small, etc. (1 pt)
- 3 pts for summary statistics
    - Summary statistics should address the research questions (1 pt)
    - Summary statistics should be calculated correctly (1 pt)
    - Summary statistics should be formatted well – not taking up pages and pages, etc. (1 pt)
- 3 pts for narrative
    - Each plot and/or R output should be accompanied by a narrative (1 pt)
    - Narrative should interpret the visuals / R output correctly (1 pts)
    - Narrative should address the research question (1 pts)

Part 4: Modeling (15 points)

Develop a Bayesian regression model to predict `audience_score` from the following explanatory variables. Note that some of these variables are in the original dataset provided, and others are new variables you constructed earlier:

- `feature_film`
- `drama`

- `runtime`
- `mpaa_rating_R`
- `thtr_rel_year`
- `oscar_season`
- `summer_season`
- `imdb_rating`
- `imdb_num_votes`
- `critics_score`
- `best_pic_nom`
- `best_pic_win`
- `best_actor_win`
- `best_actress_win`
- `best_dir_win`
- `top200_box`

Complete Bayesian model selection and report the final model.

- Carrying out the model selection correctly (5 pts)
- Model diagnostics (5 pts)
- Interpretation of model coefficients (5 pts)

Prediction (5 points)

Pick a movie from 2016 (a new movie that is not in the sample) and do a prediction for this movie using your the model you developed and the `predict` function in R.

- Correct prediction (4 pts)
- Reference(s) for where the data for this movie come from (1 pt)

Conclusion (3 points)

A brief summary of your findings from the previous sections **without** repeating your statements from earlier as well as a discussion of what you have learned about the data and your research question. You should also discuss any shortcomings of your current study (either due to data collection or methodology) and include ideas for possible future research.

- Conclusion not repetitive of earlier statements (1 pt)
- Cohesive synthesis of findings that appropriate address the research question stated earlier (1 pt)
- Discussion of shortcomings (1 pt)

Overall (6 points)

- Uploaded HTML document resulting from the Rmd template: 1 pt
- Organization: 1 pts
- Readability of the text: 2 pts
- Readability of the code: 2 pts