Rubric for Statistics 6372 Project
Name of Students: Heinen, Sanadhya, Coyne

Total Score: ___91___ out of 100 points

Following Guidelines (10 points)	
Was the final project turned in on time?	
	Total: ___10___ out of 10 points

Project set up and Data Collection and Cleaning	
(5 points) Were explanatory and response variables identified? Was the type of each variable mentioned? Was a reference given for the data set? There were two questions asked, are they being tackled separately or together?	Yes. The data dictionary was given as a reference to the variables, and a general overview of what the predictors are. However, I think it is still good practice to provide the data dictionary yourself, and include how you are treating them in terms of type (some people may treat Quality as continuous while others categorical levels). Your guys overall approach to buiding a model was well done and you articulated that was your goal, build a model for prediction.
(5 points) Was the data set described appropriately? If only part of the data were used, is the procedure for selecting that part described (Both variables or observations)? Was missing data discussed and explained how it was handled.	I wished there was more discussion about the -1s. Some people had some issues with this, but I think through your screening approach, some of the variables in question may have been discarded up front. If you included these predictors during your feature selection, you actually were only fitting your models to about half of the data points. To be more specific, was there any question whether or not if Fence was -1 if then that simply meant that the house didn't have a fence? So the missing value may make sense to be treated like a categorical level and recoded for that. Depending on whether you looked at the original data dictionary online, it may have been more obvious to consider. "For MsZoning, 79% of the data had a value of RL. MSSubClass indicated that more than 1/3 of the homes are 1 story, 1946 and Newer, All Styles. LandContour had 90% of homes listed as LVL. There were several categorical variables that could be removed using this logic. We felt these types of variables would not help in ultimately predicting the sales price of a home." Your general argument here makes sense, if a continuous predictor doesn't vary much, then it can't be very helpful in prediction. Categorical variables though is a little

	bit different story. Even if its 90% favored in one group if the mean salary drastically changes in that other 10% then it could really be helpful.
	Total: ___8___ out of 10 points

Data Analysis and Model Building	
(10 points) Were appropriate descriptive statistics shown for various aspects of the data (response variables, explanatory variables, assessment of correlations, multicolliearity assessments? Were obvious issues discussed, mentioned, and/or attempted to address.	I would have liked to seen at a minimum a basic plot and summary statistics of the response variable. Just let everyone know generally the range of the values it takes, etc. Its just the natural first step to make sure everything makes sense. Table 4 provided overall fit statistics for different model fitting which I thought was very informative.
(10 points) Was there a graphical display of the response variable and the explanatory variable or a subset of them? Was the display appropriate? Were any special or unusual features noted and explained?	Yes. These were all handled pretty well.
(10 points) Discussion of feature selection approaches (both algorithm and common sense logic). Summary of the procedures and appropriate displays to illustrate.	The overall approach to your guys providing feature selection is an interesting one. Take a concensus look at what multiple algorithms do and build a model from there. I like it. The one issue I have is some of the wording that confused me on what you guys were doing. After you guys summarized the various results of the feature selection methods, you seem to imply that a final model still needs to be made… "After our Exploratory Data Analysis and Variable Screening, the below variables were selected to remain as we proceed with selection methods in SAS to choose our final linear model." But you guys don't do anymore model building so its hard to understand what the final model is unless you dig into Figure 5. So I would try to emphasize and make it more clear exactly what the final model is and what the final predictors are.

	Total: __27__ out of 30 points
Inference (For the concise model)	
(10 points) Was the appropriate inferential procedure used? Were the null and alternative hypotheses given? Were the population and sample correctly identified? Were the results interpreted in context?	You guys do not discuss any global f-test or individual t-tests which is okay, you care about prediction, but because you guys have low VIFs and a pretty simple model, I felt that there was a lost opportunity to dive in a little more and provide a little bit of insight in how your model is actually estimating various trends and behaviors in the data. For example, if the overall quality of my house were to increase, by what amount would I expect to see in the sales price (or log sale price)?
(10 points) Were the assumptions checked with the appropriate statistics or graphics?	Yes. My only nitpick here is that the your comments and interpretations of the residual should just be structured as a paragraph in the manuscript. Since you do not reference it anywhere in the words, its quite possible that someone may simply overlook that part of the figure.
(10 points) Was the appropriate output from software shown in order to support conclusions of the inferential procedure?	Yes
	Total: __28__ out of 30 points

Conclusions (10 points)	
Model performance assessment through KAGGLE socre. Were weaknesses in the study pointed out? Were the conclusions given in the context of the appropriate population from which the sample was drawn?	Yes. What about predictors that were not available? Crime statistics, business growth, etc? you guys addressed these in the limitations section.
	Total: __10__ out of 10 points

Clarity (10 points)	

(5 points) Is the report clear and concise? Did the report have sections? Did paragraph transitions flow smoothly from one paragraph to the next? Were words overused? Were their excessive spelling or typographical errors?	Overall the report is pretty straightforward and easy to read. There are a few tranisitions (noted in previous sections) that created some confusion for me in terms of where you are going with the model building steps.
(5 points) Did graphics have appropriate titles and labeling? Were figure captions appropriate?	Yes, but make sure tables and figures are actually reference in the paper. The report needs to flow in such a way that each table and figure is supporting what you are writing. If it doesn't support what you are saying directly, then it probably shouldn't be a table or a figure. Don't let tables and figures try to do the talking for you.
	Total: __8__ out of 10 points

Additional Comments: