

# DDSAalytics - Employee Trends

*Randall Hendrickson*

*Rajat Chandna*

*Lokesh Maganti*

*Mike Kelleher*

*April 15, 2018*

## Contents

Introduction . . . . .	2
Data Description . . . . .	2
Question 1 . . . . .	2
Question 2 . . . . .	5
Appendix . . . . .	7
Question 1 SAS Code . . . . .	7
Question 1 Plots of Untransformed Data . . . . .	17
Image 001 . . . . .	17
Image 002 . . . . .	17
Image 003 . . . . .	17
Question 1 Plots of Log-Log Data . . . . .	19
Question 1 Miscellaneous Plots . . . . .	19



### House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting  
4,537 teams · 2 years to go

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#)

## Introduction

This is the final project for the semester in MSDS 6371 Statistical Foundation for Data Science, and is a group project. It is based on the Kaggle House Prices competition. Fundamentally we are seeking to answer what will the sale price of a house be based on some combination of predictive attribute measures of it. The 2 specific questions below prescribe distinct approaches to answering this question.

## Data Description

As mentioned above, the data comes from the Kaggle House Prices competition. The specific data sets we use for this project are the training data set (train.csv) to build the models, and test data set (test.csv) to cross-validate our models.

### train.csv

This dataset is roughly 450k with 1460 observations of 81 variables. To read and understand more about this specific dataset please refer to the Kaggle website at <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>.

### test.csv

This dataset, is roughly 441k with 1459 observation of 80 variables. This has one less observation: SalePrice. This is due to the fact it is meant to be used for cross-validation, and does not require this variable because we are meant to predict that with our models.

## Primary Variables

The following variables are fundamental to the answer for Question 1, do not contain missing values, and require no cleaning: \* SalePrice - the property's sale price in dollars \* GrLivArea - Above grade (ground) living area square feet \* Neighborhood - Physical locations within Ames city limits

## Question 1

### The Problem

Ames Century 21 real estate would like us to perform the following analysis of homes:

For the neighborhoods of: BrkSide, Edwards and NAmes, please find a predictive model for home SalePrice based on GrLivArea. Provide the model assumptions assessment with evidence, as well as data review and outlier analysis. Please provide estimate or estimates with confidence intervals, and a written conclusion of the relationship of GrLivArea and SalePrice within these neighborhoods.

### Build and Fit the Model

#### Untransformed Data

$$\text{SalePrice} = \text{beta0} + \text{beta1} * \text{GrLivArea} + \text{beta2} * d1 + \text{beta3} * d2 + \text{beta4} * d3 + \text{beta5} * \text{int1} + \text{beta6} * \text{int2} + \text{beta7} * \dots$$

When looking at plots of the raw data, there is a clear linear relationship between GrLivArea and SalePrice on Image 001. However, on Image 002 we can see several issues with the data: \* unequal spread on Residual vs Predicted and Studentized Residual vs Predicted, looks like the variance is increasing \* the qqplot is curved and the histogram is rather large in the middle suggesting non-normal data \* Cooks-D and Leverage show points that may have leverage and should be investigated

When looking at Image 003, most of the parameter estimates look good except for d3, int2 and int3. The VIF for all estimates are high, the smallest being 22. This is another sign that outliers should be investigated.

We need to perform outlier analysis and perform transformations to address equal deviation and normality.

## Other Transformations

Log-Linear, and Linear\_log were both experimented with. None of the violations were addressed using any of these methods. In some cases, it made the linear relationship worse with SalePrice. For brevity, no graphs are provided for these transformations.

## Log-Log Transformed Data

$$\text{SalePrice} = \text{beta0} + \text{beta1} * \text{LogGrLivArea} + \text{beta3} * \text{d2} + \text{beta4} * \text{d3} + \text{beta6} * \text{centLint2} + \text{beta7} * \text{centLint3}$$

For brevity, some of the iterative process for analyzing the data is omitted, but summarized here. It was found that there was no statistical difference between the BrkSide and Edwards neighborhoods, so these were combined into a single entity. Also, to address the high VIF values, the interaction variables were centered.

## Assumptions

When looking at the Log-Log transformed data in Image 004, the linear relationship looks much better. Also,, when looking at the fit diagnostics in Image 005, we can see that the assumptions are addressed: 1. There is a clear linear trend with explanatory variables with the LogSalePrice 1. The histogram looks better and the QQ plot as well, combined with the large number of observations, we can say the data is normal 1. All residual plots look like a good random cloud with equal variance across the graph 1. For independence, we do not know how the data were collected, and will assume that they are independent

The fit diagnostics in Image 005 and Regression statistics in Image 006 display that all issues are addressed with the data.

## Outlier Analysis

After analyzing the scatter plots, we noticed the two observations that were from homes over 4500 square feet. Since these two observations have high leverage, we decided to try to find other effects these observations might have. Looking at the group ranking of these observations we decided to remove these observations due to the effect these had on our model assumptions. We were more interested in the observations that were centered around the middle of the ranking.

Additionally, the GrLivArea of these observations for the Edwards Neighborhood is thought to be not representative of this area. These values were extremely above the third quartile. Give this we justified removing these observations.

## Comparison of Competing Models

Models	Adjusted R <sup>2</sup>	CV Press
Untransformed	0.5696	4.00198 E12
Log-Log	0.5808	97.939

Image 007 and Image 008 capture the SAS output for the Untransformed and Log-Log transformed models respectively.

