**MSDS 6372 Project 2 Description**

For Project 2 we will continue working with the housing data set from Project 1. In addition, since we will be adding to your current work, we will keep the same groups as well. There are three main objectives to this project:

1. (10%) Address the comments that were left in your last paper. There are various types of comments so here is the idea. You only need to address those comments that are constructively critical. There will be separate section of the new paper entitled “Addressing the Comments” in which you simply need to list the comment and then list what your remediation of that comment was. As an example, if it was a simple spelling error, then your comment would simply be, “Fixed the error” and then of course go and fix the error in the paper. As another example, it might be comment like, “The assumptions need to be addressed more completely in this section.” In that case, the comment may be: “We went back and added a residual plot and addressed each assumption underneath the plot.” Finally, it may be comment like, “You should consider an interaction term, it may help predictive ability.” The comment to this could be, “ We looked at interactions and didn’t find any significance and thus didn’t add it to the paper.” … or … “We will take that into consideration next time!” Again, there are various comments to address … some require action and some do not. In general, suggestions on how you “could” have done something more do not need to be acted upon whereas those that say or suggest the student “should” do something (like address the assumptions) need to be fixed / added and noted. As another example, some teams left out a final model equation or explicit explanation of estimates. This is a “should” and thus should be added to the paper. The basic idea is to have thought about important comments and have improved your analysis. Finally, there should be a lot of “Good!”/ “Well done.” comments on your paper because I like them. You do not need to address or record these comments at all.
2. Principal Components (50%)

In this section, you will attempt to better your Kaggle score through the use of principal components. In this process you must address the assumptions of principal components, produce the interpretations of the eigenvalues, screeplots, and other criterion that aid in selecting components. In addition, you must also attempt to interpret the components selected with respect to the sale price of the houses. This may be tough, but do your best. Furthermore, the investigation of the use of PCA in this project should be used in conjunction with the regression techniques you have already been using (adding categorical variables, investigating OLS, LASSO and coefficients, cross validation, variable selection, etc.) The team with the best Kaggle score this time around will again earn an extra 3 points. You only need to present ONE model but it should be the best one that you can achieve. Simply provide the same table you filled out before with your new scores as well as the code you used to generate it. As before, only techniques we have learned in this class can be used.

1. LDA (30%)

For this part, you will simply use the training set to build a classification model (using LDA) to predict the type of Foundation of houses in the Test Set. Similar to the Kaggle competition, each team will provide me with the code that will both train your final classification model and provide predictions given the test set. Only ONE model need to be developed but it should be the best one that can be achieved. The winner of this competition will earn an extra 3 points.

This section should be written up in a clear and concise manner and should include all the plots, charts, tables and explanation necessary to defend the assumptions of LDA and to validate the fit and performance of your model (confusion matrices, etc.)

1. (10%) Presentation: Spelling, Grammar, labeled plots and charts with references in the writing, consistent titles and subheadings, consistent organization of sections.

The paper you submit will be the same paper you handed in for Project 1 with the following changes / additions:

1. All “should” comments should be fixed. (All plots and tables should be labeled and referenced in the writing. This may take some time, but very worth the effort.)
2. The addition of the “Addressing the Comments” section.
3. A small conclusion summarizing the model developments after the addendum, if any.
4. The addition of the “Principal Component Analysis” section.
5. The addition of the “LDA” section.
6. Additions and changes to the Appendix. (Changed code for project 1 just needs to be updated; new code for project 2 needs to be put in a new section in the appendix.)

NOTE 1: ALL ANALYSIS MUST BE DONE IN SAS and all code must be placed in the appendix. Part of the grading process will be to run the code and verify the Kaggle score for each group.

Note 2: An extra 3 points will be awarded to the team with the model with the highest Kaggle Score. In the unlikely event of a tie will split these points.

**Required Information and SAMPLE FORMAT new material in RED**

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

Introduction **Required**

Data Description **Required**

Exploratory Analysis **Required**

Analysis of Question 1:

Restatement of Problem **Required**

Model Selection

Type of Selection

LASSO, Model Averaging

Stepwise, Forward, Backward, Mallows Cp,

Manual / Intuition

A mix of all of the above.

**At least two of the above required.**

Checking Assumptions

Residual Plots

Influential point analysis (Cook’s D and Leverage)

Comparing Competing Models

AIC, BIC, adj R2 **Required**

Interval CVPress **Required**

External Cross Validation **Required**

Parameter Interpretation

Interpretation (Verbal) **Required**

Confidence Intervals **Required**

Analysis Question 2

Restatement of Problem **Required**

Model Selection

Type of Selection

LASSO, Model Averaging

Stepwise, Forward, Backward, Mallows Cp,

Manual / Intuition

A mix of all of the above.

**At least two of the above required.**

Checking Assumptions

Residual Plots

Influential point analysis (Cook’s D and Leverage)

Comparing Competing Models

AIC, BIC, adj R2 **Required**

Interval CVPress **Required**

External Cross Validation **Required**

Kaggle Score **Required**

***Addressing the Assumptions Required***

***Principal Components Analysis Required***

***LDA Required***

Conclusion/Discussion **Required (new paragraphs of conclusion on the LDA and PCA analysis Required)**

The conclusion should reprise the questions and conclusions of the introduction,

perhaps augmented by some additional observations or details gleaned from the analysis

section. New questions, future work, etc., can also be raised here.

Appendix **Required**

Well commented SAS Code **Required**