**Introduction**

**Data Description** could be found via <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>, where it is a Kaggle competition for modeling data. The train set is the modeling and the test set is what we use the model to obtain the sales price of homes sold from 79 different categories. **Where did the data come from?** The data came from a publication from the American Statistical Association about housing sales for residential homes in Ames, Iowa (<https://ww2.amstat.org/publications/jse/v19n3/decock.pdf>). **How big is it?** The train and test data set are almost identical in size, where the train data set is 1460 rows by 81 columns and the test data set is 1460 rows by 80 columns, with the sales price missing. **How many observations?** There are 80 and 79 observations either by class category or numerical for the train and test data set, respectively. **Where can we find out more?** We can find out more by reading the pdf or going to the Kaggle website, both stated above. **What are the specific variables that we need to know to understand with respect to your analysis?** The specific variables are in the appendix.

**Analysis Question 1:**

**Restatement of Problem:** The goal is to obtain the sales price of homes based on their living room space for the neighborhoods: NAmes, Edwards, and BrkSide. To obtain more power in the performance of the modeling, all of the neighborhoods are used to model sales price based on living room space by neighborhood.

**Specify the Model:** The model that we choose is based on the steps below for sales price and living room space by neighborhood. We are using all the data from every neighborhood because their data gives us more power to get a full picture of the housing market.

**Comparing Competing Models:** Non-transformed model, log-linear model, log-log model

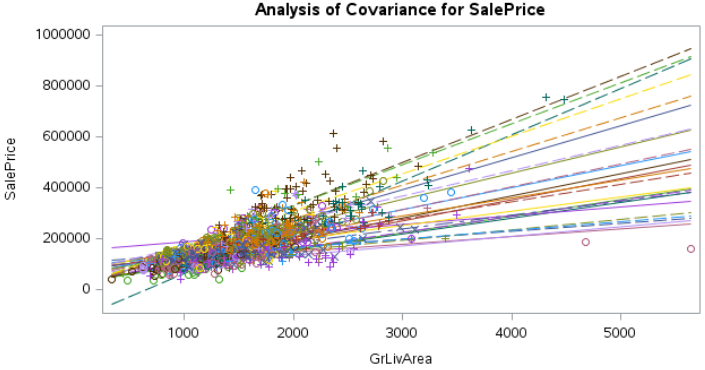


Fig 1: Sales price vs living area

When the sales price is plotted against the living area, there is a ‘fanning’ effect in the data (Fig 1), where the range in y increases as x increases and vice versa. Sales price is log transformed first because there are outliers that the multiple regression lines do not account for.

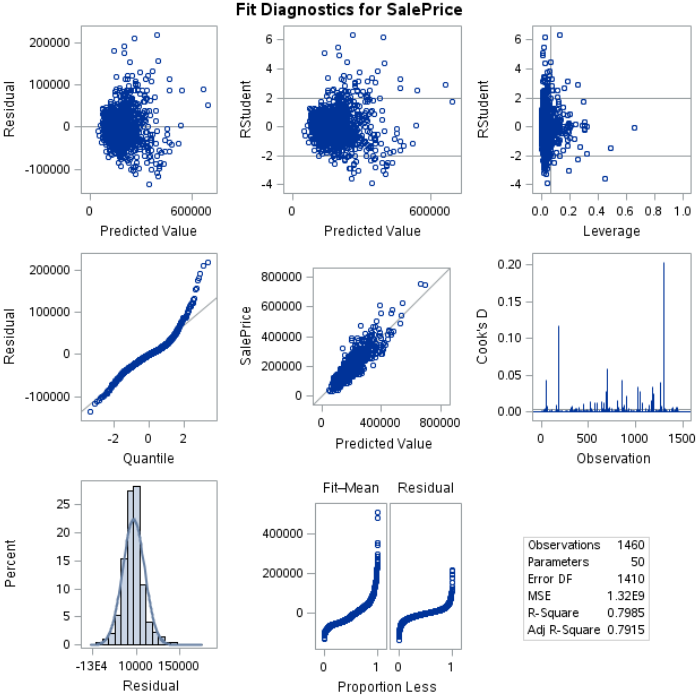


Fig 2: Diagnostic plots of regular data

Upon looking at the qq plot of the non-transformed data (Fig 2), we must do a log transformation of both the living area and the sales price because there are too many leverage points and outliers in the diagnostics for cook’s d and predicted residuals.