## Models

1.  $\text{SalePrice} = 19972 + 87.163 \cdot \text{GrLivArea} + 68382 \cdot d1 + 54705 \cdot d2 + -12416 \cdot d3 + -57.412 \cdot \text{int1} + -32.847 \cdot \text{int2} + 31.351 \cdot \text{int3}$
2.  $\text{LogSalePrice} = 5.844 + 0.845 \cdot \text{LogGrLivArea} + 0.094 \cdot d2 + 0.219 \cdot d3 + -0.268 \cdot \text{centLint2} + 0.169 \cdot \text{centLint3}$

## Parameters

- Estimates - refer to model #2 above for the parameter estimates
- Interpretation
  - Reference category: BrkSide+Edwards
    - \* beta0 - The median SalePrice of houses that are located either in the Neighborhoods of BrkSide or Edwards, and has the GrLivArea of 1
    - \* beta1 - Doubling the GrLivArea is associated with 1.796 increase of the median SalePrice, roughly a 79.6% increase in median price.
  - Category: NNames
    - \* beta2 - The difference between the median price of a house in NNames and BrkSide+Edwards neighborhoods, that has a GrLivArea of 1
    - \* beta4 - The difference in slope of the regression line of LogGrLivArea when Neighborhood is NNames and the slope when the Neighborhood is equal to BrkSide+Edwards.
  - Category: Others
    - \* beta3 - The difference between the median price of a house in Others and BrkSide+Edwards neighborhoods, that has a GrLivArea of 1
    - \* beta5 - The difference in slope of the regression line of LogGrLivArea when Neighborhood is Others and the slope when the Neighborhood is equal to BrkSide+Edwards.
- Confidence Intervals - refer to Image 007 for 95% confidence interval for the individual parameters.

## Conclusion

The relationship between SalePrice on GrLivArea of houses in the BrkSide and Edwards Neighborhoods were not found to be significantly different. In general, when you double the GrLivArea of a house in the BrkSide and Edwards Neighborhoods, the median SalePrice increases by a factor of 79.6%. The 95% confidence interval for this increase is between a 72% and 82.4% increase in median SalePrice.

The NNames Neighborhood starting median price is about 10% more than the BrkSide and Edwards Neighborhoods. A doubling of the GrLivArea of houses in this Neighborhood is associated with a 49.2% increase in the median SalePrice. The 95% confidence interval for this increase is between a 32% and 69% increase in median SalePrice.

Because this was an observational study any causal inferences made from this would be speculative. But we can say there is a correlation between SalePrice and GrLivArea for this population of homes in Ames, Iowa.

## Question 2

### The Problem

Ames Century 21 real estate would like us to perform the following analysis of homes:

Build the most predictive model for sales prices of homes in all of Ames Iowa. By using 4 models: one from forward selection, one from backwards elimination, one from stepwise selection, and one the build custom.

The custom model is one of the three preceding models . The custom model is the best in terms of being able to predict future sale prices of homes .

### Model Selection Process

1. We ran heat map of correlation between sales price and the different numerical variables
2. we shortlisted the variables that has highest correlation with Sales price.
3. Then we did log transformation on all numeric variables and again check the correlation with sales price. Then identified some more variables that has high correlation with sales price.
4. After performing above step, we identify to 10 variables that has high correlation with sales price.
5. Then we did the matrix scatter plot between these 10 variables and sales price, to visualize the correlation between sales price and these variables; and visualize the correlation between explanatory variables.
6. We also found that among categorical variable ,neighbor hood and sales condition were relevant to predict the sales price . Adding them to the model , increased adjusted  $r^2$  and decreased the RMSE.
7. We removed outlier in variable GrLivArea that has  $< 4000$  and all the relevant variable is given to subset selection process.
8. We selected the model that has lowest cp score. That are Neighborhood, Sale Condition, OverallQual, LogLotArea, LogGrLiving, LogGarArea, LogTotalBsmtSF and YearBuilt.
9. We passed the same variable to Forward/Backward/Stepwise and we got same parameters back . All were significant.
10. We ran this model in SAS to validate model assumptions and we found the observations( # 31,411,632 and 1417) having high Studentized residuals that has 5 SD away. We investigated these observations and decided to delete them. 10 . Then reran the model and found out that all the assumptions has met
11. We ran the cross validation on the training data set and then obtain relevant cv test score.
12. We added sales price column to test data set, then ran the model . The observations and Predictions are inserted into a result data set, which was presented to Kaggle and results were given by Kaggle score . The results are shared in conclusion section.

