Housing Prices Grade: 92 (A nice report generally, but I suggest you include more key SAS final outputs such as parameter estimates/CI.)+3(best Kaggle)=95

# Introduction

Traditionally, estimating the sale price of a house is based on a rough comparison to similar houses on the market. But, what if there is a better way to predict the sale price of a house? Many measurable factors affect housing prices. While a large list of variables is not useful for making a prediction, the use of these variables in a statistical model provides an easy and accurate way to answer the latter question.

This project will compare different linear regression models that can help buyers, sellers, and real estate agents predict the sale prices of different homes in Ames, Iowa. The data used to build these models includes a non-random sample of homes from Ames, Iowa. Thus, the scope of the predictions of housing prices is restricted to houses within the neighborhoods included in the dataset. While the scope of inference is limited, these models may provide a framework for predictive modeling in other housing markets.

This project uses multiple linear regression and model selection techniques to build 6 models. The first 3 competing models are optimized for easy consumption and interpretation by parties working in the home marketplace. The next 3 competing models are optimized for greatest predictive power, possibly containing much more variables and complex interactions and effects.

# Data Description

The Ames, Iowa housing data set was published by Dean De Cock in the *Journal of Statistics Education,* Volume 19, Number (3) (2011). This data was accessed via Kaggle for free at <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>. The data contains 79 explanatory variables, describing nearly every aspect of residential homes in Ames, Iowa. The data is broken up into 2 separate data sets:

1. Train data set has 1460 observations and 81 variables
2. Test data set has 1459 observations 80 variables

A complete list of the defined variable abbreviations can be found at <https://ww2.amstat.org/publications/jse/v19n3/decock/datadocumentation.txt>.

# Exploratory Data Analysis (EDA)

The first step in building a simple or complex model is exploring the data. The primary goal of EDA is determining if the explanatory and response variables meet the assumptions of multiple linear regression. This includes looking for normality in the data, variables that need transformation, averages, high leverage and high influence points. We will be focusing mainly on the EDA completed on question number 1.

The first step we took exploring the data was to create scatterplots of the sale price against all of the individual numeric explanatory variables to look for linearity (Fig. 1). This showed non-normality in the sale price, indicating the need for transformation. We used both a square root transformation for 2 simple models and a logarithmic transformation for the final simple model. This solved the majority of the linearity issues. The large number of observations should cover for any remaining non-normality. Furthermore, the pair-wise scatter plots were useful for finding explanatory variables that may need transformation and explanatory variables that exhibited collinearity with each other.

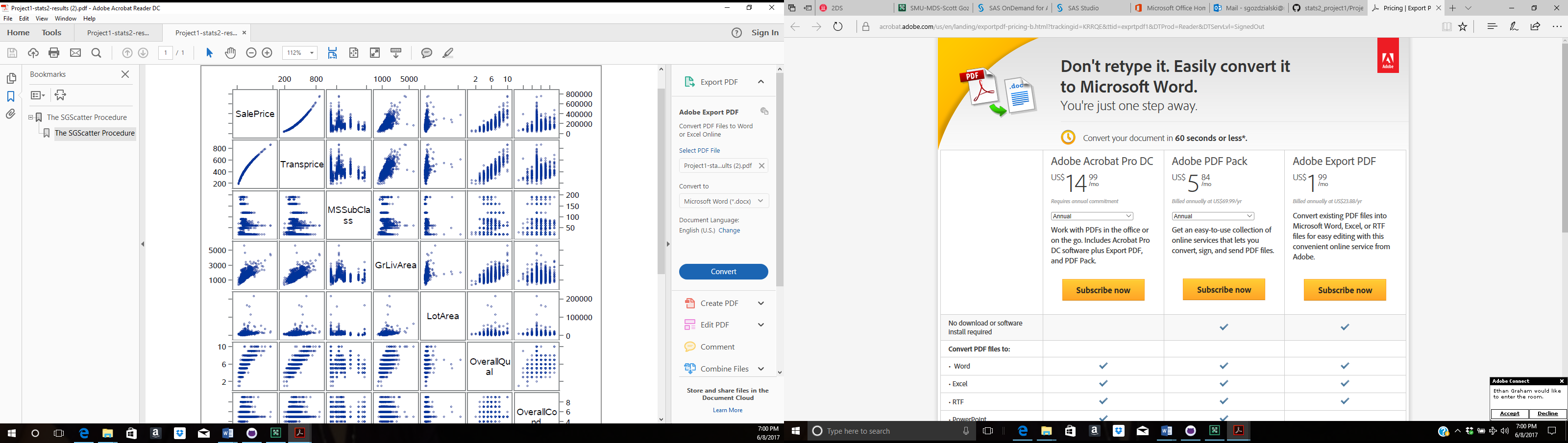


Fig 1

The PROC MEAN function in SAS was used to find the average sale price of the train data set. This average is useful to potentially replace extreme sale price values.

Finally, the PROC FREQ function was used to gain understanding of the qualitative variables. Variables with many missing values or a large majority of observations in a single category were likely to be left out of the models.

# Analysis of Question 1

The goal of question one is to create a valid model to facilitate the easy interpretation of parameters for use in helping real estate agents, contractors and prospective buyers gain insight into the important factors that influence housing prices in Ames, Iowa.

## Model Selection

### Type of Selection

1.Model one was developed by using the forward automatic selection with a limited number of variables. The square root transformation was used on sales prices to help with the non-normality of the distribution. The first step PROC SGSCATTER was used to get a matrix of the numeric variables, which was analyzed to pick the variables that demonstrated a relationship to the house’s sale price. Next, the categorical variables were chosen based on intuition while limiting them to no more than five, three were ultimately chosen. Then the variables ran through PROC GLMSELECT forward selection using 5 fold cross-validation and CV as the reason for stop. Then the numeric variables were run through PRC REG to check the VIF rating, which displayed no collinearity. Then the complete model ran the training data was sent through PROC GLM to check the F-Statistic and P-value to ensure none to the slopes were statistically zero, which it all passed. Finally, all the test data was run through PROC GLM to get the prediction results from the model.