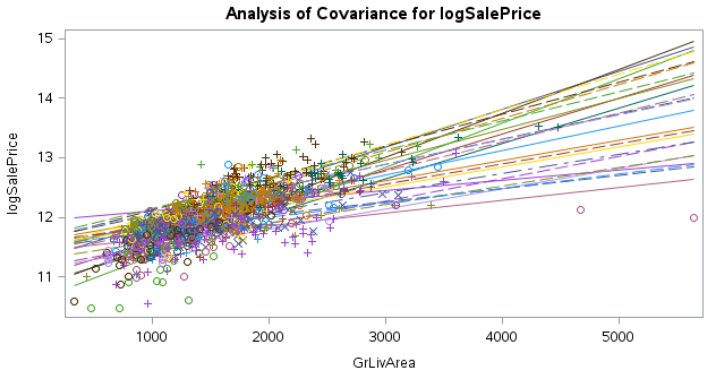


Fig 3: log(sales price) vs living area

Upon looking at the log transformed data for sale price and non-transformed data for the living area (Fig 3), we see that there are significant leverage points still. **That is why the log-log model is used and we will be using this for Analysis 1 and Analysis 2 (Fig 4)**.

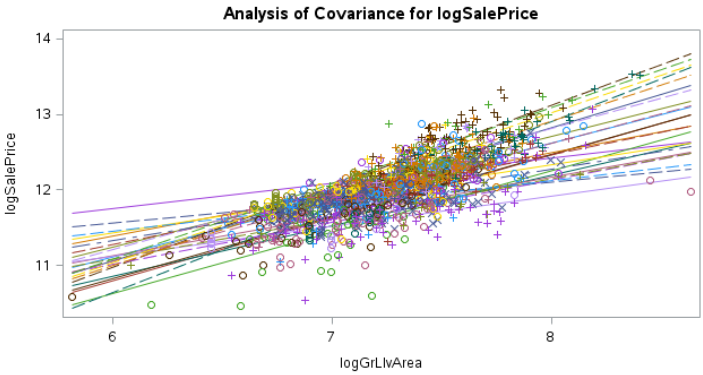


Fig 4: log log plot of sales price vs living area

**Checking Assumptions:** The model must have: 1) Linearity, 2) Residuals are normally distributed, 3) Independence, 4) Errors should have mean of 0 and have the same variance, 5) Equal SD: There is little evidence from the scatter plots of heteroscedasticity, 6) Normality of scatter plot, qq plot, and histogram of residuals

**Residual Plots:**

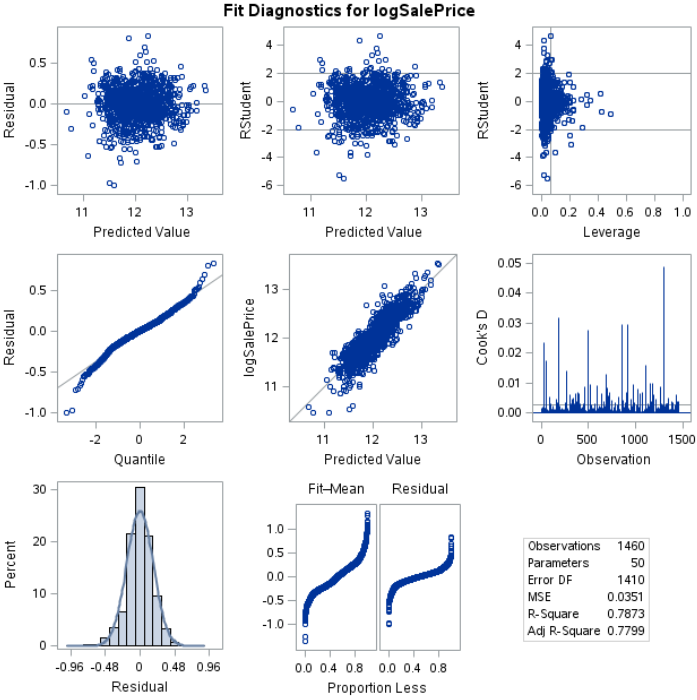
****

Fig 5: Diagnostic plots of log-log data

**Make sure and address each assumption:**

From the diagnostics panel, the qq-plot and predicted values is as linear as we could get after logging both x and y, with few outliers (1). The residual plot looks like a random cloud (5). The residual histogram shows a normal distribution with little skewness (6, 2). The data is independent because sales price is a correlation of living room space by neighborhood, but not as a result of it (3).