### Forward

#### Assumptions

The Forward selection model is satisfying all below condition after performing transformation on some of the variables. \* Linear Trend (Plots: Response vs Explanatory, Explanatory vs Explanatory) \* Normality (Residual: Scatter, QQ Plot, Histogram) \* Equal Standard Deviation (Residual Scatter Plots)

### Backward

#### Assumptions

- Linear Trend (Plots: Response vs Explanatory, Explanatory vs Explanatory)
- Normality (Residual: Scatter, QQ Plot, Histogram)
- Equal Standard Deviation (Residual Scatter Plots)

## Stepwise

### Assumptions

- Linear Trend (Plots: Response vs Explanatory, Explanatory vs Explanatory)
- Normality (Residual: Scatter, QQ Plot, Histogram)
- Equal Standard Deviation (Residual Scatter Plots)

## Custom

### Assumptions

- Linear Trend (Plots: Response vs Explanatory, Explanatory vs Explanatory)
- Normality (Residual: Scatter, QQ Plot, Histogram)
- Equal Standard Deviation (Residual Scatter Plots)

## Comparison of Models

Predictive Models	Adjusted R <sup>2</sup>	CV Press	Kaggle Score
Forward	0.8801	23.2	.201
Backward	0.8801	23.2	.201
Stepwise	0.8801	23.2	.201
Custom*	0.8801	23.2	.201

- custom method chosen was Forward(select=SL) ### Conclusion Based on above table , all the models selection is deriving similar results . Any way we are going ahead with Stepwise model for deriving conclusion.

LogSalePrice=Neighborhood SaleCondition OverallQual LogLotArea LogGrLiving LogGarArea LogTotalBsmtSF YearBuilt

This model has adjusted  $R^2 = .8801$  , CV Press=23.2 and RMSE=.021 , this model fulfill all the assumptions of the multiple linear regression , that are 1. Residual: Scatter, QQ Plot, Histogram 2. Equal Standard Deviation (Residual Scatter Plots) 3. Cooks D is looking good. 4. We assume there is independence.

**The Kaggle score is .201**






4147	▼ 717	Pierre Olivier		0.20117	1	15d
4148	▼ 717	Madhav Puri		0.20126	1	1mo
4149	new	Randall H		0.20130	2	now
<b>Your Best Entry ↑</b> You advanced 429 places on the leaderboard! Your submission scored 0.20130, which is an improvement of your previous score of 0.30871. Great job! <a href="#">Tweet this!</a>						
4150	new	Clément D		0.20131	1	3d
4151	▼ 719	Matthias Diener		0.20131	3	2mo

Figure 1: Kaggle Score

## Appendix

### Question 1 SAS Code

```

/* Data Sets:
- Train0 - Kaggle Training Data Set
- TrainRed - Kaggle Training Data Set Reduced for question 1
*/
filename CSV URL "https://dl.dropboxusercontent.com/spa/afq05cp80hp4ezn/downloads/public/msds-data/train0.csv";

data TRAIN0;
%let _EFIERR_ = 0; /* set the ERROR detection macro variable */
infile CSV
delimiter = ',' MISSOVER DSD lrecl=32767 firstobs=2 ;
/* note regex used : \d\d\d\d[ ]* */
informat Id 4. ;
informat MSSubClass $3. ;
informat MSZoning $7. ; /* #31 c (all) */
informat LotFrontage best32. ;
informat LotArea best32. ;
informat Street $4. ;
informat Alley $4. ;
informat LotShape $3. ;
informat LandContour $3. ;
informat Utilities $6. ;
informat LotConfig $7. ;
informat LandSlope $3. ;
informat Neighborhood $7. ;
informat Condition1 $6. ;
informat Condition2 $6. ;
informat BldgType $6. ;
informat HouseStyle $6. ;
informat OverallQual best32. ;
informat OverallCond best32. ;
informat YearBuilt best32. ;
informat YearRemodAdd best32. ;
informat RoofStyle $7. ;
informat RoofMatl $7. ;
informat Exterior1st $7. ;

```

informat Exterior2nd \$7. ;  
 informat MasVnrType \$7. ;  
 informat MasVnrArea best32. ;  
 informat ExterQual \$2. ;  
 informat ExterCond \$2. ;  
 informat Foundation \$6. ;  
 informat BsmtQual \$2. ;  
 informat BsmtCond \$2. ;  
 informat BsmtExposure \$2. ;  
 informat BsmtFinType1 \$3. ;  
 informat BsmtFinSF1 best32. ;  
 informat BsmtFinType2 \$3. ;  
 informat BsmtFinSF2 best32. ;  
 informat BsmtUnfSF best32. ;  
 informat TotalBsmtSF best32. ;  
 informat Heating \$5. ;  
 informat HeatingQC \$2. ;  
 informat CentralAir \$1. ;  
 informat Electrical \$5. ;  
 informat \_1stFlrSF best32. ;  
 informat \_2ndFlrSF best32. ;  
 informat LowQualFinSF best32. ;  
 informat GrLivArea best32. ;  
 informat BsmtFullBath best32. ;  
 informat BsmtHalfBath best32. ;  
 informat FullBath best32. ;  
 informat HalfBath best32. ;  
 informat BedroomAbvGr best32. ;  
 informat KitchenAbvGr best32. ;  
 informat KitchenQual \$2. ;  
 informat TotRmsAbvGrd best32. ;  
 informat Functional \$4. ;  
 informat Fireplaces best32. ;  
 informat FireplaceQu \$2. ;  
 informat GarageType \$7. ;  
 informat GarageYrBlt best32. ;  
 informat GarageFinish \$3. ;  
 informat GarageCars best32. ;  
 informat GarageArea best32. ;  
 informat GarageQual \$2. ;  
 informat GarageCond \$2. ;  
 informat PavedDrive \$1. ;  
 informat WoodDeckSF best32. ;  
 informat OpenPorchSF best32. ;  
 informat EnclosedPorch best32. ;  
 informat \_3SsnPorch best32. ;  
 informat ScreenPorch best32. ;  
 informat PoolArea best32. ;  
 informat PoolQC \$2. ;  
 informat Fence \$5. ;  
 informat MiscFeature \$4. ;  
 informat MiscVal best32. ;  
 informat MoSold best32. ;  
 informat YrSold best32. ;



```

informat SaleType $5. ;
informat SaleCondition $7. ;
informat SalePrice best32. ;
format Id 4. ;
format MSSubClass $3. ;
format MSZoning $2. ;
format LotFrontage best12. ;
format LotArea best12. ;
format Street $4. ;
format Alley $4. ;
format LotShape $3. ;
format LandContour $3. ;
format Utilities $6. ;
format LotConfig $7. ;
format LandSlope $3. ;
format Neighborhood $7. ;
format Condition1 $6. ;
format Condition2 $6. ;
format BldgType $6. ;
format HouseStyle $6. ;
format OverallQual best12. ;
format OverallCond best12. ;
format YearBuilt best12. ;
format YearRemodAdd best12. ;
format RoofStyle $7. ;
format RoofMatl $7. ;
format Exterior1st $7. ;
format Exterior2nd $7. ;
format MasVnrType $7. ;
format MasVnrArea best12. ;
format ExterQual $2. ;
format ExterCond $2. ;
format Foundation $6. ;
format BsmtQual $2. ;
format BsmtCond $2. ;
format BsmtExposure $2. ;
format BsmtFinType1 $3. ;
format BsmtFinSF1 best12. ;
format BsmtFinType2 $3. ;
format BsmtFinSF2 best12. ;
format BsmtUnfSF best12. ;
format TotalBsmtSF best12. ;
format Heating $5. ;
format HeatingQC $2. ;
format CentralAir $1. ;
format Electrical $5. ;
format _1stFlrSF best12. ;
format _2ndFlrSF best12. ;
format LowQualFinSF best12. ;
format GrLivArea best12. ;
format BsmtFullBath best12. ;
format BsmtHalfBath best12. ;
format FullBath best12. ;
format HalfBath best12. ;

```

```

format BedroomAbvGr best12. ;
format KitchenAbvGr best12. ;
format KitchenQual $2. ;
format TotRmsAbvGrd best12. ;
format Functional $4. ;
format Fireplaces best12. ;
format FireplaceQu $2. ;
format GarageType $7. ;
format GarageYrBltn best12. ;
format GarageFinish $3. ;
format GarageCars best12. ;
format GarageArea best12. ;
format GarageQual $2. ;
format GarageCond $2. ;
format PavedDrive $1. ;
format WoodDeckSF best12. ;
format OpenPorchSF best12. ;
format EnclosedPorch best12. ;
format _3SsnPorch best12. ;
format ScreenPorch best12. ;
format PoolArea best12. ;
format PoolQC $2. ;
format Fence $5. ;
format MiscFeature $4. ;
format MiscVal best12. ;
format MoSold best12. ;
format YrSold best12. ;
format SaleType $5. ;
format SaleCondition $7. ;
format SalePrice best12. ;
input
  Id
  MSSubClass
  MSZoning $
  LotFrontage $
  LotArea
  Street $
  Alley $
  LotShape $
  LandContour $
  Utilities $
  LotConfig $
  LandSlope $
  Neighborhood $
  Condition1 $
  Condition2 $
  BldgType $
  HouseStyle $
  OverallQual
  OverallCond
  YearBuilt
  YearRemodAdd
  RoofStyle $
  RoofMatl $

```

Exterior1st \$  
 Exterior2nd \$  
 MasVnrType \$  
 MasVnrArea \$  
 ExterQual \$  
 ExterCond \$  
 Foundation \$  
 BsmtQual \$  
 BsmtCond \$  
 BsmtExposure \$  
 BsmtFinType1 \$  
 BsmtFinSF1  
 BsmtFinType2 \$  
 BsmtFinSF2  
 BsmtUnfSF  
 TotalBsmtSF  
 Heating \$  
 HeatingQC \$  
 CentralAir \$  
 Electrical \$  
 \_1stFlrSF  
 \_2ndFlrSF  
 LowQualFinSF  
 GrLivArea  
 BsmtFullBath  
 BsmtHalfBath  
 FullBath  
 HalfBath  
 BedroomAbvGr  
 KitchenAbvGr  
 KitchenQual \$  
 TotRmsAbvGrd  
 Functional \$  
 Fireplaces  
 FireplaceQu \$  
 GarageType \$  
 GarageYrBltd \$  
 GarageFinish \$  
 GarageCars  
 GarageArea  
 GarageQual \$  
 GarageCond \$  
 PavedDrive \$  
 WoodDeckSF  
 OpenPorchSF  
 EnclosedPorch  
 \_3SsnPorch  
 ScreenPorch  
 PoolArea  
 PoolQC \$  
 Fence \$  
 MiscFeature \$  
 MiscVal  
 MoSold

```

YrSold
SaleType $
SaleCondition $
SalePrice
;
if _ERROR_ then call symputx('_EFIERR_',1); /* set ERROR detection macro variable */
run;

/* ***** */
/* DATA CLEANUP - use proc means to show missing values*/
/* ***** */
title "Train0 - Data Set Missing Values";
proc means data=train0 nmiss n; run;

/* fix missing values */
data train0;
  set train0;
  if missing(LotFrontage) then LotFrontage = 0;
run;

title "Train0 - Data Set Fixed Missing Values";
proc means data=train0 nmiss n; run;

title "Train0 - First 10";
proc print data=train0 (obs=10);
run;
title "";
/* ***** */
/* END DATA CLEANUP - after second proc means, no missing values */
/* ***** */
/*****/
/** train0 is for Question 2 **/
/** trainraw3 is Q1 Unmodified **/
/**      / Untransformed Data **/
/*****/
data train0raw;
  set train0 (keep=Id Neighborhood GrLivArea SalePrice);

  select (Neighborhood);
    when ("NAMES")   Nei=Neighborhood;
    when ("Edwards") Nei=Neighborhood;
    when ("BrkSide") Nei=Neighborhood;
    otherwise Nei="Others";
  end;

  if Nei="Edwards" then d1=1; else d1=0;
  if Nei="NAMES"   then d2=1; else d2=0;
  if Nei="Others"  then d3=1; else d3=0;

  int1 = d1 * GrLivArea;
  int2 = d2 * GrLivArea;
  int3 = d3 * GrLivArea;
run;

