Housing Prices

# 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. 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 will provide a great framework for predictive modeling in other housing markets.

This project uses multiple linear regression and model selection techniques to build 6 total models that fall evenly into two categories. These categories are essentially simple and complex. The first goal is to develop 3 competing models that are simple and therefore more easily interpretable by parties working in the home marketplace. The second goal is to develop 3 competing models that have the greatest predictive power even if they include a much more complex number of variables.

# 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:

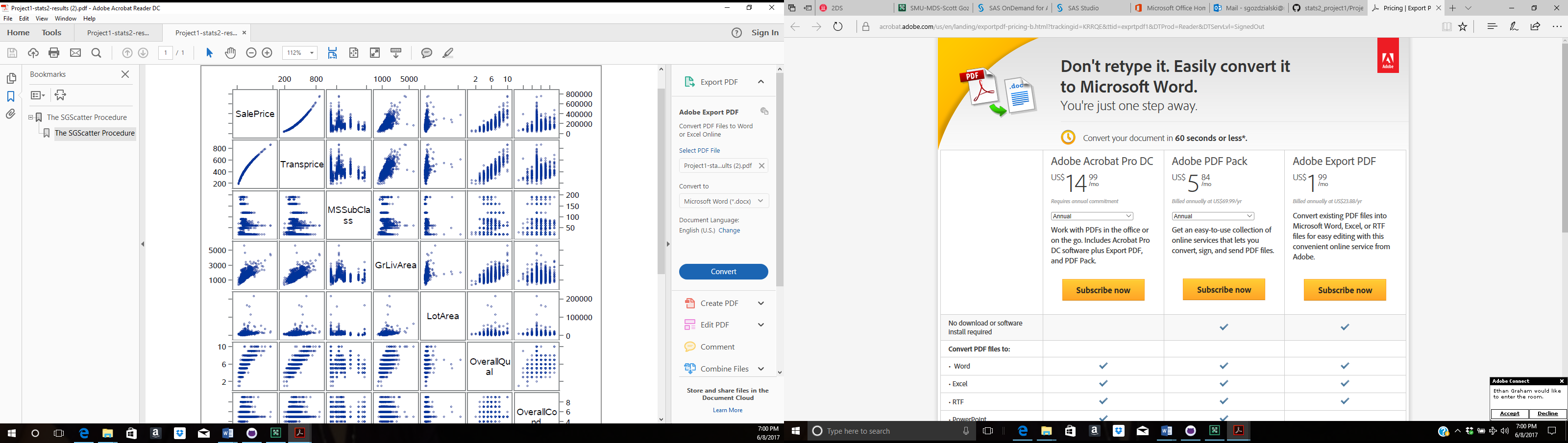
1. Train data set has 1460 observations and 81 columns
2. Test data set has 1459 observations 80 columns

Furthermore, 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, being a simple model we did not want to complicate thing, and left it with the little issue of non-normality knowing the number of observations should correct that last bit of 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.



Next, we ran used a the PROC MEAN function in SAS 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 a large number of missing values or a large majority of observations in a single category were considered to be left out of the model.

# Analysis of Question 1

The goal of question one was 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

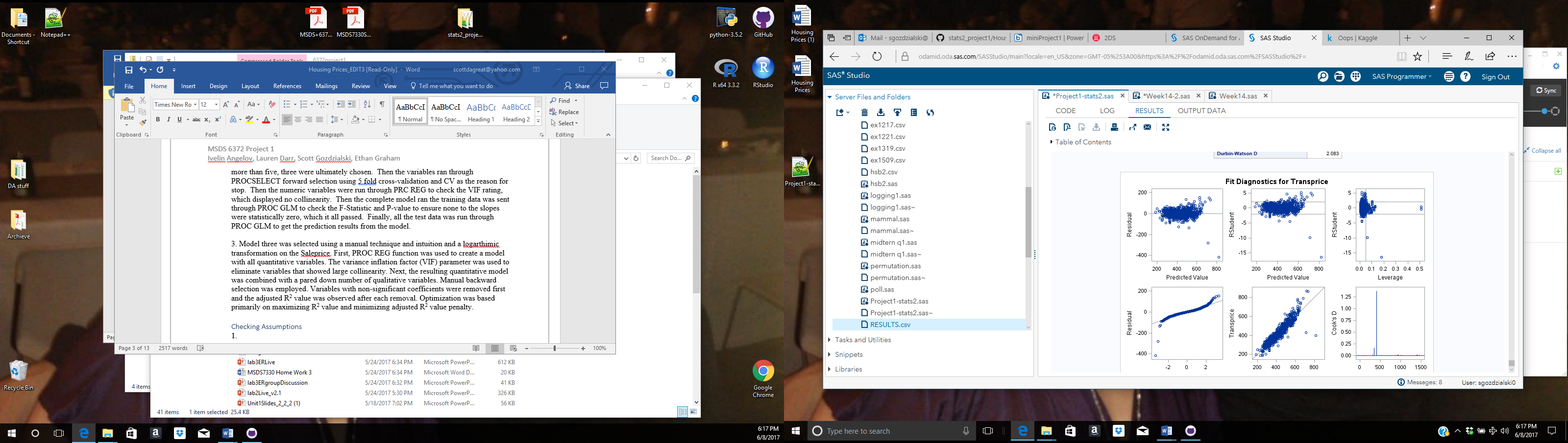
### Type of Selection

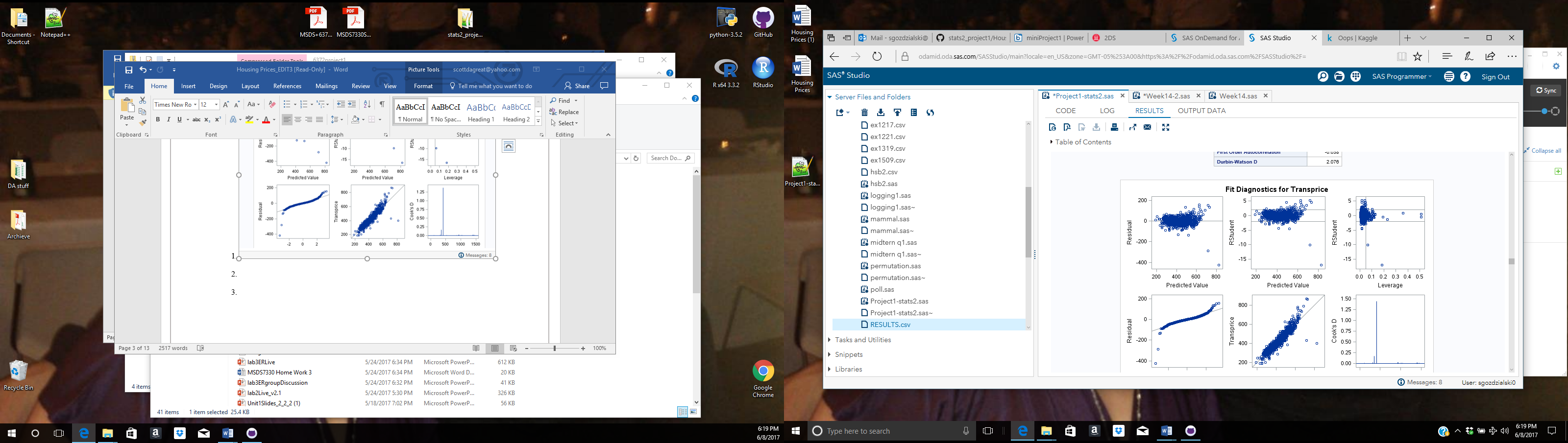
1.Model one was developed by using the forward automatic selection with a limited number of variables. The square root transformation was used on sales prices to help with the non-normality of the distribution. The first step PROC SGSCATTER was used to get a matrix of the numeric variables, which was analyized 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.

