BHSEC Queens
Statistics Spring 2016
Class 20: March 16

Name: _____

MIDTERM PROJECT: HOLLYWOOD MOVIE CONSULTING

This project is essentially "Lab 4," but it is individual and it is more open-ended. If you get a higher grade on this project than your Lab 1, 2, 3 average, I will make it worth more than one lab to reflect your improvement over time.

WELCOME TO HOLLYWOOD!

You work as a market researcher for an executive at Paramount Pictures. Your boss, Brad Grey, has just gotten a treasure trove of data. Since he is a busy man and you spent your job interview talking about how great your statistics class was at BHSEC, he has assigned you to do the analysis. He tells you that some of the unpaid interns have compiled the numbers from every movie from every studio from 2009 to 2011. He would like you to review the data to help him make good choices about which movies will make big bucks (BIG bucks) for the company. The movie currently in the works is called "'Love... Never Heard of It" about the greatest women's ping-pong champion of all time, Fu Laoshi, who used unorthodox and controversial (and some say illegal!) methods to retain her ping-pong crown. It is a biographical drama with a score of 59 on Rotten Tomatoes (a review website) and preview audiences have given it a 76 for its fantastic portrayal of "one of the most beguiling and transfixing characters of our time." It is scheduled to be released in 3145 theaters and its budget was 64 million dollars. He would also like you to suggest what kind of movie he should make next, including how much to spend on it and how many theaters it should be released in. Lastly, he would like your recommendation about what the best predictors are for how well a movie will do.

You should do a thorough job on this project from your boss because he gets cranky when things aren't done well. Deal honestly with the data. Don't inflate, massage, or distort the impression your boss should get and be honest about the strength of your models. Even though he's your boss, when he's wrong based on your data, it's your job on the line. This means you should try many combinations for what might make good predictors for how well a movie might do. Also, you may want to consider subsets of the data. Perhaps you might restrict your analysis to drama or comedy, 2009 movies, or even those with a Rotten Tomatoes score above 75? Be imaginative! Be careful with your analysis and make sure to report how good your model is and cite evidence for your conclusions. However, if you can't find a good linear relationship, state so and list all the relationship you tried.

Brad Grey needs this report for Thursday (March 24), which is when he will be talking to the big investors at Paramount. He would like at least three things in this report:

1. A commentary on the data gathering methods and how it might affect the reliability of your subsequent analysis in parts 2 and 3 below.
2. An appraisal of how "Love... Never Heard of It" will do in theaters. How confident are you in your prediction?
3. A recommendation for what kind of movie to make next, what its budget should be, and how many theaters it should be released in.

He will need your report typed and all your images (histograms, boxplots, comparative boxplots, scatterplots, etc.) and supporting evidence clearly labeled, and the talking points for him clearly laid out in your conclusions (he's "not a numbers guy").

IMPORT THE DATA

- Go to the course website at `classroom.google.com` and find today's post. The PDF you see is a digital copy of this handout. No source code necessary, but download the data and the R Markdown file.
- In RStudio, import the data and change the name to `hollywood`.
- Make sure to attach the data: `attach(hollywood)`.

REMEMBER: SOURCE, IMPORT, ATTACH.

VARIABLES:

The information in this data set is from all movies released by Hollywood based production companies between 2009 and 2011. The data was hand compiled by using boxofficemojo.com, imdb.com, TheNumbers.com, and Wikipedia.com by David McCandless. Here are the variables:

`film`	Title of the movie
`studio`	Studio that released the movie
`rottenscore`	Rotten Tomatoes rating (critics' ratings)
`audscore`	Preview audience rating (via Rotten Tomatoes website)
`story`	General theme (one of 21 themes)
`genre`	Action, Adventure, Animation, Comedy, etc
`openingtheaters`	Number of screens for opening weekend
`avgopening`	Average box office income per theater (opening weekend)
`DOMgross`	Gross income for domestic viewers (in millions)
`FORgross`	Gross income for foreign viewers (in millions)
`WWgross`	Gross income for all viewers worldwide (in millions)
`budget`	Production budget (in millions)
`percentprofit`	WWgross divided by budget
`grossopening`	Opening weekend gross (in millions)
`year`	year the movie was released

RSTUDIO REFERENCE:

You are encouraged to look through past labs, but to start, here are some commands you may find useful:

- Subset a whole data set: `newdata <- subset(data, conditions)`. The conditions can be `>`, `<`, or `==`. Categorical variable values must be quotes (e.g. `"Adventure"`).
- Barplot: `barplot(table(x))`
- Mosaic plot: `plot(table(x,y))`
- Scatterplot: `plot(x, y)`
- Scatterplot of all variables: `pairs(dataset)`
- Correlation: `cor(x, y)`
- Coefficient of determination: square the correlation.
- Titles: add the `main` argument, like so: `plot(x, y,main="What a nice title!")`

- Labels: add the `xlab` and `ylab` arguments:
`plot(x, y, xlab="Look at this x-axis!", ylab="Look at that y-axis!")`
- Linear model: `lm(y~x)`. *Notice how the variables are in the opposite order.* The output gives the intercept first and slope second.
- Plotting a line: `abline(a, b, col="red")`. This command draws a line with intercept `a` and slope `b` *on whatever plot has most recently been made.* For example, `abline(3,2,col="blue")` will draw the line y=3+2x on the current plot. *If no plot is open, an error may occur.*. HINT: it may also help to Google the `curve` function.
- If you want to find a model that is not linear (e.g. exponential), talk to me and I'll help you through it.

GRADING AND GUIDANCE:

Due to the open-ended nature of this project, it will be graded a little differently and since you can take it in so many different directions, there is no exact point-by-point rubric, but here are some things I will look for:

1. Data Gathering
   (a) Discuss any biases in the data set
   (b) Discuss ways to improve the data gathering methods
   (c) Discuss how the data gathering methods might affect the analysis of and conclusions based on the data set
2. Describing Data
   (a) Display data and choose summary statistics appropriately
   (b) Describe data visuals appropriately
   (c) Interpret data visuals and summary statistics appropriately
3. Linear Regression
   (a) Comment on the strength of linear associations appropriately
   (b) Interpret and use linear regression models appropriately
   (c) Deal with outliers and influential points appropriately (where applicable)
   (d) Work with subsets of data appropriately (where applicable)
   (e) Straighten relationships appropriately (where applicable)
4. Data Analysis Habits
   (a) Create and follow your own questions
   (b) Explores multiple variables (or subsets of variables) for relationships
   (c) Consider evidence for and against a claim (avoid "cherry picking")
5. Communication Habits
   (a) Expression of ideas is engaging and makes report enjoyable to read
   (b) The integration of visuals and discussion is well balanced
   (c) The expression of ideas is succinct yet thorough
   (d) The organization of the written work helps to communicate ideas