**Influential point analysis (Cook’s D and Leverage):** The rstudent values have the majority of the data within range of the limits, with the center value as 0 (4). There are some data points still causing leverage in the data. The cook’s D has some data points that would influence the linear regression of the data, but when doing further exploration, those data points are not caused by the 3 neighborhoods that we are trying to analyze.

**Adjusted R^2:** Full model is 0.787; Forwards is 0.7799; Backwards is 0.7799; Stepwise is 0.7680

**Interval CVPress:** Forwards elimination is the best considering CV press and adjusted R^2 (Fig 6)

**Forwards: 53.3 backwards: 53.5 Stepwise: 55.3**

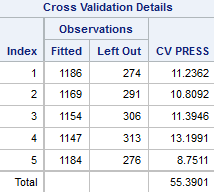
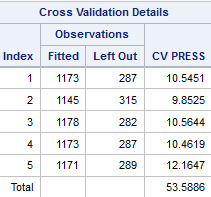
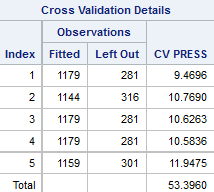
****

Fig 6: CV press results of eliminations

**Parameter Interpretation:** From the full model: (Fig 7, 8, 9)

ln(sales price) = β\_0 + β\_1\*names + β\_(…)\*neighborhood(…) + β\_x+1\*ln(living room space) + β\_x+2\*ln(living room space)\*names + β\_x+3\*ln(living room space)\*neighborhood(…)

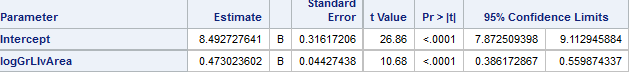


Fig 7: Names intercept and slope

Names has an intercept of 8.49, with a 95% confidence interval from 7.87 to 9.11.

Names has a slope of 0.47, with a 95% confidence interval from 0.38 to 0.55.

**ln(Names sales price) = 8.49 + 0.47\*ln(living room space)**

**Names sales price = e^8.49\***(**living room space)^0.47**

**Names sales price = 4865.86607325\*(living room space)^0.47**

**The predicted median of Names sales price with a living room space of 1 ft^2 = $4865. Doubling the living room space multiplies the predicted median Names sales price by 2^0.47 = 1.38510946811. In other words, it increases by 38%. For 100 ft^2 in Names, the sales price would be $42,379.**

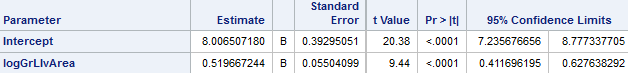
****

Fig 8: Edwards intercept and slope

**Edwards has an intercept of 8, with a 95% confidence interval from 7.23 to 8.77.**

**Edwards has a slope of 0.51, with a 95% confidence interval from 0.41 to 0.62.**

**ln(Edwards sales price) = 8 + 0.51\*ln(living room space)**

**Edwards sales price = e^8\*(living room space)^0.51**

**Edwards sales price = 2980.95798704\*(living room space)^0.51**

**The predicted median of Edwards sales price with a living room space of 1 ft^2 = $2980. Doubling the living room space multiplies the predicted median Edwards sales price by 2^0.51 = 1.4240501956. In other words, it increases by 42%. For 100 ft^2 in Edwards, the sales price would be $31,214.**

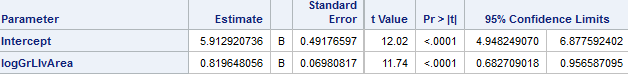
****

Fig 9: Brkside intercept and slope

**Brkside has an intercept of 5.91, with a 95% confidence interval from 4.94 to 6.87.**

**Brkside has a slope of 0.81, with a 95% confidence interval from 0.68 to 0.95.**

**ln(sales price) = 5.91 + 0.81\*ln(living room space)**

**Brkside sales price = e^5.91\*(living room space)^0.81**

**Brkside sales price = 368.706155409\*(living room space)^0.81**

**The predicted median of Brkside sales price with a living room space of 1 ft^2 = $368. Doubling the living room space multiplies the predicted median Brkside sales price by 2^0.81 = 1.75321144263. In other words, it increases by 75%. For 100 ft^2 in Brkside, the sales price would be $15,370.**