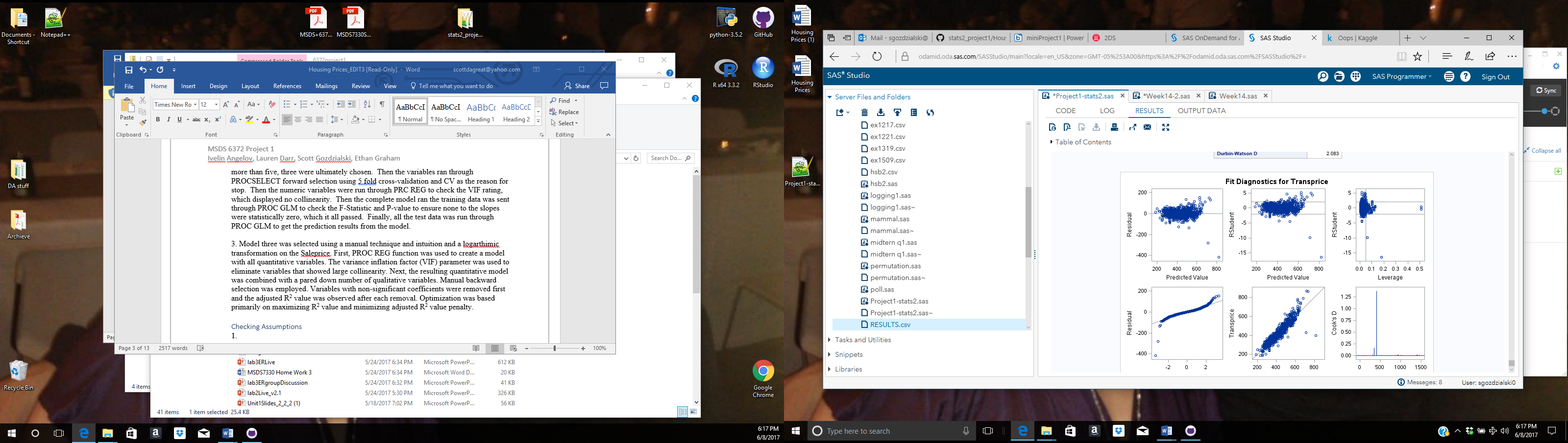
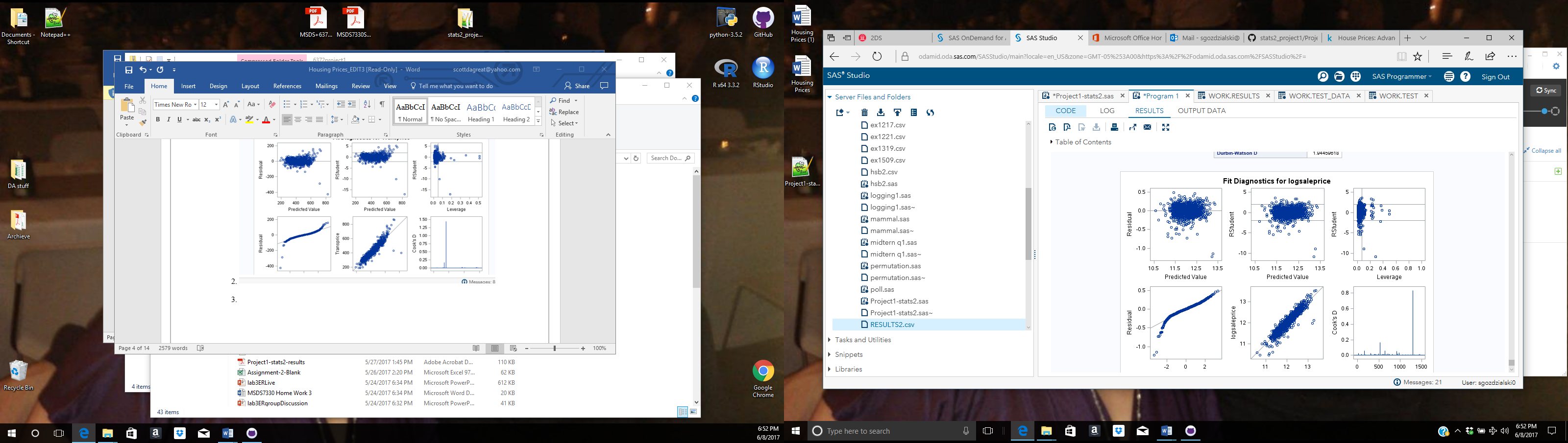
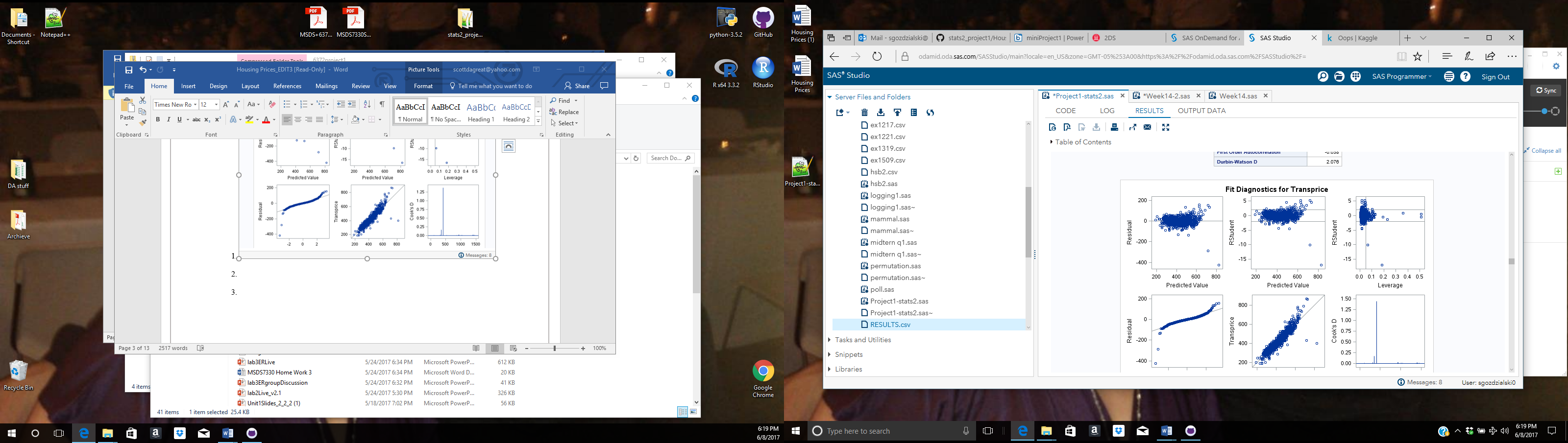
2. Model two was developed by using the LASSO automatic selection with a limited number of variables. The square root transformation was used on sales prices to help with the non-normality of the distribution. The model was more inclusive than model one, but was run through the same process as model one, list again in this paragraph. The first step proc sgscatter was used to get a matrix of the numeric variables, which was analyzed to pick the variables that looked like they were related to the houses sale price. Next the categorical variables were chosen based on intuition while limiting them to no more than five, three were ultimately chosen. Then the variables ran through PROC GLMSELECT forward selection using 5 fold cross-validation and CV as the reason for stop. Then the numeric variables were run through PRC REG to check the VIF rating, which displayed no collinearity. Then the complete model ran the training data was sent through PROC GLM to check the F-Statistic and P-value to ensure none to the slopes were statistically zero, which it all passed. Finally, all the test data was run through PROC GLM to get the prediction results from the model.

3. Model three was selected using a manual technique, intuition and a logarithmic transformation on the sale price. First, PROC REG function was used to create a model with all quantitative variables. The variance inflation factor (VIF) parameter was used to eliminate variables that showed large collinearity. Next, the resulting quantitative model was combined with a pared down number of qualitative variables. Manual backward selection was employed. Variables with non-significant coefficients were removed first and the adjusted R2 value was observed after each removal. Optimization was based primarily on maximizing R2 value and minimizing adjusted R2 value penalty.

### Checking Assumptions

During the model building process assumptions were checked throughout the model selection. During the exploratory data analysis, the normality of sale price was checked using PROC UNIVARIATE. Sale price had a left skew and required transformation. Normality was also checked during the model building process. The histogram indicated relative normality for the final model. Also, the plot of residuals showed a relative random scatter suggesting equal variance of the error. The final product of each model is shown below. All the QQ plots look good for a simple model, the all have some deviation but nothing that would indicate an issue. All the scatter plot have a nice random cloud with a few points of high leverage/high influence but nothing that will break our model.

Model 1 Model 2 Model 3

## Comparing Competing Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test Set Models** | R2 | Adjusted R2 | PRESS | Kaggle Score |
| Model 1 Forward/Square root | 0.875822 | NA | 1492288.316 | 0.14070 |
| Model 2 LASSO/Square root | 0.878173 | NA | 1475403.393 | 0.13919 |
| Model 3 Manual/log | 0.883353 | 0.879 | 31.29747974 | 0.14831 |

The second model is chosen as the model used for interpretation purposes because of the high R-square, the low PRESS score and the lowest Kaggle score. Model three has the highest R-square and the lowest PRESS score but for some reason has the highest Kaggle score making it the most unpredictive model.

