**Project Description and Grading Rubric**

**NOTE: Most groups will have three members. It is expected that work will be divided fairly among group members. If this is not the case, grades may be reduced for some members. You should be able to work out minor differences in contributions, but if there is serious inequity, please let me or the TA know as early as possible so we can help you work out a working solution. After handing in the project, you will fill out a group evaluation, describing your collaborative process.**

**Proposal:**

I will provide a number of data sets. If you would like to use other data, we can meet to decide if it is appropriate. Sign up for your data set on the google doc (link on Blackboard). Once you have your data, decide on some research questions.

Type a short proposal (one full page, double spaced), explaining your research question, which data you will be using, why your study is interesting, and some general ideas that you would like to analyze. That is, describe some relationships that would like to investigate, graphs that you plan to make, machine learning techniques you are considering. Suggest some of your hypotheses. For example, you may plan to look at data on films to see if newer films make more money as a percent of budget, than older films.

Your proposal should be in a word document, or a pdf. Upload a copy of the proposal to your Project section in Brightspace.

**Draft:**

**After your proposal has been approved**, begin to explore your data set, using R. Following the instructions for Data Analysis and Writeup below, prepare a draft with **all of the parts** noted, **except** the Machine Learning portion. The more finished your draft, the better feedback I can give you before you submit the final version.

The draft should be a pdf file or word document and you should upload it to the Project section in Blackboard.

**The Data Analysis and Writeup:**

Your presentation will have five parts: An Introduction, Data Visualizations, Machine Learning Analysis, Conclusions, and Limitations/Recommendations:

1. **Introduction:** At the beginning of your paper, you must describe the data, in a paragraph. Note the following:

* What is the source of the data? Where and when was it created?
* If it is a sample, from what population was it drawn, and how was the sample selected?
* Do you suspect any sampling bias?
* Was it an experiment or an observational study?
* How were measurements taken, or questions asked?
* Do you suspect any bias in the questions or measurements?
* Why is this data of interest to you, and why should the class find it interesting?
* What kind of data cleaning was necessary (R code for this must show…)

1. **Data Visualizations:** Write R code to create some relevant descriptive graphs, using techniques that we’ve used in class (ggplot, maybe dplyr). About **4 or 5 graphs** should be plenty, depending on complexity.

For each graph and numerical summary, write a paragraph or two summarizing what you see, and suggesting some implications. For example, describe patterns that you observe in a graph, and suggest why they make sense, given what you know about the subject, or if they are unexpected. Do you think there is a cause-effect relationship between any variables? Explain your reasoning.

1. **Machine Learning Methods:** Apply one or more (preferably 2) of the modeling or other machine learning techniques for classification and one or more modelling techniques for regression that we’ve learned in class. Describe your reasons for using the method – why it is appropriate. Describe the conclusions of each particular analysis. Include more visualizations, as needed.
2. **Conclusions:** Write some overall conclusions – an overall summary of what you learned from your visualizations and analysis. Summarize in one paragraph.
3. **Limitations / Recommendations:** Write a paragraph describing some of the limitations that are inherent in your study. Also discuss ideas for future research that might build on the work you did in this project. Summarize in one paragraph.

When you have everything assembled, put your code and text in an Rmd script that will help you present your results. The R code in your Rmd script **must have comments** that explain parts of the R code. The **paper created by your Rmd script must show the R code**, even setup and data transformation code (e.g., do not use include=FALSE). Remove extra warnings and messages, as you did on homework. Text should be written using complete sentences with correct spelling and grammar. Knit the script, and save as a pdf or html (or as a Word document, if you prefer).

**Submit the pdf or html or Word version** by uploading it to the Project section in blackboard. Be sure that your **data** is also uploaded if you’re not using one of the provided data sets. Also upload your **Rmd script.**

**Group Evaluation after submission**

The day after the project is due, go to the Project Description link on Brightspace, and click the link to the Group Evaluation. This is worth a few points towards your project grade.

Please read the rubric in the Project section in Brightspace which will tell you the criteria that I will use to grade this project. Feel free to contact me if you have questions.