**Conclusion:** mostly stated above, limited room because of the 7 page rule.

**A short summary of the analysis:** It seems that the forward elimination is the best (or tied with backwards elimination) when evaluating sales price from living room space by neighborhood. The houses with the lowest starting price have the highest slopes when it comes to price based on living room space. The predictive equations based on models are stated above with starting prices and slopes, as well as the 100 ft^2 conversion. For more analysis of the data, more categories have to be considered. Note because of random seed, analysis changes upon running in SAS giving different results, which is why the full model is used.

**Analysis Question 2:**

**Restatement of Problem:** The goal is to create a predictive model of the sales price of homes based on all attributes of the train data while eliminating observational columns that are nonfactor towards the prediction of sales price from the test data. Low Kaggle score, low CV press, and high adjusted R value are used to determine the best model.

**Model Selection:** From Analysis 1, the log-log plot is used with the reasons provided above. The model that we choose is based on the steps below for predicting sales price based on numerical and categorical data.

**Type of Selection:** we explored logging and interaction and removing cross validating everything. We chose categorical data as classes and numerical data as modeling. Some numerical data such as year could be categorical. We use the 3 elimination procedures to find the most parameters to find the most important levels of class variables. From there, we reduced them down and tried not to over-fit or under-fit the data.

**Stepwise:** We havethe classes as:Neighborhood MSZoning Street Alley LotFrontage LotShape LandContour Utilities LotConfig LandSlope Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType and SaleCondition. The model is based on log\_SalePrice = log\_GrLivArea Neighborhood BldgType BsmtFinSF1 BsmtFullBath Condition2 KitchenQual LotArea OverallCond OverallQual SaleCondition and YearBuilt, with no interaction terms.

**Forward:** We have the classes as: Neighborhood MSZoning Street Alley LotFrontage LotShape LandContour Utilities LotConfig LandSlope Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType and SaleCondition. The model is based on log\_SalePrice = log\_GrLivArea Neighborhood BedroomAbvGr BldgType BsmtFinSF1 BsmtFullBath Condition2 FullBath Functional GarageCars Heating KitchenAbvGr KitchenQual LotArea OverallCond OverallQual SaleCondition ScreenPorch TotalBsmtSF WoodDeckSF YearBuilt and YearRemodAdd, with no interaction terms.

**Backward:** We have the classes as: Neighborhood MSZoning Street Alley LotFrontage LotShape LandContour Utilities LotConfig LandSlope Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType and SaleCondition. The model is based log\_SalePrice = log\_GrLivArea Neighborhood BedroomAbvGr BsmtFinSF1 BsmtFinSF2 BsmtFullBath BsmtUnfSF Condition2 FullBath Functional GarageArea Heating KitchenAbvGr KitchenQual LotArea MSSubClass OverallCond OverallQual SaleCondition ScreenPorch WoodDeckSF and YearBuilt, with no interaction terms.

**CUSTOM stepwise:** Neighborhood MSZoning Street Alley LotFrontage LotShape LandContour Utilities LotConfig LandSlope Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType and SaleCondition. The model is based on log\_SalePrice = log\_GrLivArea d1 d2 d3 d4 d6 d7 d8 d9 d11 d12 d13 dBT\_1 dBT\_5 BsmtFinSF1 BsmtFullBath dC2\_5 dC2\_3 dKQ\_2 dKQ\_3 LotArea OverallCond OverallQual dSC\_1 dSC\_2 dSC\_3 dSC\_6 YearBuilt, with no interaction terms.