## Parameter Interpretation

The sale price prediction is based on many factors, we limited the number of factors for our simple model to neighborhood and house type plus 10 different numeric factors with the results being the square root of the sale price. The predicted sale price is therefore the predicted value of the model squared. The model equations is as described. The predicted square root (sales price) = -1169.38 +15.20(OverallQual) + 0.053(grLivArea) + 14.85(BsmtFullBath) + 13.26(GarageCars) + 8.48(OVerallCond) + 0.015(TotBsmtSF) + 0.51(YearBlt) +0.00043(LotArea) + 0.16(YearRemodAdd) + 27.79 (BldgType 1Fam) + 20.68(BldgType2fmCon) + 6.69(BldgType Duplex) – 19.33(BldgType Twnhs) -15.85(Blmngtn) – 35.83 (Blueste) – 32.70(BrDale) – 34.36(BrkSide) -16.53(ClearCr) -31.98(CollgCr) -1.8(Crawford) -47.61(Edwards) – 37.51(Gilbert) – 51.73(IDOTTR) – 33.56 (MeadowV) – 40.77(Mitchel) – 37.20 (Names) + 5.91 (NoRidge) + 22.70 Nridght) – 50.83(OldTown) – 40.57(Sawyer) – 35.43(SawyerW) – 12.46(Somerset) + 29.17(StoneBr) – 21.11(Timber). The base (intercept) of the square root of the sale price is $-1169, this may seem odd, but this holding all other variables at zero. The reference neighborhood is Veeneker (the neighborhood that all other neighborhoods are compared against) with nothing changing for that neighborhood with all other factors held the same. The square root of the home sale price for house based on the neighborhood are $-15.84 for Bloomington Heights, $-35.84 for Bluestem, $-30.70 for Brairdale, $-34.36 for Brookside, $-16.54 for Clear Creek, $-31.99 for College Creek, $-1.81 for Crawford, $ -47.61 for Edwards, $-37.51 for Gilbert, -51.73 for Iowa DOT and RR, $-33.56 for Meadow village, -40.77 for Mitchell, $-37.20 for North Ames, $-17.76 for Northpark Villa, -40.67 for Northwest Ames, $5.92 for Northridge, $22.70 for Northridge Heights, $-50.83 for Old Town, $-40.15 for South & West of Iowa State University, -$40.57 for Sawyer, $-35.42 for Sawyer West, $-12.47 for Somerset, $29.17 for Stone Brook, and $-21.11 for Timerland holding all other factor constant.

Next, the type of structure variable plays a part in the prediction of the square root of the home sale price. The reference structure type is Townhouse end unit causing no change in the house price. The square root for the home price based on structure type is $20.79 for single family, $20.69 for a two-family conversion, $6.70 for a duplex, and $-19.33 for a Townhouse inside unit holding all other factors constant. The square root of predicted home sale price increase by $1 for every 0.51 years based upon the year built holding all other factors constant. The square root of predicted home sale price increase by $15.20 for every for every level of the overall quality of the home on a 10-point scale, holding all other factors constant. The square root of predicted home sale price increase by $0.05 for every square foot of above ground living space, holding all other factors constant. The square root of predicted home sale price increase by $14.85 for each additional full bathroom in basement, holding all other factors constant. The square root of predicted home sale price increase by $13.29 for car the garage can hold, holding all other factors constant. The square root of predicted home sale price increase by $8.48 for every for every level of the overall condition of the home on a 10-point scale, holding all other factors constant. The square root of predicted home sale price increase by $0.02 for every square foot in the basement, holding all other factors constant. The square root of predicted home sale price increase by $0.51 for the year in which the house was built, holding all other factor constant. The square root of predicted home sale price increase by $0.0004 for every square foot of lot space the home sits upon, holding all other factors constant. The square root of predicted home sale price increase by $0.16 for the year in which the home was remodeled, holding all other factors constant.

|  |  |  |  | **95% Confidence Limits** | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Intercept** |  |  |  | -1451.245493 | | | | | -887.522872 | |
| **OverallQual** |  |  |  | 13.097117 | | | | | 17.294514 | |
| **GrLivArea** |  |  |  | 0.048662 | | | | | 0.057717 | |
| **Neighborhood Blmngtn** |  |  |  | -39.801368 | | | | | 8.119226 | |
| **Neighborhood Blueste** |  |  |  | -82.141524 | | | | | 10.472446 | |
| **Neighborhood BrDale** |  |  |  | -57.918758 | | | | | -7.487286 | |
| **Neighborhood BrkSide** |  |  |  | -54.993485 | | | | | -13.727013 | |
| **Neighborhood ClearCr** |  |  |  | -38.100192 | | | | | 5.031535 | |
| **Neighborhood CollgCr** |  |  |  | -50.942741 | | | | | -13.028334 | |
| **Neighborhood Crawfor** |  |  |  | -22.224338 | | | | | 18.599631 | |
| **Neighborhood Edwards** |  |  |  | -67.073316 | | | | | -28.149610 | |
| **Neighborhood Gilbert** |  |  |  | -57.268885 | | | | | -17.765057 | |
| **Neighborhood IDOTRR** |  |  |  | -73.416963 | | | | | -30.060178 | |
| **Neighborhood MeadowV** |  |  |  | -57.730460 | | | | | -9.393420 | |
| **Neighborhood Mitchel** |  |  |  | -60.874269 | | | | | -20.667493 | |
| **Neighborhood NAmes** |  |  |  | -55.995454 | | | | | -18.398166 | |
| **Neighborhood NPkVill** |  |  |  | -45.296463 | | | | | 9.780776 | |
| **Neighborhood NWAmes** |  |  |  | -60.140720 | | | | | -21.195677 | |
| **Neighborhood NoRidge** |  |  |  | -14.782178 | | | | | 26.618395 | |
| **Neighborhood NridgHt** |  |  |  | 2.993785 | | | | | 42.411431 | |
| **Neighborhood OldTown** |  |  |  | -71.008971 | | -30.645290 | |
| **Neighborhood SWISU** |  |  |  | -62.931204 | -17.374694 | |
| **Neighborhood Sawyer** |  |  |  | -60.226287 | -20.919424 | |
| **Neighborhood SawyerW** |  |  |  | -55.213535 | -15.636151 | |
| **Neighborhood Somerst** |  |  |  | -32.049984 | 7.112102 | |
| **Neighborhood StoneBr** |  |  |  | 7.296945 | 51.042529 | |
| **Neighborhood Timber** |  |  |  | -41.796882 | -0.424184 | |
| **Neighborhood Veenker** |  |  |  | . | . | |
| **BsmtFullBath** |  |  |  | 11.494352 | 18.204413 | |
| **GarageCars** |  |  |  | 10.329701 | 16.250942 | |
| **OverallCond** |  |  |  | 6.699208 | 10.264146 | |
| **TotalBsmtSF** |  |  |  | 0.010459 | 0.020359 | |
| **BldgType 1Fam** |  |  |  | 20.153040 | 35.427061 | |
| **BldgType 2fmCon** |  |  |  | 6.926584 | 34.447701 | |
| **BldgType Duplex** |  |  |  | -4.891689 | 18.282530 | |
| **BldgType Twnhs** |  |  |  | -31.313279 | -7.349547 | |
| **BldgType TwnhsE** |  |  |  | . | . | |
| **YearBuilt** |  |  |  | 0.382671 | 0.641449 | |
| **LotArea** |  |  |  | 0.000247 | 0.000613 | |
| **YearRemodAdd** |  |  |  | 0.041785 | 0.272046 | |

The 95% confidence intervals were calculated for every predicted value. The confidence intervals suggested 95% confidence that the mean square root of each predicted home price fell between +/- $61. This would translate to a 95% confidence interval for the sale price with a width of $14,884.

