Assignment 6 (term): Data Analytics (Spring 2016) 30% (25% written+5%oral)
Due: Written on TUESDAY May 3, 2016 (by 2pm ET). Presentation sent after you present.

Submission method: written and presentation (after you present it) by email to pfox@cs.rpi.edu **and** Rahul Divekar divekr@rpi.edu

Please use the following file naming for electronic submission:
DataAnalytics2016_A6_YOURFIRSTNAME_YOURLASTNAME.xxx, etc.

Late submission policy: **this assignment is due at the end of term. If you are more than a week late it is likely that you will not have your grade for this assignment included in your final grade before they need to be submitted.**

Note: Your presentation for this assignment should be the result of your own individual work. Take care to avoid plagiarism ("copying"), and include references to all web resources, texts, and class presentations. You may discuss the project with other students, but do not take written notes during these discussions, and do not share your written assignment or presentation before the class they are presented in.

General assignment: Your term projects should fall within the scope of a data analytics problem of the type you have worked with in class/ labs, or know of yourself – the bigger the data the better. This means that the work must go beyond just making lots of figures. You should develop the project to indicate you are thinking of and exploring the relationships and distributions within your data to lead to optimized predictive models. Start with a hypothesis, claim, or questions. Think of one or more ways to construct model(s)[1], find or collect the necessary data, and do both preliminary analysis, detailed modeling, validation, summary (interpretation) and (if any) resulting decisions.

Note: You do not have to come up with a positive result, i.e. disproving the hypothesis is just as good. Please use the section numbering below for your written submission for this assignment.

Guidance: Topics, scope and general nature – please use the opportunity in Assignment 5 (project proposals) and seek feedback from the instructor and your classmates.

1. Introduction (2%)
   Describe your motivation, initial hypothesis/ idea that you wanted to investigate, and if applicable any prior work, interest in the topic (like an intro for a paper, with references). Min. 1/2 page.

2. Data Description (3%)

[1] **NOTE: graduate students must develop at least two different types of models, not just change the number of variables for a given model.**

Describe how you determined which datasets you used in this project, the criteria, source, data and information-types in detail, associated documentation and any other supporting materials. Min. 1/2 page text (+graphics if applicable).

3. Analysis (5%)
   Explore the statistical aspects of your datasets. Perform any transformations, interpolations, smoothing, cleaning, etc. required on the data, to begin to explore your hypothesis/ questions. Analyze the distributions; provide summaries of the relevant statistics and plots of any fits you made. Discuss and specify or estimate possible sources of error, uncertainty or bias in the data you used (or did not use). Min. 2 pages text + graphics.

4. Model Development and Application of model(s) (12%)
   Identify what types of models you used to describe the data (regression, classification, clustering, etc.), patterns/ trends you found, visual approaches that helped you choose models, and or variables (type/ number) in the model, other parameter choices or settings for the models (e.g. distance metrics, kernels, etc.). Apply the models to assess model performance (i.e. predict). Discuss the confidence in your results including any statistic measures. Discuss how you validated your models and performed any optimization (give details). Min. 6 pages text + graphics.

5. Conclusions and Discussion (3%)
   Describe your conclusions; interpret the results, predictions you made, the models and their characteristics, and a give summary of what changed as you went through the project (data, analysis, model choices, etc.), what you would do next, or do differently in a subsequent exploration. Min. 1 page text + graphics (optional).

References – websites, papers, packages, data refs, etc. should be included at the end. Include your R scripts! (e.g. in a zip file).

6. Oral presentation (5%). Suggest these slides (limit your presentation to 5 mins):
   a. Title (with your name)
   b. Problem area – what you wanted to explore/ solve/ predict and why, and what you wanted to predict?
   c. The data – where it came from, why it was applicable and the preliminary assessments you made.
   d. How you conducted your analysis: distribution, pattern/ relationship and model construction. What techniques did you use/ not use and why?
   e. How did you apply the model? How did you optimize, account for uncertainties?
   f. What did you predict and what decisions (prescriptions) were possible. What was the outcome?

Graphical Representations
   Provide graphical representations related to each of questions 2, 3, and 4, at least. Ensure all figures are numbered, legible, fully explained and annotated.

The final document should be a minimum of 8 pages of writing (but can be more). All graphics should be within your written assignment unless they are very large. Large graphics files should be sent as a separate attachment (e.g. in a zip file).