Congratulations on getting a job as a data scientist at Paramount Pictures! Please see the Data Analysis Project for your assignment. Below you will find the files that you will need.

Your boss has just acquired data about how much audiences and critics like movies as well as numerous other variables about the movies. This dataset is provided below, and it includes information from Rotten Tomatoes and IMDB for a random sample of movies.

Your boss is interested in learning what attributes make a movie popular. She is also interested in learning something new about movies. She wants your team to figure it all out.

As part of this project you will complete exploratory data analysis (EDA), modeling, and prediction.

The specific modeling task you need to complete is as follows: 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 will need to construct in the data manipulation section using the mutate function in dplyr:

- feature_film: *"yes" if title_type is Feature Film, "no" otherwise*
- drama: *"yes" if genre is Drama, "no" otherwise*
- runtime
- mpaa_rating_R: *"yes" if mpaa_rating is R, "no" otherwise*
- thtr_rel_year
- oscar_season: *"yes" if movie is released in November, October, or December (based on thtr_rel_month), "no" otherwise*
- summer_season: *"yes" if movie is released in May, June, July, or August (based on thtr_rel_month), "no" otherwise*
- imdb_rating

- imdb_num_votes
- critics_score
- best_pic_nom
- best_pic_win
- best_actor_win
- best_actress_win
- best_dir_win
- top200_box

All analysis must be completed using the R programming language via RStudio, and your write-up must be an R Markdown document. To help you get started we provide a template Rmd file below (see Rmd template in the Required files section below). Download this file, and fill in each section.

NOTE: If you have previously completed Course 3 in the Statistics with R Specialization (Linear Regression and Modeling) you should be familiar with this dataset. While the dataset is the same, the tasks you are asked to carry out for this project are different.

IMPORTANT: Analyses completed using software other than R, or not written up using R Markdown, will receive a 0 on the project regardless of their content.

Required files

- Data - Save this file in the same directory as the Rmd template (provided below).

NOTE: If you are using Chrome as your browser you might need to change the .gz at the end of the extension to .Rdata in the file you downloaded.

```
movies.Rdata
```

- Codebook - Review this file to find out what each column in the data represents.

```
movies_codebook.html
```

- Rmd template - You must use this template to write up your project. Save the data and this file in the same directory.

```
bayesian_project.Rmd
```

- Assessment rubric - You might want to review the assessment rubric while working on your project so that you have some idea of how your peers will evaluate your work. Please review this carefully before starting your project and refer to it regularly as you are working on your project.

bayesian_project_rubric.html

More information on the data

The data set is comprised of 651 randomly sampled movies produced and released before 2016.

Some of these variables are only there for informational purposes and do not make any sense to include in a statistical analysis. It is up to you to decide which variables are meaningful and which should be omitted. For example information in the the `*actor1*` through `*actor5*` variables was used to determine whether the movie casts an actor or actress who won a best actor or actress Oscar.

You might also choose to omit certain observations or restructure some of the variables to make them suitable for answering your research questions.

When you are fitting a model you should also be careful about collinearity, as some of these variables may be dependent on each other.

Source: Rotten Tomatoes and IMDB APIs.

More information on model selection

You may choose to use any of the Bayesian model selection techniques presented in this course, however you should justify your choice. Note that there are many other model selection techniques that are beyond the scope of this course, and those should not be used in this project.

Note that you have a very specific task on hand: predict audience_score based on the explanatory variables listed above. Also note that you first need to create some of these explanatory variables based on existing variables in the dataset.

Frequently Asked Questions

Do I have to use R for my project? Yes. While there are other statistical packages and/or programming languages that may be perfectly appropriate for your project, since one of the goals of this course is to learn R, all analysis **must** be completed in R

and using the Rmd template provided above. Projects completed using other statistical packages and/or programming languages will receive a 0 on the project.

Where can I find a list of R commands that might be useful for the project? Refer to the previous labs and see the RStudio cheatsheets for dplyr, ggplot2, and RMarkdown.

Who am I writing for? Write as if you are explaining your results to whomever would be interested in your research question, whether this is another scholar in your field or peers sharing your interest in the topic. This audience may not have taken a statistics course. You must be statistically accurate and use correct statistical terminology, but must also explain your conclusions in a way that anyone can understand.

Does my project have to be written in English? Yes, your project must be written in English; this is the only way to ensure that the students who are assigned to review your project can understand it.

What is a peer assessment? Peer Assessment is when students in a course evaluate a fellow student's work. First, each student submits an assignment. Then, the students who have submitted an assignment are given other students' assignments to evaluate, according to provided criteria. Finally, each student receives a grade that is based on the other students' evaluations.

Can I use a paper I've worked on for another course or purpose? No. Please create a unique project for this course. Do not use your master's thesis, work you have published elsewhere or work you have submitted for another course. In the past, students who have submitted work they used elsewhere were reported as submitting plagiarized work.

What if I think the project I am assessing has been plagiarized?

1. Assess the project according to the Evaluation/Feedback directions.
2. Report plagiarism to Coursera https://learner.coursera.help/hc/en-us/articles/209818863-Coursera-Honor-Code

How do I avoid plagiarism?

"In an instructional setting, plagiarism occurs when a writer deliberately uses someone else's language, ideas, or other original (not common-knowledge) material without acknowledging its source." - The Council of Writing Program Administrators.

Therefore, please give credit for all of the sources you have used. Copying and pasting from a site without giving the source is plagiarism, and will be reported. For more information, see this tutorial on avoiding plagiarism. In your own project, give credit to

all sources you used, even if you have paraphrased them or if they are your own work but published elsewhere.

Mark as completed