# Analysis of Question 2

The goal of question two was to create a valid model with the maximum predictive power. This model may be more complex than the first models. However, even a complex model should meet the assumptions of multiple linear regression and attempt to minimize bias.

Assumptions for all models

1. Linear relationship

All continuous variables are manually tested for linearity with the explanatory variable based on the scatter plots. Most of the variables are transformed in order to have better linear relationship. Only the variables who show at-least some linear relationship are included.

1. Homoscedasticity. This assumption requires that the variance of error terms are similar across the independent variables.

**Following plots are used to validate the above assumptions**

|  |  |  |  |
| --- | --- | --- | --- |
| ../../Desktop/Unknown.png  MSSubClass (Not linear) | ../../Desktop/Unknown.png  LotArea | ../../Desktop/Unknown.png  OverallQual | ../../Desktop/Unknown.png  OverallCond |
| ../../Desktop/Unknown.png  YearRemodAdd | ../../Desktop/Unknown.png  BsmtFinSF1 | ../../Desktop/Unknown-1.png  BsmtFinSF2 (Not linear) | ../../Desktop/Unknown.png  BsmtUnfSF (Not linear) |
| ../../Desktop/Unknown.png  TotalBsmtSF | ../../Desktop/Unknown-1.png  \_1stFlrSF | ../../Desktop/Unknown.png  \_2ndFlrSF | ../../Desktop/Unknown-1.png  LowQualFinSF (not linear) |
| ../../Desktop/Unknown.png  GrLivArea | ../../Desktop/Unknown-1.png  BsmtFullBath | ../../Desktop/Unknown.png  BsmtHalfBath (not linear) | ../../Desktop/Unknown.png  FullBath |
| ../../Desktop/Unknown.png  HalfBath | ../../Desktop/Unknown-1.png  BedroomAbvGr | ../../Desktop/Unknown-2.png  KitchenAbvGr (not linear) | ../../Desktop/Unknown.png  TotRmsAbvGrd |
| ../../Desktop/Unknown-1.png  Fireplaces | ../../Desktop/Unknown.png  GarageCars | ../../Desktop/Unknown-2.png  GarageArea | ../../Desktop/Unknown.png  WoodDeckSF |
| ../../Desktop/Unknown-1.png  OpenPorchSF (not linear) | ../../Desktop/Unknown.png  EnclosedPorch (not linear) | ../../Desktop/Unknown.png  \_3SsnPorch (not linear) | ../../Desktop/Unknown.png  ScreenPorch (not linear) |
| ../../Desktop/Unknown.png  PoolArea (not linear) | ../../Desktop/Unknown.png  MiscVal (not linear) | ../../Desktop/Unknown-1.png  MoSold (not linear) | ../../Desktop/Unknown.png  YrSold (not linear) |

1. Multivariate Normality. Multiple regression assumes that the variables are normally distributed. However, because all models exclude outliers with cooks’d larger than 4/1460 and because of the large number of training samples (1460) we can assume that the central limit theorem will be in affect and this assumption is validated.
2. No Multicollinearity. This assumption assumes that the independent variables are not highly correlated with each other. This assumption is tested by the Variance Inflation Factor (VIF) statistic. VIF > 10 indicates multicollinearity.

## Model Selection

### Type of Selection/Checking Assumptions

1. Model one was develop as a base line for creating a maximum predictive model. It uses the following chain of techniques

* The train and test datasets are imported and datatypes are normalized
* Most of the continuous variables are transformed based on previously created distribution charts. The goal of the transformations is to yield roughly normal distributions.
* Linearity with SalePrice is checked manually with scatterplots, and few of the variables are not included in the final dataset.
* The dataset includes 43 categorical variables. Because of the high number, they are transformed to dummy variables automatically with PROC GLMMOD
* High-influential points are automatically filtered based on a rule of having cook’s distance less than 4/N (N-number of training observations). Nine observations are dropped after this technique.
* An automatic feature selection method PRC GLMSELECT with a STEPWISE algorithm is used for selecting the final set of features. Five-fold cross validation is used to ensure better performance on the test dataset. All continuous variables and all dummy variables (299 in total) are fed into the feature selection. Forty-three variables are selected from the feature selection.
* A linear regression is fitted with all the selected features and the test dataset portion of the dataset is exported to a CSV file.

1. The second model was developed to test if variables with polynomial transformation of degree 2 will perform better. The intuition for this comes from scatter plots created to test the linearity of all continuous variables. Some of the variables seem to fit a curved line better. The model uses the following chain of techniques

* The train and test datasets are imported and datatypes are normalized
* Most of the continuous variables are transformed based on previously created distribution charts. The goal the transformation is having roughly normal distributions.
* Linearity with SalePrice is checked manually with scatterplots and few of the variables are not included in the final dataset.
* The dataset includes 43 categorical variables. Because of the high number, they are transformed to dummy variables automatically with PROC GLMMOD.
* High-influential points are automatically filtered based on rule of having cook’s distance less than 4/N (N-number of training observations). Nine observations are dropped after this technique.
* An automatic feature selection method PRC GLMSELECT with a STEPWISE algorithm is used for selecting the final set of features. Five-fold cross validation is used to ensure better performance on the test dataset. All continuous variables are raised to order of 2. The feature selection method is used to select best performing variables. 34 variables are selected out of 303.
* A linear regression is fitted with all the selected features and the test dataset portion of the dataset is exported to a CSV file.

1. After the model with polynomial transformation of degree 2 showed better performance, the next model was developed to test if cross features will have improvement. The intuition for this comes from logic that combinations are important. For example, a house with four bedrooms can be way more valuable if it has more than one bathroom, but bathrooms by itself don’t have such a great impact on the price. The model uses the following chain of techniques:

* The train and test datasets are imported and datatypes are normalized
* Most of the continuous variables are transformed based on previously created distribution charts. The goal the transformation is having roughly normal distributions.
* Linearity with SalePrice is checked manually with scatterplots and few of the variables are not included in the final dataset.
* The dataset includes 43 categorical variables. Because of the high number, they are transformed to dummy variables automatically with PROC GLMMOD
* High-influential points are automatically filtered based on rule of having cook’s distance less than 4/N (N-number of training observations). Nine observations are dropped after this technique.
* An automatic feature selection method PRC GLMSELECT with a LASSO algorithm is used for selecting the final set of features. Only 200 steps are allowed for the LASSO algorithm and five-fold cross validation is used to ensure better performance on the test dataset. All continuous variables are raised to order of 2 and two-way combinations between all features are created. The feature selection method is used to select best performing variables. All features with two-way combination and second order transformation have count of about 44,500. The 200 step LASO will select 159 out of them.
* A linear regression is fitted with all the selected features and the test dataset portion of the dataset is exported to a CSV file.

### Comparing Competing Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test Set Models** | R2 | Adjusted R2 | Kaggle  Score | PRESS |
| Model 1 | 0.948 | 0.9463 | 0.12883 | 475399 |
| Model 2 | 0.9433 | 0.9419 | 0.13413 | 490601 |
| Model 3 | 0.9567 | 0.9510 | 0.12360 | N/A |

# Conclusion/Discussion

The Ames, Iowa housing data provided a unique opportunity to build statistical models that can predict housing prices based on a combination of measured house traits. It also provided an opportunity to compare models with different end goals in mind. In the first question, three competing models were built with the end goal of yielding simple models that can be easily interpreted by home buyers, home sellers, etc. The highest performing simple model on Kaggle had a R2 value of 0.878173. With 10 quantitative variables and two categorical variables, the best simple model can be explained based on its individual parameters. And, if 87.8% of housing price variation can be explained by this combination of variables, real estate agents could potentially gather these particular data points on new homes with relative ease and predict their potential prices. In essence, the simple model is more functional because less data would need to be gathered on a single home in order to yield a fairly good estimation.

