# STAT 6210 - Project #1 - Due March 22, Tuesday

**Due date**: March 22, 2016, 6:10 PM

This project has two components. In the first section, you will download a data set and build a predictor on R. In the second section, you're asked to write a project proposal.

## 1 Data Analysis (50 points)

*All of this takes place in a galaxy far far away (i.e. I made up the data).*

Your client, Graystone Broadcasting Inc., is a major commercial broadcasting network on the planet of Caprica. Their largest revenue comes from Friday Night Live, a live comedy show that features different celebrities every week. Graystone Broadcasting is interested in understanding how the different elements that make up every episode impact the viewer count for that week. They have contacted you to help them make accurate prediction of the viewer count based on the major factors of the show.

Also, they believe that **some of the shows might be outliers** but they have good hope that you can **identify these, remove them from the dataset** and help them find a simple formula to estimate the viewer count of other episodes.

Your client does not expect you to investigate the causal relationships for the variables. They also believe that variables that are commonly assumed to be good indicators might not be really good predictors for viewer count.

The dataset is available on Blackboard. The columns in the dataset are not labeled (headers indicate the column number) but the description of each column is given below.

1. Advertisement Spending (in million cubits)
2. Average Score of Director on cmdb.com
3. Director Experience (# of episodes made)
4. Average Score of Writers' Previous Shows on cmdb.com
5. Average Writer Experience (# of episodes made)
6. Guest Artist's Caprican Billboard Top 20 Appearances
7. Number of Guest Artist's Fans (as counted from theCapricanFacebook.com)

8. Guest Artist's Experience with Similar Shows (on a scale of 100%)
9. Average Score of Guest Actor on cmdb.com
10. Number of Guest Actor's Fans (as counted from theCapricanFacebook.com)
11. Guest Actor's Experience with Similar Shows (on a scale of 100%)
12. Average Score of Guest Actress on cmdb.com
13. Number of Guest Actress's Fans (as counted from theCapricanFacebook.com)
14. Guest Actress's Experience with Similar Shows (on a scale of 100%)
15. Categorical: Surprise Appearance by a Non-Actor Celebrity (1,0)
16. Categorical: Holiday Season Show (1,0)

Your job is to create a formula to help Graystone Broadcasting to predict the viewer count for episodes. You are also asked to present diagnostics: calculate p-values of the variables in your final model, and using tests for normality (along with other tests we have seen so far), explain how meaningful these p-values are.

While this dataset has been collected without your advice, Graystone Broadcasting also asked you what information might be needed to improve their prediction accuracy if needed.

### Held out data

Graystone Broadcasting takes this project very seriously, and they have contacted other statistical consulting companies. To see which consultant does best, they will evaluate the models on 100 data points that they held out. They will then pick the model with the lowest test data MSE. After formulating your final model, **use this model and predict the responses for the held out data. Upload your predictions to Blackboard for these 100 samples (Xtest) as a csv file.**

*Your model's performance on the held out data will determine 30% of your grade from this section.*

## 2 Project Proposal (50 points)

Write a 1-2 page project proposal. The proposal should answer the following questions:

- What is your statistical problem, why is it important?
- What statistical methods can be applied to solve this problem? (Google will be your best friend here)
- What R packages do you need to implement those methods?
- How will you collect your data? Is it available online?

Your answer to the first question should at least be 2 paragraphs long, and probably even longer. Good statistical problems can be hardly described with 300 words.

## Resources for Datasets

- Kaggle.com
- UCI Machine Learning Data Repo (http://archive.ics.uci.edu/ml/datasets.html)
- Health and Medical Care Data Archive (http://www.icpsr.umich.edu/icpsrweb/HMCA/archive.jsp)
- US Census (http://www.census.gov/)
- Yahoo! Finance, Reuters, or Bloomberg (for Financial Data)
- Twitter
- Yelp
- R packages (http://www.public.iastate.edu/~hofmann/data_in_r_sortable.html)
- http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html
- http://www.kdnuggets.com/datasets/index.html