A02: Classification Models
Due date: 11:59PM Saturday July 18th

Problem Statement

In this assignment you will apply the idea of classification to your data sets. How can classification be useful? You should start by spending some time thinking about the "response" (y) variable you are interested in. When you dealt with regression, some of your predictors (x's) may have been categorical. Now you are dealing with a modelling technique that requires that the response (y) variable is categorical. Could using data to come up with some sort of classification be useful or interesting? As an example, consider classifying the traffic level of ambulances at hospitals in Fredericton. Although the requests for the number of ambulances in a county is a discrete variable, you can experiment with models that would classify the number as low concern (1-2), moderate concern (3-5), and concerning (6+). Classifying the "ambulance situation" at a particular hospital helped administrators decide if they should divert the EHS vehicles to another hospital or not, and if so, to which hospital. If you need to, you can take numerical data and make it categorical. Indeed, there may be definitions for useful categories out there that people commonly use that relate already to your dataset of interest.

You can reuse dataset used in A01 or choose a new dataset. **A reminder also that you are not expected to have all the answers, just very good questions and observations at this point.**

Expectations

A convincing argument will likely include:

- Five or more charts
- Three or more classification models
    - Of course, the number of charts and models depends on your argument, how in-depth you go into each image and model. If you have only one very sophisticated model and go very in-depth, perhaps one model is sufficient.

Guidelines

1. State your "thesis" clearly and at the introduction and conclusion in your report. What are you trying to prove? (For example, perhaps you suggest that Masks do not help reduce the spread of the virus. If this is your thesis statement, state it plainly up front, then prove it in the body of your writing, and repeat this at the conclusion and summarize briefly your main points.)
2. Take time to explain the models and the conclusions that can be drawn from them.
    a. Do all the technical work you need to do but put extra analysis in the appendix at the end
    b. Use the body of the report to show pictures of interesting data, introduce model summaries, explain which variables are significant, and attempt to interpret the data.
    c. Interpretation: by this I mean explain why your results are interesting or important and how they could be potentially useful to key decision makers, health care administrators.
3. Include your R studio Script as an appendix to your submission
4. Technical hints:
    a. Determine error rates
    b. Try and create models that use more than one predictor and see if the error rates decrease as the model becomes more complex

Submission

1. You will create your solution in a tracked GIT repository.
2. The commit to be evaluated will include the message "Final Submission"
3. You will submit link to the solution git repository in D2L as submission to the assignment.

Grading Rubric

A02 - Classification Grading Rubric				
	Very poor Value: 0%	**Poor** Value: 0.5	**Average** Value: 1.5	**Good** Value: 2.5
Plots and Interpretation 25%	Mostly inappropriate or simplistic plots and interpretations	Plots and interpretations "shot from the hip" and show promise but too little detail and deep understanding not shown	Student didn't always use best plots for the context. Interpretations solid, but could have gone further	Student has used the appropriate plots for the situation, has made them easily readable, and formed solid conclusions from them.
Models: the technical work 25%	Only 1 or 2 categories on the far right completed at all or several completed insufficiently well.	Only 2 -4 categories to the right completed only reasonably well	At least 3 of the categories described to the right done quite well	1. Classification categories make sense. 2. Several different models discussed. 3. test and training data used, 4. null classifier used as benchmark for good error rates
Models: their Interpretation 25%	Claims made about models have great errors in logic or understanding	Interpretations are incorrect in a significant way	Interpretations are imprecise but generally correct	Student has accurately interpreted their models and show that they understand well key relationships of variables. They also do not over- or under- state claims or conclusions
Readability and Argumentation 25%	The student's thesis is hard to understand. The visuals and models do not hold together very well.	The marker can make a guess at the student's thesis with some work, but the blog doesn't hold together very well. Graphs and models are scattered in their application to the main point	The thesis statement is fairly clear but the blog "meanders" a little and there are a few points here and there that distract from the main point	It is obvious what the student is trying to prove and all models and diagrams help support their case