The second analysis question had the end goal of creating a most predictive model. The most successful model in this category could explain 95.67% of the variation in sale price. However, this predictive power was achieved at the price of simplicity. The highest performer on Kaggle includes combination variables and a final selection of 159 feature variables. If put to use, this model would require extensive data gathering for new homes. However, it is useful when a more accurate prediction needs to be made on future Ames, IA housing prices. In conclusion, multiple linear regression and model selection prove to be flexible, yet powerful techniques to building predictive models. And, just like most statistical techniques, they are most effective when they are used under the correct assumptions.

# PCA

## Introduction and Assumptions

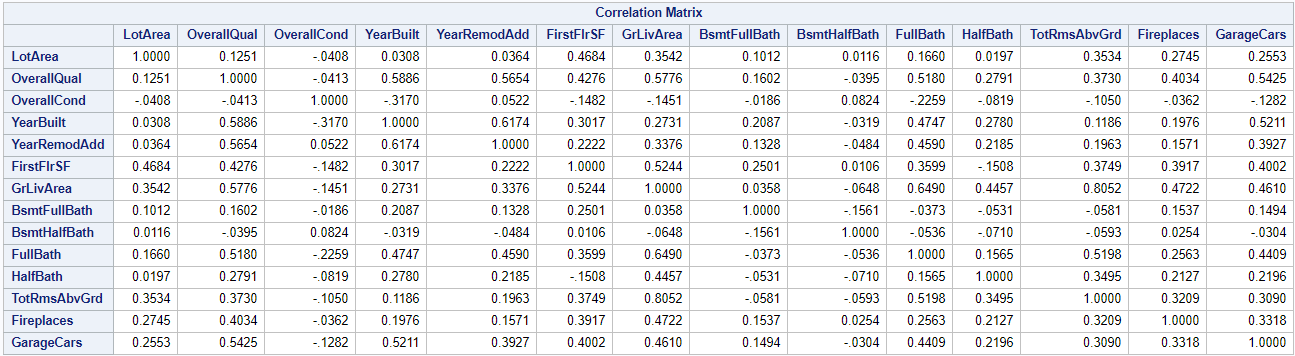
Principal components analysis (PCA) is primarily used in situations where there are a lot of explanatory variables, and many of them redundantly explain variance in the response variable. PCA reduces the explanatory variables into principal components that group correlated variables together in order to explain the most variance.

In the case of the Kaggle housing data, PCA is an appropriate tool because there is a large number of variables (~80). Since exploration of the Kaggle housing variables has previously been completed, this analysis starts with the pared down set of 14 quantitative variables. Variables excluded from the analysis include qualitative variables and variables that have a large number of missing values.

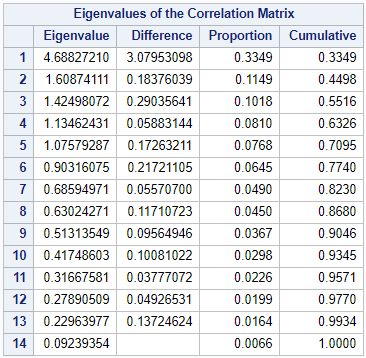
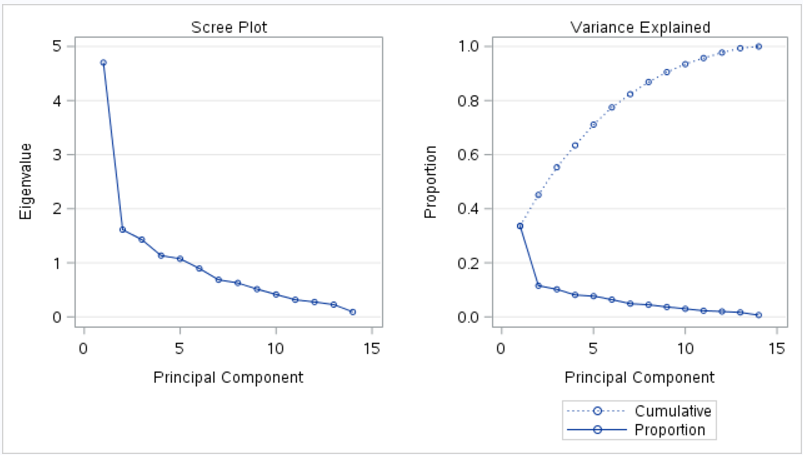
The assumptions of principal components analysis do not differ from the assumptions of multiple linear regression. The primary concerns are linear relationships between each explanatory variable and the response variable and normal distribution of each variable. For this PCA the response variable ‘SalePrice’ was transformed by a common logarithm. The 14 explanatory variables were added to 1 and then transformed by a natural logarithm. It was necessary to add 1 to every value for observations with a value 0. High influential points were excluded from the analysis on the basis of cook’s D > 4/1460 (the number of observations). Finally, the explanatory variables were standardized by subtraction of their respective means and division by their respective standard deviations. PCA with raw data tends to emphasize those variables with the highest variance. Also, raw data is only ideal if all variables have the same unit of measure. Standardization accounts for different units of measure and allows for all variables to be considered regardless of their variance size.

## PCA Analysis

The ‘Princomp’ procedure was used to perform a principal components analysis. The correlation matrix gives the first indicators of which explanatory variables are most strongly correlated. Not all variables had any medium-strong correlations. For example, overall condition is not strongly correlated to any of the other variables. On the opposite hand, total rooms above ground has a strong positive relationship with greater living area (0.8052).

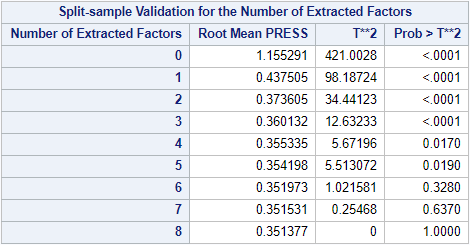


Because there are 14 explanatory variables, there are 14 dimensions to the plot of these variables. It is difficult to understand 14 dimensions, but luckily the majority of the variance is explained by only 8 of these dimensions. The table below lists the 14 components and their eigenvalues. Eigenvalues relate how much variance is in each eigenvector or direction. In order to choose how many principal components to build a model with, the below scree plots were assessed. The scree plots are a visualization of the proportion of variance and cumulative variance explained. The plot of the left makes it clear that the first principal component explains the largest proportion of variance. However, 8 principal components were chosen to build a model because the plot of the right shows that at about the 8th component the cumulative variance explained starts to level off.

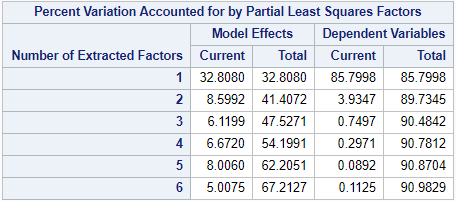
## Partial Least Squares Regression

Partial least squares analysis is a regression method particularly suited for principal component analysis. PLS was performed with a maximum number of factors set to 8. A five-fold split cross-validation approach was used to determine the optimal number of factors to include in the model. Because the root mean PRESS statistic continued to decrease from factor 7 to factor 8, all 8 factors were included in the model. The minimum root mean PRESS was 0.3253. However, it is apparent that the root mean PRESS with 8 factors is not very different from root mean PRESS with 4+ factors. The ‘cvtest’ option was added to the PLS model to test for significant difference of root mean PRESS statistic with the addition of each factor. The result, as seen in the below table, indicates that the root mean PRESS statistic is not significantly different for 6 factors as compared to having 7 or 8. Therefore, it is preferable to choose the model with 6 components rather than 8. This removal of two components improved the Kaggle score.



