MSDS 6372 Project 1 Description

There are two main objectives for Project 1. Since each group will be using their own data set, there will be a little flexibility in what needs to be delivered. Below is a summary of what is absolutely necessary as part of your report.

Objective 1: Display the ability to build regression models using the skills and discussions from Unit 1 and 2.

- Perform your regression analysis and report the predictive ability of your model using a test set or some other means through CV. Be sure to provide metrics if you compare multiple models.
- Provide interpretation of the regression model including hypothesis testing, interpretation of regression coefficients, and confidence intervals. Practical vs Statistical significance.

Logistical Considerations.

- Depending on the groups data set (number of predictors and complexity in your final model), it may make more sense to produce two models. One model that may not be the absolute best for prediction, but would be good for interpretation and hypothesis testing. Another model could achieve the opposite, high predictability but difficult in providing interpretation. For smaller data sets it may not be required and one model can provide both. Do what makes sense given your own data set and your own model building process.

Objective 2: Perform a secondary analysis using the tools from Unit 3 or Unit 4. You only need to do one, pick the one that makes the most sense.

- Two Way ANOVA - I required you guys to make sure there are at least 2 categorical variables and this is that reason. Using the two categorical predictors in your data set, pretend that these are the only variables available to you and preform a Two Way ANOVA analysis. (Of course this would not the correct thing to do given we have access to other variables, but illustrate your ability to perform the analysis in this exercise).
- Time Series - Many times in regression analysis the observational rows in the data set are collected over time and not really mentioned or recorded. Create an additional variable in your data set denoted “time” and code it from 1 to the total sample size (It can be the entire data set, or the training or the test set). Use your knowledge and understanding of time series to assess if the assumption of independent errors of your final regression model is actually a valid assumption.

Logistical Considerations.

- If you are dealing with a large data set, time series may get bogged down a little bit computationally, but to answer the question I don’t think you will have an issue.
- If you are dealing with a large data set, make sure to consider practical vs statistical significant differences in your findings of a Two Way ANOVA analysis.

Additional details

NOTE 1: ALL ANALYSIS MUST BE DONE IN SAS OR R and all code must be placed in the appendix of your report.

Required Information and SAMPLE FORMAT
Required deliverables in the complete report. The format of your paper (headers, sections, etc) is flexible although should contain the following information.

PAGE LIMIT: I do not necesarrily require a page limit, but you should definitely be shooting for know more than 7 pages written. It of course can blow up quite larger than that due to graphics and tables, but good projects are clear, concise, to the point. You do not need to show output for every model you considered. (You may put supporting plots/charts/tables etc. in the appendix if you want, just make sure you label and reference them appropriately.)

Introduction **Required**

Data Description **Required**

Exploratory Analysis **Required**

Addressing Objective 1:
- Restatement of Problem and the overall approach to solve it **Required**

- Model Selection **Required**
  - Type of Selection
    - **Optional**: LASSO, RIDGE, ELASTIC NET etc. and Model Averaging
Stepwise, Forward, Backward, Mallows Cp,
Manual / Intuition
A mix of all of the above.

  - Checking Assumptions **Required**
    - Residual Plots
Influential point analysis (Cook's D and Leverage)

  - **Optional:** Comparing Competing Models
    - **Optional**: (AIC, BIC, adj R2
Interval CVPress
External Cross Validation**)**

- Parameter Interpretation
  - Interpretation **Required**
  - Confidence Intervals **Required**

- Final conclusions from the analyses of Objective 1 **Required**

  In addition to overall conclusions, feel free to include additional insights or conerns gleaned from the analysis. What needs to be done next or how could we do it better next time?

Addressing Objective 2

State what route you are going to take 2way ANOVA or Time series and summarize the goal. **Required**

Main Analysis Content **Required**

This will depend on the route you take. I'm leaving it open here to see what you do.

Conclusion/Discussion **Required**

The conclusion should reprise the questions and conclusions of objective 2.

Appendix **Required**

Well commented SAS/R Code **Required**

Graphics and summary tables (Can be placed in the appendix or in the written report itself.