```

```

proc sort data=train0raw;
  by descending GrLivArea;
run;

proc means data=train0raw;
  var GrLivArea d1 d2 d3;
run;

proc sgscatter data=train0raw;
  plot SalePrice*GrLivArea / group=Nei reg;
run;
/*****
/*****      Model Building      *****/
/* Residual Plot showing possible unequal spread) */
/* Model for Untransformed Data */
proc glm data=train0raw plots=all;
  class Nei(ref="BrkSide");
  model SalePrice=GrLivArea d1 d2 d3 int1 int2 int3 | Nei / solution VIF;
  TITLE "Non Transformed";
run;

proc reg data=train0raw plots=all;
  model SalePrice = GrLivArea d1 d2 d3 int1 int2 int3 / VIF;
  TITLE "Un-Transformed Data";
run;
/*****
/***** Question 1 Tranformed Data *****/
/*****
/*****
/***** train0 is for Question 1 - final model data **/
/***** outliers are removed as well *****/
/*****
data trainRed;
  set train0 (keep=Id Neighborhood GrLivArea SalePrice);

  select (Neighborhood);
    when ("NAMES") Nei=Neighborhood;
    when ("Edwards") Nei="EdwardP";
    when ("BrkSide") Nei="EdwardP";
    otherwise Nei="Others";
  end;

  /* Remove outliers */
  if GrLivArea < 4500;
  /* make sure we do not include missing values */
  if GrLivArea > 0;
  /* make sure we do not include missing values and we dont take log of zero */
  if SalePrice > 0;
    logSales=log(SalePrice);
    logLiv=log(GrLivArea);

  if Nei="NAMES" then d2=1; else d2=0;
  if Nei="Others" then d3=1; else d3=0;

```

```

l_int2 = d2 * logLiv;
l_int3 = d3 * logLiv;

centLint2=(logLiv - 7.2660236) * (d2 - 0.1543210);
centLint3=(logLiv - 7.2660236) * (d3 - 0.7386831);
run;

proc sort data=trainRed;
  by descending logLiv;
run;

proc means data=trainRed;
  var logLiv d2 d3;
run;

proc sgscatter data=trainRed;
  plot logSales*logLiv / reg;
  TITLE "Log-Log Transformed Data";
run;

/*****
/*****      Model Building      *****/

proc glm data=trainRed plots=all;
  class Nei(ref="EdwardP");
  model logSales=logLiv | Nei / solution;
  TITLE "Log-Log Transformed Data";
run;

proc reg data=trainRed plots=all;
  model logSales=logLiv d2 d3 centLint2 centLint3 / VIF;
  TITLE "Log-Log Transformed Data";
run;

/*****
/*****      Model Comparison      *****/
proc reg data=trainRaw plots=all;
  model SalePrice = GrLivArea d1 d2 d3 int1 int2 int3 / PRESS VIF CLM P;
  TITLE "Un-Transformed Data";
run;

proc reg data=trainRed plots=all;
  model logSales=logLiv d2 d3 centLint2 centLint3 / PRESS VIF CLM P;
  TITLE "Log-Log Transformed Data";
run;