## Interpretation

Overall, the 6 component model accounts for 91.0% of the variance in the dependent variable log10SalePrice. And, it accounted for 67.2% of the variance within the independent variables (see table below).



The interpretation of each principal component in terms of the original variables is not straightforward. However, the principal components can be roughly broken down by variables with a correlation matrix between the principal components and the original variables. Below is the correlation table. Where an absolute correlation coefficient was equal to or greater than approximately 0.44 a red box was placed around it. This threshold is based on the largest correlation coefficient for principal component 6. Based on this threshold, the contributing variables for each component were surmised.

**Principal Component 1:** The first principal component offers explanation of the variance of lnSalePrice by the variables OverallQual, YearBuilt, YearRemodAdd, FirstFlrSF, GrLivArea, FullBath, TotRmsAbvGrd, Fireplaces, and GarageCars. With a correlation coefficient of 0.844 the greater living area explains the majority of the variance. Since all of these variables are positive it is expected that as one increases the others will too.

**Principal Component 2:** The second component will increase when the transformed LotArea variable increasesand the transformed YearBuilt decreases.

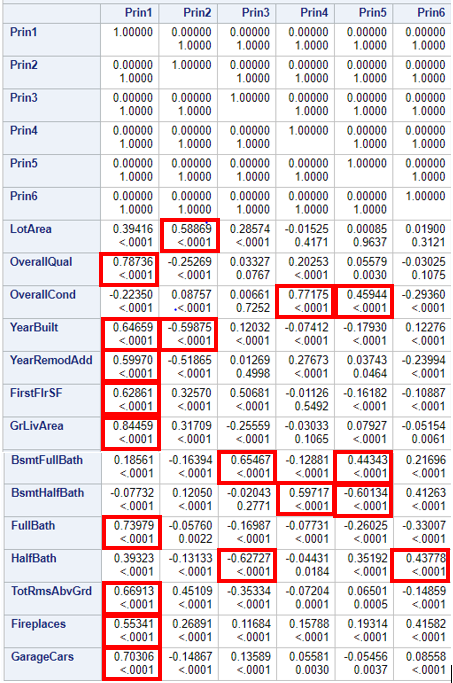
**Principal Component 3:** The third component will increase when the transformed value of BsmtFullBath increases and when the transformed value of HalfBath decreases.

**Principal Component 4:** The fourth component will increase when the transformed number of OverallCond increases and the transformed value of BsmtHalfBath increases.

**Principal Component 5:** The fifth component will increase when the transformed value of OverallCond increases, the transformed value of BsmtFullBath increases, and the transformed value of BsmtHalfBath decreases.

**Principal Component 6:** The final component will increase when the transformed value of HalfBath increases.

PCA allowed for the discovery of correlations between explanatory variables that were not necessarily intuitively predicted. For example, the number of fire places and the year built could explain some of the same variance in the log transformation of the sale price. PCA proved to be an efficient means of predicting the Sale Price of homes in Ames, Iowa by reducing the amount of redundant predictor variables.



# Comparing Competing Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test Set Models** | R2 | Adjusted R2 | Kaggle  Score | Minimum Root Mean Press |
| PCA Model | 0.9084 | N/A | 0.14193 | 0.3514 |
| LDA Model |  |  |  |  |

# Appendix

1. Question 1, Model 2

\*Data download and cleaning;  
\*importing train.CSV;  
Proc import datafile = "/home/sgozdzialski0/train.csv"  
out = train  
replace;  
delimiter=',';  
getnames=yes;  
run;  
\*importing the test data;  
proc import datafile = "/home/sgozdzialski0/test.csv"  
out= test  
replace;  
delimiter = ',';  
getnames = yes;  
run;  
   
   
\*Square root transformaiton on saleprice. To fix the none normality;  
data train2;  
set train;  
where saleprice > 100;  
Transprice = sqrt(saleprice);  
run;  
   
   
\* imputing datafile into test data set and adding saleprice to test dataset;  
data test;  
set test;  
saleprice =.;  
run;  
  
proc sort data=train2;  
by neighborhood;  
run;   
  
  
proc means data= train2;  
var saleprice;  
run;  
  
  
data Train3;  
set train2 test;  
Transprice = sqrt(saleprice);  
run;  
\*Running glmselect on variables that look good from the EDA, adding neighborhood and buildingtype  
because different nieghborhoods have different price ranges and differnet building type affect the   
cost;  
\*Using LASSO with Cross validation level 5 on the training data;  
proc glmselect data = train2 plot =all;  
class Neighborhood housestyle bldgtype;  
model Transprice = overallcond Grlivarea Neighborhood TotalBsmtSF Overallqual bsmtfullbath   
garagecars yearbuilt bldgtype housestyle lotarea YearRemodAdd TotalBsmtSF  
/selection=LASSO(stop=cv) cvmethod=random(5) hierarchy=single showpvalues;  
run;  
\*Model selected is Transprice = overallqual grLivArea Neighborhood BsmtFullbath GarageCars   
Overallcond TotalBsmtSF bldgType Yearbuilt LotArea yearRemodAdd;   
\*Running proc reg with VIF on all numeric variable to check for multicollinearity.   
All variables look good.;  
proc reg data= train2;  
model Transprice = overallqual grLivArea BsmtFullbath GarageCars Overallcond   
TotalBsmtSF Yearbuilt LotArea yearRemodAdd/VIF;  
run;  
  
\*Running GLM on the training data one last time to chekc all of the F scores and P-values to see  
if any of the slopes are zero. All variables look good.;  
proc GLM data = train2 plot = all;  
class Neighborhood bldgtype;  
model Transprice = overallqual grLivArea Neighborhood BsmtFullbath GarageCars Overallcond   
TotalBsmtSF bldgType Yearbuilt LotArea yearRemodAdd/Cli solution;  
run;  
  
\*Running final LASSO selected GLM on test data to get predictions;  
proc GLM data = train3 plot = all;  
class Neighborhood bldgtype;  
model Transprice = overallqual grLivArea Neighborhood BsmtFullbath GarageCars Overallcond   
TotalBsmtSF bldgType Yearbuilt LotArea yearRemodAdd/Cli solution;  
output out = result p = predict;  
run;  
  
\*Cleaning all unneed records of the front of the results.;  
data results;  
set result;  
predict = predict\*predict;  
if Predict > 0 then SalePrice = Predict;  
if Predict < 0 then SalePrice = 195000;  
keep id SalePrice;  
where id > 1460;  
run;  
  
\*Final Kaggle score of LASSO selection was 0.13919;

Question 2, Model 3

/\*

\* Import the datasets

\* train has 1460 obs and 81 columns

\* test has 1459 obs 80 columns

Char vars:

Alley BldgType BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtQual CentralAir Condition1 Condition2 Electrical ExterCond ExterQual Exterior1st Exterior2nd Fence FireplaceQu Foundation Functional GarageCond GarageFinish GarageQual GarageType Heating HeatingQC HouseStyle KitchenQual LandContour LandSlope LotConfig LotShape MSZoning MasVnrType MiscFeature Neighborhood PavedDrive PoolQC RoofMatl RoofStyle SaleCondition SaleType Street Utilities

Num vars:

BedroomAbvGr BsmtFinSF1 BsmtFinSF2 BsmtFullBath BsmtHalfBath BsmtUnfSF EnclosedPorch Fireplaces FullBath GarageArea GarageCars GrLivArea HalfBath KitchenAbvGr LotArea LowQualFinSF MSSubClass MiscVal MoSold OpenPorchSF OverallCond OverallQual PoolArea ScreenPorch TotRmsAbvGrd TotalBsmtSF WoodDeckSF YearBuilt YearRemodAdd YrSold \_1stFlrSF \_2ndFlrSF \_3SsnPorch

Extra vars:

Id

response variable:

SalePrice

\*/

**proc** **import**

datafile='/home/iangelov0/kaggle/train.csv'

out=train

dbms=CSV

replace;

getnames=yes;

datarow=**2**;

guessingrows=**2000**;

**proc** **import**

datafile='/home/iangelov0/kaggle/test.csv'

out=test

dbms=CSV

replace;

getnames=yes;

datarow=**2**;

guessingrows=**2000**;

/\*

\* Some vars in the test dataset have char data type instead of numeric

\* This will change their data type

\* Ugly but I don't know a better way to change data types

\*/

**data** test;

set test;

new1 = input(BsmtFinSF1, **8.**);

drop BsmtFinSF1;

if new1=. then new1=**0**;

rename new1=BsmtFinSF1;

new2 = input(BsmtFinSF2, **8.**);

drop BsmtFinSF2;

if new2=. then new2=**0**;

rename new2=BsmtFinSF2;

new3 = input(BsmtUnfSF, **8.**);

drop BsmtUnfSF;

if new3=. then new3=**0**;

rename new3=BsmtUnfSF;

new4 = input(TotalBsmtSF, **8.**);

drop TotalBsmtSF;

if new4=. then new4=**0**;

rename new4=TotalBsmtSF;

new5 = input(BsmtFullBath, **8.**);

drop BsmtFullBath;

if new5=. then new5=**0**;

rename new5=BsmtFullBath;

new6 = input(BsmtHalfBath, **8.**);

drop BsmtHalfBath;

if new6=. then new6=**0**;

rename new6=BsmtHalfBath;

if new6=. then new6=**0**;

rename new6=BsmtHalfBath;

new7 = input(GarageArea, **8.**);

drop GarageArea;

if new7=. then new7=**0**;

rename new7=GarageArea;

new8 = input(GarageCars, **8.**);

drop GarageCars;

if new8=. then new8=**0**;

rename new8=GarageCars;

/\*

\* Append the test dataset on the end of train

\*/

**data** dataset;

set train test;

/\*

\* Drop problematic vars

\*/

**data** dataset;

set dataset;

drop GarageYrBlt LotFrontage MasVnrArea;

/\*

\* Replace character missing values with 'NA'

\*/

**data** dataset;

set dataset;

array change \_character\_;

do over change;

if missing(change) then change='NA';

end;

if missing(change) then change='NA';

end;

/\*

\* Output some data

\*/

**proc** **print** data=dataset (obs=**10**);

**proc** **means** data=dataset;

**proc** **means** data=dataset N NMISS;

/\*

\* Variable transformations of the continious vars

\*/

/\* LotAreaTrans OverallQualTrans OverallCondTrans YearRemodAddTrans BsmtFinSF1 BsmtFinSF1Flag \*/

/\* TotalBsmtSFTrans TotalBsmtSFFlag \_1stFlrSFTrans \_2ndFlrSFFlag \_2ndFlrSFFlag GrLivAreaTrans \*/

/\* BsmtFullBath FullBath HalfBath BedroomAbvGrTrans KitchenAbvGr TotRmsAbvGrd Fireplaces \*/

/\* GarageCars GarageArea WoodDeckSF WoodDeckSFFlag \*/

**data** dataset;

set dataset;

SalePriceSqrt=sqrt(SalePrice);

drop SalePrice;

LotAreaTrans = log(LotArea);

OverallQualTrans = log(OverallQual);

OverallCondTrans = log(OverallCond);

YearRemodAddTrans = log(YearRemodAdd );

BsmtFinSF1Flag = **0**;

if BsmtFinSF1 = **0** then do;

BsmtFinSF1 = **700**;

BsmtFinSF1Flag = **1**;

end;

TotalBsmtSFTrans = sqrt(TotalBsmtSF);

TotalBsmtSFFlag = **0**;

if TotalBsmtSFTrans = **0** then do;

TotalBsmtSFTrans = sqrt(TotalBsmtSF);

TotalBsmtSFTrans = **30**;

TotalBsmtSFFlag = **1**;

end;

\_1stFlrSFTrans = log(\_1stFlrSF);

\_2ndFlrSFFlag = **0**;

if \_2ndFlrSF = **0** then do;

\_2ndFlrSF = **750**;

\_2ndFlrSFFlag = **1**;

end;

GrLivAreaTrans = log(GrLivArea);

BedroomAbvGrTrans = log(BedroomAbvGr);

WoodDeckSFFlag = **0**;

if WoodDeckSF = **0** then do;

WoodDeckSF = **200**;

WoodDeckSFFlag = **1**;

end;

keep Id SalePriceSqrt LotAreaTrans OverallQualTrans OverallCondTrans YearRemodAddTrans BsmtFinSF1 BsmtFinSF1Flag TotalBsmtSFTrans TotalBsmtSFFlag \_1stFlrSFTrans \_2ndFlrSFFlag \_2ndFlrSFFlag GrLivAreaTrans BsmtFullBath FullBath HalfBath BedroomAbvGrTrans KitchenAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF WoodDeckSFFlag Alley BldgType BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtQual CentralAir Condition1 Condition2 Electrical ExterCond ExterQual Exterior1st Exterior2nd Fence FireplaceQu Foundation Functional GarageCond GarageFinish GarageQual GarageType Heating HeatingQC HouseStyle KitchenQual LandContour LandSlope LotConfig LotShape MSZoning MasVnrType MiscFeature Neighborhood PavedDrive PoolQC RoofMatl RoofStyle SaleCondition SaleType Street Utilities;

/\*

\* Create dummy variables

\*/

**proc** **glmmod** data=dataset outdesign=dummies noprint;

class Alley BldgType BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtQual CentralAir Condition1 Condition2 Electrical ExterCond ExterQual Exterior1st Exterior2nd Fence FireplaceQu Foundation Functional GarageCond GarageFinish GarageQual GarageType Heating HeatingQC HouseStyle KitchenQual LandContour LandSlope LotConfig LotShape MSZoning MasVnrType MiscFeature Neighborhood PavedDrive PoolQC RoofMatl RoofStyle SaleCondition SaleType Street Utilities;

model Id = Alley BldgType BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtQual CentralAir Condition1 Condition2 Electrical ExterCond ExterQual Exterior1st Exterior2nd Fence FireplaceQu Foundation Functional GarageCond GarageFinish GarageQual GarageType Heating HeatingQC HouseStyle KitchenQual LandContour LandSlope LotConfig LotShape MSZoning MasVnrType MiscFeature Neighborhood PavedDrive PoolQC RoofMatl RoofStyle SaleCondition SaleType Street Utilities;

/\*

\* Merge the dummy vars with the dataste

\*/

**data** dataset;

merge dataset dummies;

/\*

\* Don't need the char vars anymore because they are represented by dummies

\* This will only keep numerical variables in the dataset

\*/

**data** dataset;

set dataset;

KEEP \_numeric\_;