**Checking Assumptions:** The model must have: 1) Linearity, 2) Residuals are normally distributed, 3) Independence, 4) Errors should have mean of 0 and have the same variance, 5) Equal SD: There is little evidence from the scatter plots of heteroscedasticity, 6) Normality of scatter plot, qq plot, and histogram of residuals

**Residual Plots:**

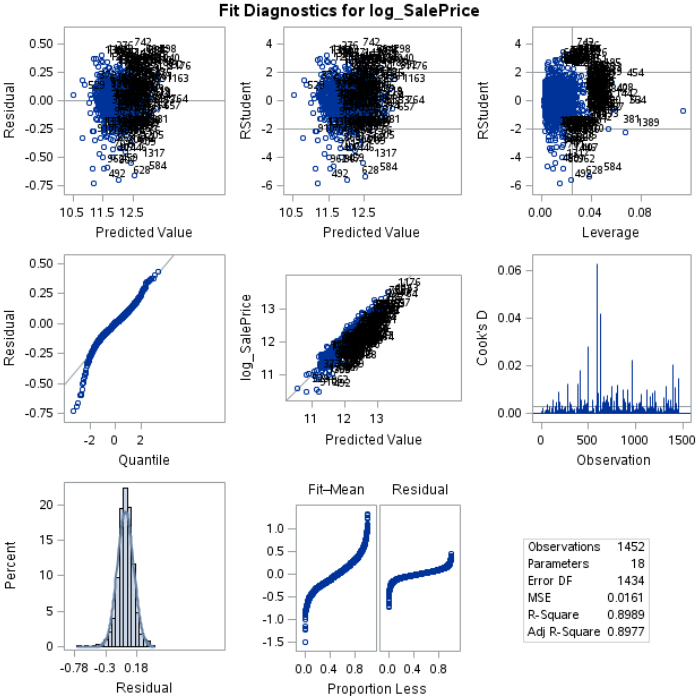
****

Fig 10: Diagnostic plots of log-log data

**Make sure and address each assumption:** From the diagnostics panel (Fig 10), the qq-plot and predicted values is as linear as we could get after logging both x and y, with few outliers (1). The residual plot looks like a random cloud (5). The residual histogram shows a normal distribution with little skewness (6, 2). The data is independent because sales price is a correlation of living room space by neighborhood, but not as a result of it (3).

**Influential point analysis (Cook’s D and Leverage):** The rstudent values have the majority of the data within range of the limits, with the center value as 0 (4). There are some data points still causing leverage in the data. The cook’s D has some data points that would influence the linear regression of the data, but when doing further exploration, those data points are not caused by the 3 neighborhoods that we are trying to analyze.

**Comparing Competing Models:**

****

Fig 11: Kaggle score

Here is the screenshot of the kaggle score when completing this assignment (Fig 11).

Table 1: Elimination results

|  |  |  |  |
| --- | --- | --- | --- |
| Predictive Models | Adjusted R^2 | CV PRESS | Kaggle Score |
| Forward | 0.9190 | 10.0200 | 0.12794 |
| Backward | 0.9067 | 26.7069 | 0.12108 |
| Stepwise | 0.8916 | 28.4796 | 0.13071 |
| CUSTOM | 0.8994 | 23.6992 | 0.13326 |

**Conclusion: A short summary of the analysis:** Backwards produced the best Kaggle score while Forward produced the best CV press, meaning that when compared to our own data, Forward elimination is the best process, but for predicting data, backwards perhaps could result in better predictions (Table 1). We are not presumptuous with one data over another. We used interaction and it did not seem to have any parameters eventually selected by automatic selection. We only included the best parameters by contribution and we checked variance inflation factors to ensure no redundancy in our model. We checked diagnostics for REG and GLM to see if any influential points and all assumptions hold when doing the prediction. **Our best Kaggle score is 0.12108.**

**Appendix:**

**Variable meaning:**

• SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.

• MSSubClass: The building class

• MSZoning: The general zoning classification

• LotFrontage: Linear feet of street connected to property

• LotArea: Lot size in square feet

• Street: Type of road access

• Alley: Type of alley access

• LotShape: General shape of property

• LandContour: Flatness of the property

• Utilities: Type of utilities available

• LotConfig: Lot configuration

• LandSlope: Slope of property

• Neighborhood: Physical locations within Ames city limits

• Condition1: Proximity to main road or railroad

• Condition2: Proximity to main road or railroad (if a second is present)

