Assignment 4: Data Analytics (Spring 2016) (15% written)
Due: FRIDAY March 11, 2016 (written by 5pm ET)

Submission method: written document 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_A4_YOURFIRSTNAME_YOURLASTNAME.xxx

Late submission policy: first time with valid reason – no penalty, otherwise 20% of score deducted each late day

Note: Your report 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 problems with other students, but do not take written notes during these discussions, and do not share your written solutions.

General assignment: Pattern, trend, relations: model development and evaluation of housing (Brooklyn, Manhattan, Queens) datasets (http://aquarius.tw.rpi.edu/html/DA/*sales*). The weighting score for each question is included below. Please use the question numbering below for your written responses for this assignment.

Please include code (fragments and/or scripts) and the plots you generate for the questions below.

1. For any **one** of the Brooklyn, Manhattan, Queens sales datasets, perform the following:
    a. Describe the type of patterns or trends you might look for and how you plan to model them. Describe any exploratory data analysis you performed. Include plots and other descriptions. Min. 2-3 sentences (2%)
    b. Pick one or more models (these need not be restricted to the models you've learned so far [multivariate regression, KNN, K-Means]) to explore the chosen data. Interpret the model fits and indicate significance. Describe any cleaning you had to do and why. Min. 2-3 sentences (3%)
2. For your chosen dataset:
    a. Apply the model(s) to predict quantities of interest (that you choose). Describe (contingency table) or plot the predictions. Min. 2-3 sentences (ugrad 5%, grad 3%)
    b. Examine the fit(s). Perform a significance test that is suitable for the variables you are investigating and describe the results. Min. 2-3 sentences (ugrad 4%, grad 3%)
    c. Discuss any observations you had about the datasets/ variables, other data in the dataset and/or your confidence in the result. Min 1-2 sentences (1%)
3. Graduate 6xxx-level question (3%). Draw conclusions from this study – about the model type and suitability/ deficiencies. Describe what worked and why/ why not. Min. 4-5 sentences