**proc** **contents** data=dataset;

/\*

\* Get cooks'd for all continious vars

\* influance will contain the result

\*/

**proc** **reg** data=dataset plots=diagnostics;

model SalePriceSqrt=BsmtFinSF1 BsmtFullBath FullBath HalfBath KitchenAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF LotAreaTrans OverallQualTrans OverallCondTrans YearRemodAddTrans BsmtFinSF1Flag TotalBsmtSFTrans TotalBsmtSFFlag \_1stFlrSFTrans \_2ndFlrSFFlag GrLivAreaTrans BedroomAbvGrTrans WoodDeckSFFlag

/ noprint;

output out=influance cookd=cookd NOPRINT;

/\*

\* Remove all columns but cookd

\* Just to keep the dataset clean

\*/

**data** influance;

set influance;

keep cookd;

/\*

\* Merge the two datasets

\* dataset will now have a variable cookd

\*/

**data** dataset;

\*/

**data** dataset;

merge dataset influance;

/\*

\* Remove all very large cooks'd

\* A rule of thumb cutoff is 4/N (N - obs)

\*/

**data** dataset;

set dataset;

where cookd < **4**/**1460**;

drop cookd;

**proc** **print** data=dataset (obs=**10**);

**proc** **means** data=dataset;

/\*

\* Get only train part of the data

\*/

**data** train;

set dataset;

where SalePriceSqrt ^= .;

/\*

\* Variable Selection

\* other methods: LASSO stepwise(choose=BIC) selection=LASSO(stop=CV) cvMethod=RANDOM(20)\*

\*/

**proc** **glmselect** data=train;

model SalePriceSqrt=

BsmtFinSF1 BsmtFullBath FullBath HalfBath KitchenAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF LotAreaTrans OverallQualTrans OverallCondTrans YearRemodAddTrans BsmtFinSF1Flag TotalBsmtSFTrans TotalBsmtSFFlag \_1stFlrSFTrans \_2ndFlrSFFlag GrLivAreaTrans BedroomAbvGrTrans WoodDeckSFFlag Col1 -- Col276

/ selection=LASSO(stop=CV) cvMethod=RANDOM(**3**);

/\*

\* Fit the linear regression

\* and predict SalePriceSqrt for the test part of the dataset

\*/

**proc** **glm** data=dataset;

model SalePriceSqrt=BsmtFinSF1 BsmtFullBath FullBath HalfBath KitchenAbvGr Fireplaces GarageCars GarageArea WoodDeckSF LotAreaTrans OverallQualTrans OverallCondTrans YearRemodAddTrans BsmtFinSF1Flag TotalBsmtSFTrans GrLivAreaTrans WoodDeckSFFlag Col16 Col19 Col22 Col34 Col38 Col39 Col41 Col43 Col69 Col72 Col76 Col119 Col130 Col137 Col140 Col150 Col160 Col165 Col173 Col177 Col199 Col216 Col217 Col225 Col226 Col227 Col231 Col232 Col237 Col256 Col261;

output out=regout(where=(SalePriceSqrt=.)) p=predicted;

**proc** **print** data=regout (obs=**10**);

/\*

\* Create the final dataset

\*/

**data** submission;

set regout;

/\* since we took square root now have to addjuct back \*/

SalePrice = predicted \* predicted;

keep Id SalePrice;

**proc** **print** data=submission (obs=**10**);

/\*

\* Export the result to csv

\*/

**proc** **export** data=submission dbms=csv

outfile="/home/iangelov0/kaggle/submission.csv"

replace;

**run**;

Appendix-Project 2

\*/Principal Components Analysis/\*

proc import

datafile='/home/iangelov0/kaggle/train.csv'

out=train

dbms=CSV

replace;

getnames=yes;

datarow=2;

guessingrows=2000;

proc import

datafile='/home/iangelov0/kaggle/test.csv'

out=test

dbms=CSV

replace;

getnames=yes;

datarow=2;

guessingrows=2000;

data test;

set test;

new1 = input(BsmtFinSF1, 8.);

drop BsmtFinSF1;

if new1=. then new1=0;

rename new1=BsmtFinSF1;

new2 = input(BsmtFinSF2, 8.);

drop BsmtFinSF2;

if new2=. then new2=0;

rename new2=BsmtFinSF2;

new3 = input(BsmtUnfSF, 8.);

drop BsmtUnfSF;

if new3=. then new3=0;

rename new3=BsmtUnfSF;

new4 = input(TotalBsmtSF, 8.);

drop TotalBsmtSF;

if new4=. then new4=0;

rename new4=TotalBsmtSF;

new5 = input(BsmtFullBath, 8.);

drop BsmtFullBath;

if new5=. then new5=0;

rename new5=BsmtFullBath;

new6 = input(BsmtHalfBath, 8.);

drop BsmtHalfBath;

if new6=. then new6=0;

rename new6=BsmtHalfBath;

new7 = input(GarageArea, 8.);

drop GarageArea;

if new7=. then new7=0;

rename new7=GarageArea;

new8 = input(GarageCars, 8.);

drop GarageCars;

if new8=. then new8=0;

rename new8=GarageCars;

data dataset;

set train test;

SalePrice = log10(SalePrice);

data continuous;

set dataset;

drop Id;

drop SalePrice;

array vars \_numeric\_;

do over vars;

vars=log(vars+1);

end;

data continuous;

merge dataset continuous;

proc print data=continuous (obs=10);

proc means data=continuous;

/\*

\* Get cooks'd for all continious vars

\* influence will contain the result

\*/

proc reg data=continuous plots=diagnostics;

model SalePrice=LotArea OverallQual OverallCond YearBuilt YearRemodAdd \_1stFlrSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath TotRmsAbvGrd Fireplaces GarageCars

/ noprint;

output out=influence cookd=cookd NOPRINT;

/\*

\* Remove all columns but cookd

\* Just to keep the dataset clean

\*/

data influence;

set influence;

keep cookd;

/\*

\* Merge the two datasets

\* dataset will now have a variable cookd

\* Remove all very large cooks'd

\* A rule of thumb cutoff is 4/N (N - obs)

\* I use 5/N because only want to exclude more extreme outliers

\*/

data dataset;

merge dataset continuous influence;

data dataset;

set dataset;

where cookd < 5/1460;

drop cookd Intercept;

/\*

\* Standartize the continuous variables

\*/

PROC STANDARD data=dataset mean=0 std=1 out=dataset;

var LotArea OverallQual OverallCond YearBuilt YearRemodAdd \_1stFlrSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath TotRmsAbvGrd Fireplaces GarageCars;

run;

/\*

\* Output some data

\*/

proc print data=dataset (obs=10);

proc means data=dataset;

proc means data=dataset N NMISS;

/\*

\* Get PCA analysis

\*/

proc princomp data=dataset;

var LotArea OverallQual OverallCond YearBuilt YearRemodAdd \_1stFlrSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath TotRmsAbvGrd Fireplaces GarageCars;

run;

/\*

\* create regression with PCA

\*/

ods graphics on;

proc pls data=dataset cv=SPLIT(4) nfac=8 cvtest(1234);

model SalePrice = LotArea OverallQual OverallCond YearBuilt YearRemodAdd \_1stFlrSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath TotRmsAbvGrd Fireplaces GarageCars;

output out=regout(where=(SalePrice=.)) p=predicted;

run;

/\*

\* Create the final dataset

\*/

data submission;

set regout;

/\* since we took square root now have to addjuct back \*/

SalePrice = 10 \*\* predicted;

keep Id SalePrice;

/\*

\* Export the result to csv

\*/

proc export data=submission dbms=csv

outfile="/home/iangelov0/kaggle/submission.csv"

replace;

/\*

\* Kaggle score: 0.14266

\*/

run;

quit;