```

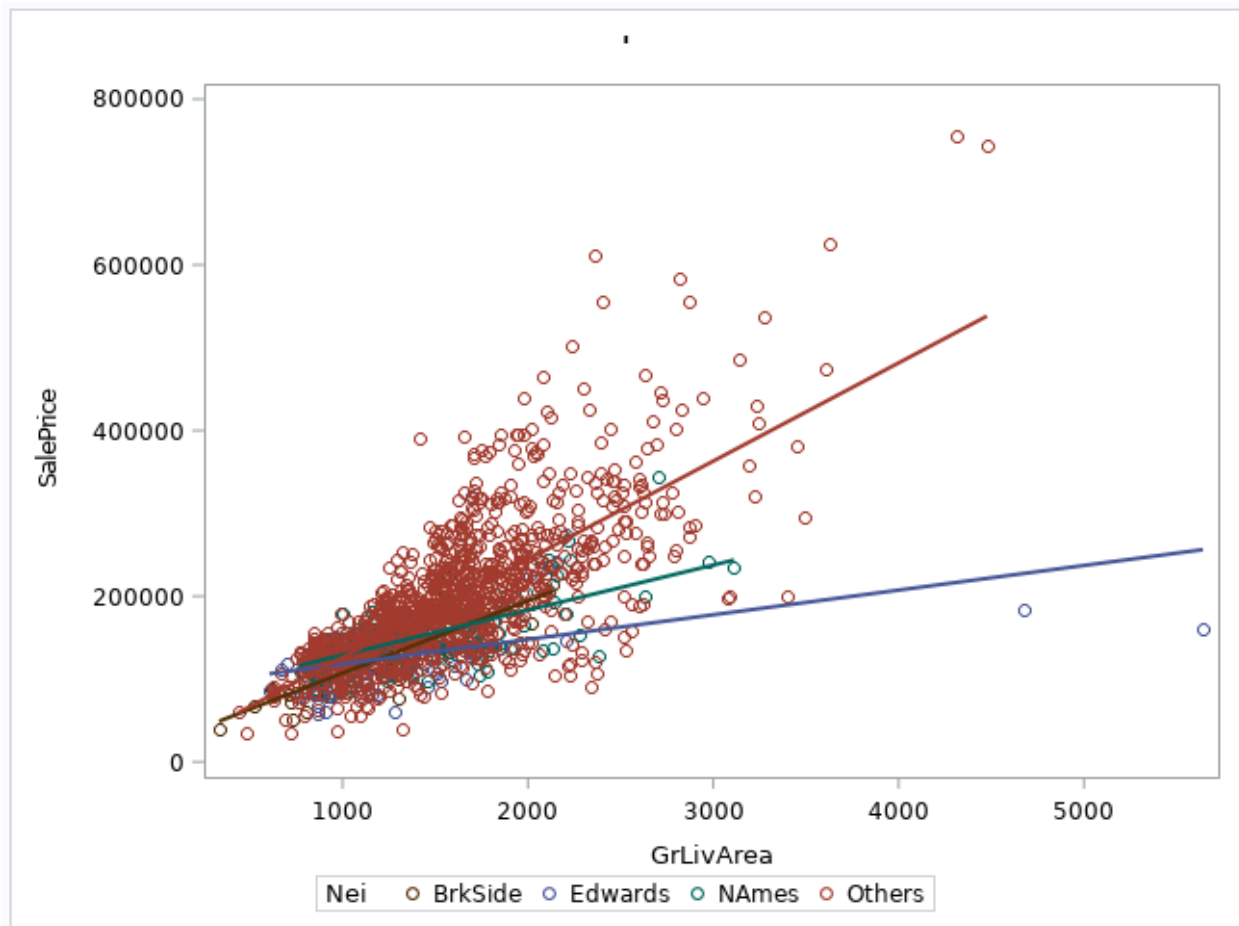


Figure 2: Neighborhoods

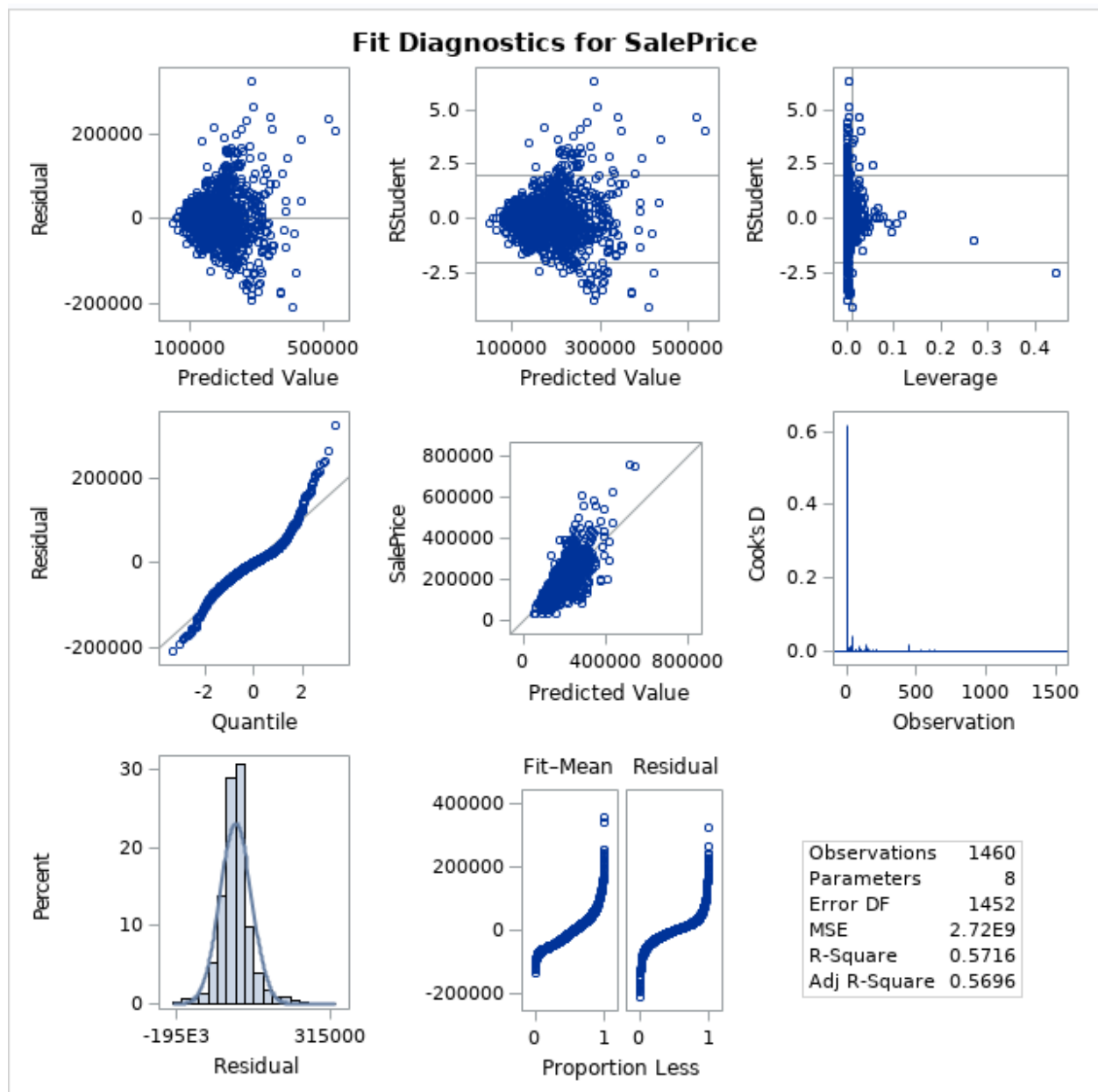


Figure 3: Fit Diagnostics



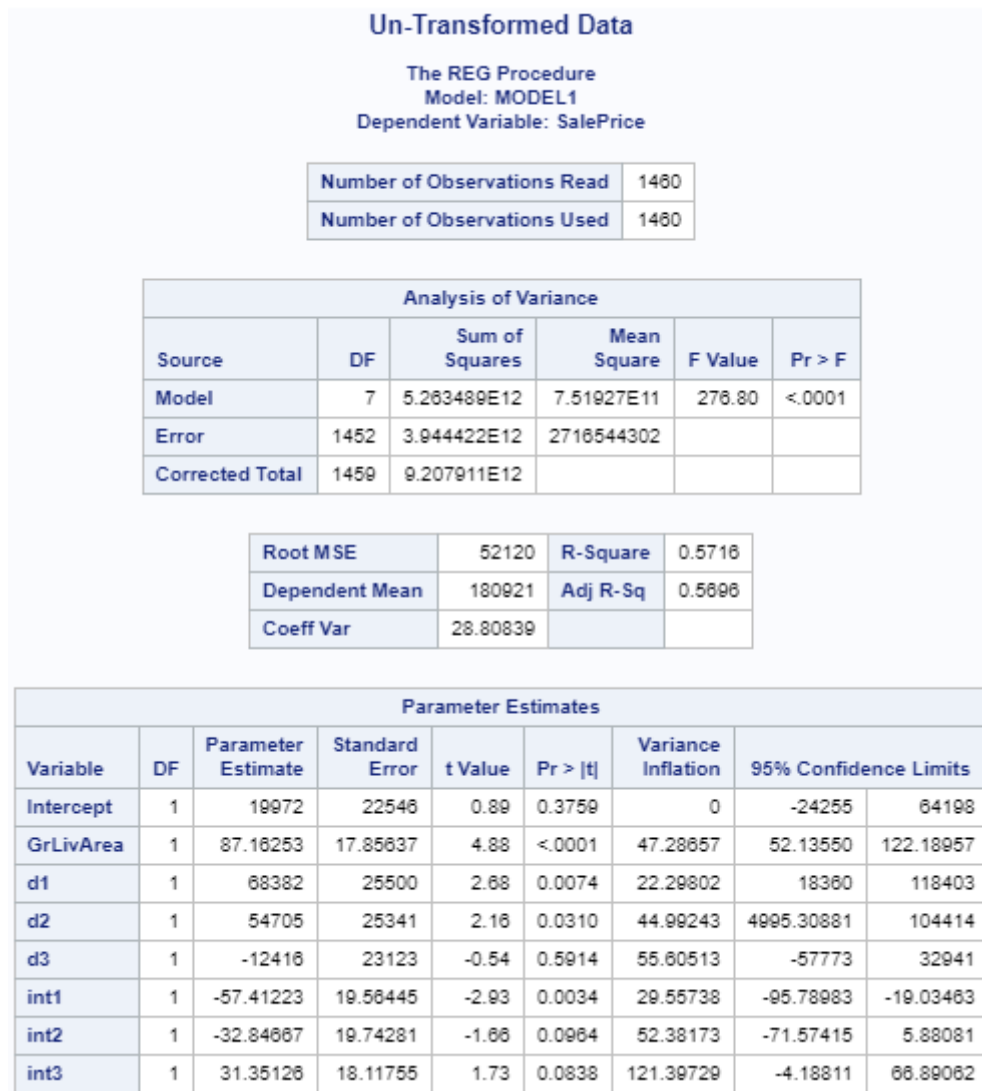


Figure 4: MLR Proc Reg

## Question 1 Plots of Untransformed Data

Image 001

Image 002

Image 003



Figure 5: Neighborhoods

## Question 1 Plots of Log-Log Data

Image 004

Image 005

Image 006

## Question 1 Miscellaneous Plots

Image 007

Image 008

---

## Question 2

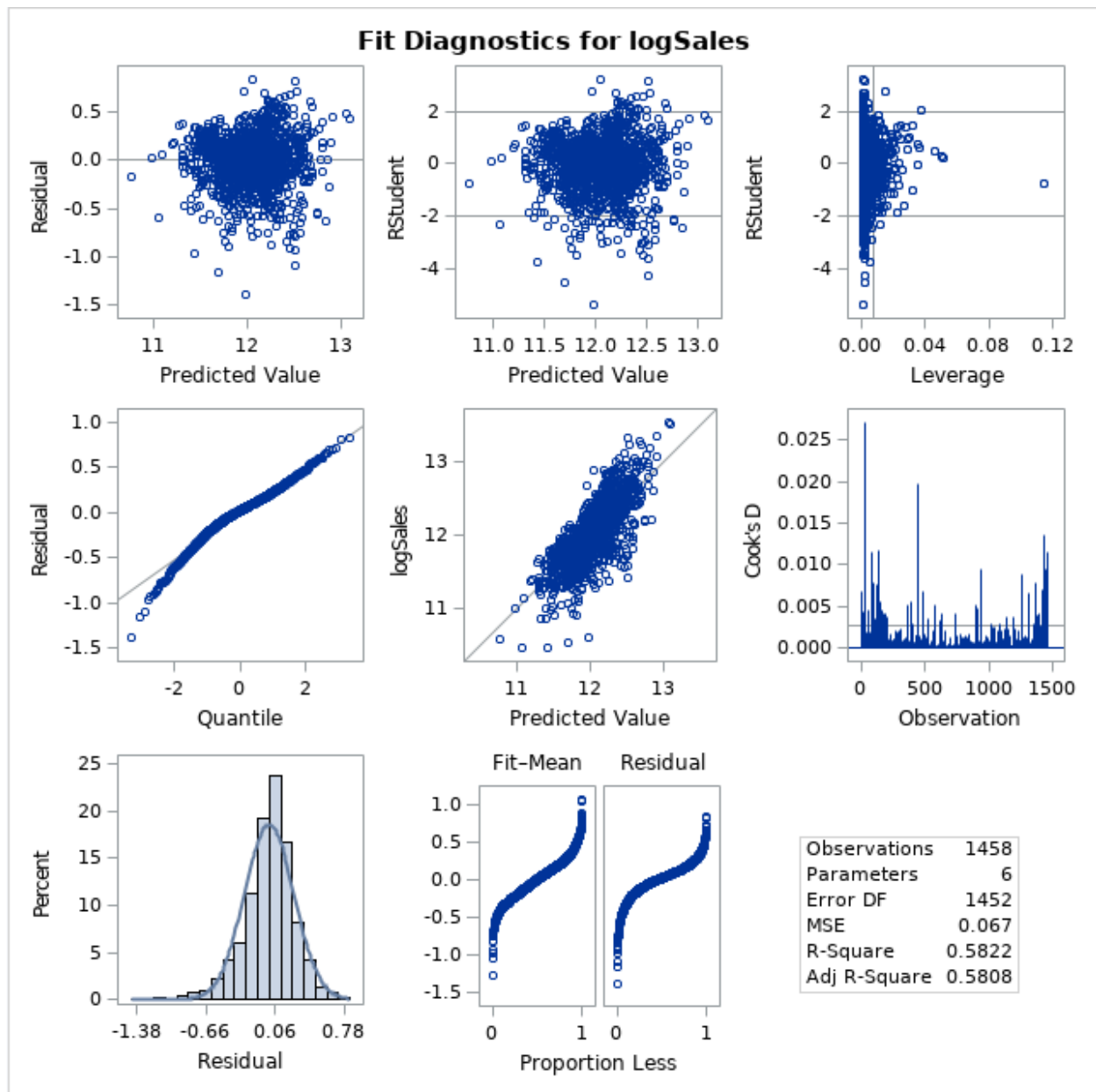


Figure 6: Fit Diagnostics

### Log-Log Transformed Data

The REG Procedure  