• BldgType: Type of dwelling

• HouseStyle: Style of dwelling

• OverallQual: Overall material and finish quality

• OverallCond: Overall condition rating

• YearBuilt: Original construction date

• YearRemodAdd: Remodel date

• RoofStyle: Type of roof

• RoofMatl: Roof material

• Exterior1st: Exterior covering on house

• Exterior2nd: Exterior covering on house (if more than one material)

• MasVnrType: Masonry veneer type

• MasVnrArea: Masonry veneer area in square feet

• ExterQual: Exterior material quality

• ExterCond: Present condition of the material on the exterior

• Foundation: Type of foundation

• BsmtQual: Height of the basement

• BsmtCond: General condition of the basement

• BsmtExposure: Walkout or garden level basement walls

• BsmtFinType1: Quality of basement finished area

• BsmtFinSF1: Type 1 finished square feet

• BsmtFinType2: Quality of second finished area (if present)

• BsmtFinSF2: Type 2 finished square feet

• BsmtUnfSF: Unfinished square feet of basement area

• TotalBsmtSF: Total square feet of basement area

• Heating: Type of heating

• HeatingQC: Heating quality and condition

• CentralAir: Central air conditioning

• Electrical: Electrical system

• 1stFlrSF: First Floor square feet

• 2ndFlrSF: Second floor square feet

• LowQualFinSF: Low quality finished square feet (all floors)

• GrLivArea: Above grade (ground) living area square feet

• BsmtFullBath: Basement full bathrooms

• BsmtHalfBath: Basement half bathrooms

• FullBath: Full bathrooms above grade

• HalfBath: Half baths above grade

• Bedroom: Number of bedrooms above basement level

• Kitchen: Number of kitchens

• KitchenQual: Kitchen quality

• TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

• Functional: Home functionality rating

• Fireplaces: Number of fireplaces

• FireplaceQu: Fireplace quality

• GarageType: Garage location

• GarageYrBlt: Year garage was built

• GarageFinish: Interior finish of the garage

• GarageCars: Size of garage in car capacity

• GarageArea: Size of garage in square feet

• GarageQual: Garage quality

• GarageCond: Garage condition

• PavedDrive: Paved driveway

• WoodDeckSF: Wood deck area in square feet

• OpenPorchSF: Open porch area in square feet

• EnclosedPorch: Enclosed porch area in square feet

• 3SsnPorch: Three season porch area in square feet

• ScreenPorch: Screen porch area in square feet

• PoolArea: Pool area in square feet

• PoolQC: Pool quality

• Fence: Fence quality

• MiscFeature: Miscellaneous feature not covered in other categories

• MiscVal: $Value of miscellaneous feature

• MoSold: Month Sold

• YrSold: Year Sold

• SaleType: Type of sale

• SaleCondition: Condition of sale

**Well commented SAS Code for Analysis 1 and 2:**

**Analysis 1**

FILENAME REFFILE '/home/yaoy890/sasuser.v94/my\_courses/Stats 1 files/train.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=train;

GETNAMES=YES;

RUN;

FILENAME REFFILE '/home/yaoy890/sasuser.v94/my\_courses/Stats 1 files/test.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=test;

GETNAMES=YES;

RUN;

data test;

set test;

SalePrice = .;

;

\*combine

data train2;

set train test;

run;

\*transform

data train3;

set train2;

logSalePrice = Log(SalePrice);

logGrLIvArea = Log(GrLIvArea);

run;

\*test train

proc gplot data=train3;

plot logsaleprice\*logGrLIvArea=Neighborhood;

run;

quit;

\*the three selection procs

proc glm data = train3 plots=diagnostics;

class Neighborhood (REF = "NAmes");

model logsaleprice=logGrLIvArea | Neighborhood /solution clparm ;

output out = t student=res cookd = cookd h = lev;

run; quit;

proc glm data = train3 plots=diagnostics;

class Neighborhood (REF = "Edwards");

model logsaleprice=logGrLIvArea | Neighborhood /solution clparm ;

run; quit;

proc glm data = train3 plots=diagnostics;

class Neighborhood (REF = "BrkSide");

model logsaleprice=logGrLIvArea | Neighborhood /solution clparm ;

run; quit;

proc glmselect data = train3 ;

class Neighborhood;

where Neighborhood in ("BrkSide", "Edwards", "NAmes") ;

model logsaleprice=logGrLIvArea | Neighborhood / selection= none details=summary cvdetails = cvpress showpvalues;

run;