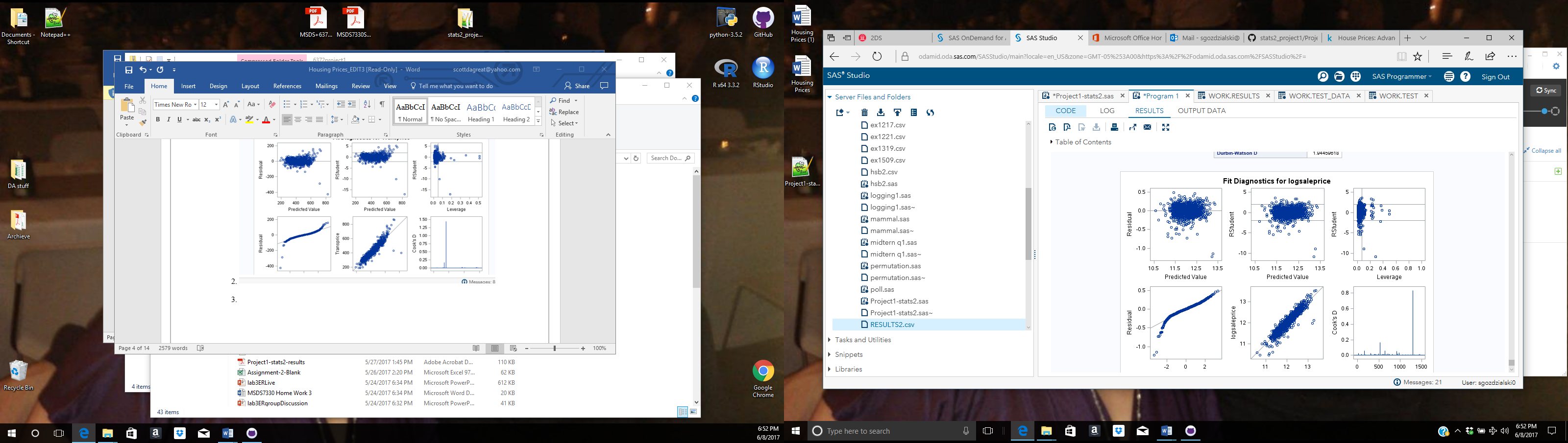
2. Model two was developed by using the LASSO automatic selection with a limited number of variables. The square root transformation was used on sales prices to help with the non-normality of the distribution. The model was more inclusive than model one but was run through the same process as model one, list again in this paragraph. The first step proc sgscatter was used to get a matrix of the numeric variables, which was analyized 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 and intuition and a logarthimic transformation on the Saleprice. 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

1. 

2. 

1. 

Model three 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.

### 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 |  | 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

1.

# 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.

## Model Selection

### Type of Selection

1.

2.

3.

### Checking Assumptions

1.

2.

3.

### Comparing Competing Models

# Conclusion/Discussion

# Appendix

/\*

\* 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 Col2 Col3 Col4 Col5 Col6 Col7 Col8 Col9 Col10 Col11 Col12 Col13 Col14 Col15 Col16 Col17 Col18 Col19 Col20 Col21 Col22 Col23 Col24 Col25 Col26 Col27 Col28 Col29 Col30 Col31 Col32 Col33 Col34 Col35 Col36 Col37 Col38 Col39 Col40 Col41 Col42 Col43 Col44 Col45 Col46 Col47 Col48 Col49 Col50 Col51 Col52 Col53 Col54 Col55 Col56 Col57 Col58 Col59 Col60 Col61 Col62 Col63 Col64 Col65 Col66 Col67 Col68 Col69 Col70 Col71 Col72 Col73 Col74 Col75 Col76 Col77 Col78 Col79 Col80 Col81 Col82 Col83 Col84 Col85 Col86 Col87 Col88 Col89 Col90 Col91 Col92 Col93 Col94 Col95 Col96 Col97 Col98 Col99 Col100 Col101 Col102 Col103 Col104 Col105 Col106 Col107 Col108 Col109 Col110 Col111 Col112 Col113 Col114 Col115 Col116 Col117 Col118 Col119 Col120 Col121 Col122 Col123 Col124 Col125 Col126 Col127 Col128 Col129 Col130 Col131 Col132 Col133 Col134 Col135 Col136 Col137 Col138 Col139 Col140 Col141 Col142 Col143 Col144 Col145 Col146 Col147 Col148 Col149 Col150 Col151 Col152 Col153 Col154 Col155 Col156 Col157 Col158 Col159 Col160 Col161 Col162 Col163 Col164 Col165 Col166 Col167 Col168 Col169 Col170 Col171 Col172 Col173 Col174 Col175 Col176 Col177 Col178 Col179 Col180 Col181 Col182 Col183 Col184 Col185 Col186 Col187 Col188 Col189 Col190 Col191 Col192 Col193 Col194 Col195 Col196 Col197 Col198 Col199 Col200 Col201 Col202 Col203 Col204 Col205 Col206 Col207 Col208 Col209 Col210 Col211 Col212 Col213 Col214 Col215 Col216 Col217 Col218 Col219 Col220 Col221 Col222 Col223 Col224 Col225 Col226 Col227 Col228 Col229 Col230 Col231 Col232 Col233 Col234 Col235 Col236 Col237 Col238 Col239 Col240 Col241 Col242 Col243 Col244 Col245 Col246 Col247 Col248 Col249 Col250 Col251 Col252 Col253 Col254 Col255 Col256 Col257 Col258 Col259 Col260 Col261 Col262 Col263 Col264 Col265 Col266 Col267 Col268 Col269 Col270 Col271 Col272 Col273 Col274 Col275 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**;