Model: MODEL1  
Dependent Variable: logSales

Number of Observations Read	1458
Number of Observations Used	1458

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	135.52713	27.10543	404.65	<.0001
Error	1452	97.26129	0.06698		
Corrected Total	1457	232.78842			

Root MSE	0.25881	R-Square	0.5822
Dependent Mean	12.02401	Adj R-Sq	0.5808
Coeff Var	2.15247		

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation	95% Confidence Limits	
Intercept	1	5.84426	0.15384	37.99	<.0001	0	5.54249	6.14604
logLiv	1	0.82456	0.02150	38.36	<.0001	1.09724	0.78239	0.86672
d2	1	0.09422	0.03083	3.06	0.0023	2.70082	0.03374	0.15471
d3	1	0.21898	0.02556	8.57	<.0001	2.74544	0.16884	0.26913
centLint2	1	-0.26810	0.09080	-2.95	0.0032	2.24360	-0.44622	-0.08998
centLint3	1	0.16896	0.07138	2.37	0.0181	2.17647	0.02894	0.30898

Figure 7: MLR Proc Reg

Sum of Residuals	0
Sum of Squared Residuals	3.944422E12
Predicted Residual SS (PRESS)	4.00198E12

Figure 8: Untransformed R<sup>2</sup> and PRESS

Sum of Residuals	4.48477E-12
Sum of Squared Residuals	97.26129
Predicted Residual SS (PRESS)	97.93883

Figure 9: Log-Log Transformed R<sup>2</sup> and PRESS