\*exploration

data train4;

set train3;

If Neighborhood = "BrkSide" then Neighborhood = "aBrkSide";

If Neighborhood = "Edwards" then Neighborhood = "aEdwards";

If Neighborhood = "NAmes" then Neighborhood = "aNAmes";

run;

proc glmselect data = train3 plots = coefficientpanel ;

class Neighborhood (REF = "BrkSide");

model logsaleprice = logGrLIvArea | Neighborhood / selection = stepwise(stop = CV) cvdetails = CVPRESS showpvalues ;

run;

**Analysis 2:**

FILENAME REFFILE '/home/yaoy890/sasuser.v94/my\_courses/Stats 1 files/train.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=train;

GETNAMES=YES;

RUN;

FILENAME REFFILE '/home/yaoy890/sasuser.v94/my\_courses/Stats 1 files/test.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=test;

GETNAMES=YES;

RUN;

data test;

set test;

SalePrice = .;

;

\*combine

data train2;

set train test;

run;

\*transform

data train\_test\_log\_fullQ; set train2;

log\_SalePrice = log(SalePrice);

log\_GrLivArea = log(GrLivArea); run;

\*proc reg stuff

data reg\_test1Q; set train\_test\_log\_fullQ;

if Neighborhood = "BrkSide" then d1 = 1; else d1 = 0;

if Neighborhood = "Edwards" then d2 = 1; else d2 = 0;

if Neighborhood = "NAmes" then d3 = 1; else d3 = 0;

if Neighborhood = "BrDale" then d4 = 1; else d4 = 0;

if Neighborhood = "NridgHt" then d5 = 1; else d5 = 0;

if Neighborhood = "OldTown" then d6 = 1; else d6 = 0;

if Neighborhood = "Sawyer" then d7 = 1; else d7 = 0;

if Neighborhood = "SawyerW" then d8 = 1; else d8 = 0;

if Neighborhood = "Somerst" then d9 = 1; else d9 = 0;

if Neighborhood = "StoneBr" then d10 = 1; else d10 = 0;

if Neighborhood = "Gilbert" then d11 = 1; else d11 = 0;

if Neighborhood = "Mitchel" then d12 = 1; else d12 = 0;

if Neighborhood = "NWAmes" then d13 = 1; else d13 = 0;

int1 = d1\*GrLivArea; int2 = d2\*GrLivArea; int3 = d3\*GrLivArea; int4 = d4\*GrLivArea;

int5 = d5\*GrLivArea; int6 = d6\*GrLivArea; int7 = d7\*GrLivArea; int8 = d8\*GrLivArea;

int9 = d9\*GrLivArea; int10 = d10\*GrLivArea; int11 = d11\*GrLivArea; int12 = d12\*GrLivArea;

int13 = d13\*GrLivArea; intA = d1\*log\_GrLivArea; intB = d2\*log\_GrLivArea; intC = d3\*log\_GrLivArea; intD = d4\*log\_GrLivArea;

intE = d5\*log\_GrLivArea; intF = d6\*log\_GrLivArea; intG = d7\*log\_GrLivArea; intH = d8\*log\_GrLivArea; intI = d9\*log\_GrLivArea;

intJ = d10\*log\_GrLivArea; intK = d11\*log\_GrLivArea; intL = d12\*log\_GrLivArea; intM = d13\*log\_GrLivArea; run;

data reg\_test1Q\_3; set reg\_test1Q\_2;

if \_n\_ = 524 then delete;

if \_n\_ = 826 then delete; run;

data reg\_test1Q\_4; set reg\_test1Q\_3;

if \_n\_ = 31 then delete;

if \_n\_ = 314 then delete;

if \_n\_ = 336 then delete; run;

data reg\_test1Q\_5; set reg\_test1Q\_4;

if \_n\_ = 249 then delete; run;

data reg\_test1Q\_6; set reg\_test1Q\_5;

if \_n\_ = 702 then delete; run;

data reg\_test1Q\_7; set reg\_test1Q\_6;

if \_n\_ = 1292 then delete; run;

\*\*\*\*\*EXPLORATORY CODE TRUNCATED…..WE HAVE 28 PAGES OF THIS;

proc glmselect data = reg\_test1Q\_6 plots = coefficientpanel;

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

\*where Neighborhood in ("BrkSide", "Edwards", "NAmes", "BrDale", "NridgHt", "OldTown", "Sawyer", "SawyerW", "Somerst", "StoneBr", "Gilbert", "Mitchel", "NWAmes");

model log\_SalePrice = log\_GrLivArea Neighborhood BedroomAbvGr BldgType BsmtFinSF1 BsmtFullBath Condition2 FullBath Functional GarageCars Heating KitchenAbvGr KitchenQual LotArea OverallCond OverallQual SaleCondition ScreenPorch TotalBsmtSF WoodDeckSF YearBuilt YearRemodAdd / selection = **forward**(choose = CV include = 1) cvdetails = CVPRESS showpvalues;

output out = results p = Predict;

run;

proc glmselect data = reg\_test1Q\_6 plots = coefficientpanel;

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

\*where Neighborhood in ("BrkSide", "Edwards", "NAmes", "BrDale", "NridgHt", "OldTown", "Sawyer", "SawyerW", "Somerst", "StoneBr", "Gilbert", "Mitchel", "NWAmes");

model log\_SalePrice = log\_GrLivArea Neighborhood BedroomAbvGr BsmtFinSF1 BsmtFinSF2 BsmtFullBath BsmtUnfSF Condition2 FullBath Functional GarageArea Heating KitchenAbvGr KitchenQual LotArea MSSubClass OverallCond OverallQual SaleCondition ScreenPorch WoodDeckSF YearBuilt / selection = **backward**(choose = CV include = 1) cvdetails = CVPRESS showpvalues;

output out = results p = Predict;

run;

proc glmselect data = reg\_test1Q\_6 plots = coefficientpanel;

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

\*where Neighborhood in ("BrkSide", "Edwards", "NAmes", "BrDale", "NridgHt", "OldTown", "Sawyer", "SawyerW", "Somerst", "StoneBr", "Gilbert", "Mitchel", "NWAmes");

model log\_SalePrice = log\_GrLivArea Neighborhood BldgType BsmtFinSF1 BsmtFullBath Condition2 KitchenQual LotArea OverallCond OverallQual SaleCondition YearBuilt / selection = **stepwise**(choose = CV include = 1) cvdetails = CVPRESS showpvalues;

output out = results p = Predict;

run;

\*\*\*\*\*;

\*custom

proc glmselect data = reg\_test1Q\_7 plots = coefficientpanel;

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

\*where Neighborhood in ("BrkSide", "Edwards", "NAmes", "BrDale", "NridgHt", "OldTown", "Sawyer", "SawyerW", "Somerst", "StoneBr", "Gilbert", "Mitchel", "NWAmes");

model log\_SalePrice = log\_GrLivArea d1 d2 d3 d4 d6 d7 d8 d9 d11 d12 d13 dBT\_1 dBT\_5 BsmtFinSF1 BsmtFullBath dC2\_5 dC2\_3 dKQ\_2 dKQ\_3 LotArea OverallCond OverallQual dSC\_1 dSC\_2 dSC\_3 dSC\_6 YearBuilt / selection = stepwise(choose = CV include = 12) cvdetails = CVPRESS showpvalues;

output out = results p = Predict;

run;

\*\*\*\*\*;

\*custom2

proc reg data = reg\_test1Q\_7 plots(label)=(CooksD RStudentByLeverage) outest= logSP\_stats\_reg4 edf;

model log\_SalePrice = log\_GrLivArea d2 d6 d8 d11 d13 dBT\_1 dBT\_5 BsmtFinSF1 BsmtFullBath dC2\_5 dKQ\_3 LotArea OverallCond OverallQual dSC\_1 dSC\_2 dSC\_3 YearBuilt / VIF adjrsq;

run; quit;

\*custom 3

proc glm data = reg\_test1Q\_7 plots = all outstat = logSP\_stats\_glm4;

model log\_SalePrice = log\_GrLivArea d2 d6 d8 d11 d13 dBT\_1 dBT\_5 BsmtFinSF1 BsmtFullBath dC2\_5 dKQ\_3 LotArea OverallCond OverallQual dSC\_1 dSC\_2 dSC\_3 YearBuilt/ solution;

output out = results p = Predict;

run;