

Table of Contents

Appendix – SAS and R Code for Project	2
Analysis 1	2
Analysis 2	3
Appendix 1: Vanilla Regression – Raw Data Model Original	7
Appendix 2: Vanilla Regression Model: Log-Linear	8
Appendix 3: Vanilla Regression Model: Linear - Log	9
Appendix 4: Vanilla Regression Model: Log-Log	10
Appendix 5: Vanilla Regression Model: Outliers Addressed*	11
Appendix 6: Vanilla Regression Model: Log-Linear (Outliers Addressed)	12
Appendix 7: Vanilla Regression Model: Linear-Log (Outliers Addressed)	13
Appendix 8: Vanilla Regression Model: Log-Log (Outliers Addressed)	14
Appendix 9: Analysis 2 Checking Assumptions	15
Appendix 10 : Forward Selection SAS Results	16
.....	16
Appendix 11 : Analysis 2 Backwards Elimination	17
.....	17
.....	17
.....	17
Appendix 12 : Analysis 2 Stepwise Selection	18
Appendix 13 : Analysis 2 Custom Selection	18
.....	19
.....	19
APPENDIX 14 : Kaggle Submission Score	20

Appendix – SAS and R Code for Project

Analysis 1

```
FILENAME REFFILE '/folders/myfolders/sasuser.v94/test.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=test;
    GETNAMES=YES;
RUN;

FILENAME REFFILE '/folders/myfolders/sasuser.v94/train.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=train;
    GETNAMES=YES;
RUN;

/*Filter only relevant Neighborhoods for Q1. NAmes, Edwards, BrkSide*/
data neighborhoodSS;
    set train;
    where Neighborhood = 'NAmes' or Neighborhood = 'Edwards' or Neighborhood='BrkSide';
run;

/* Create GrLivAreaper100 variable, log Transform the SalesPrice and GrLivAreaper100 to re-run model */
data logneighborhoodSS;
    set neighborhoodSS;
    GrLivAreaper100 = GrLivArea/100;
    logSalePrice = log(SalePrice);
    logGrLivAreaper100 = log(GrLivArea/100);
run;

/*proc print data = logneighborhoodSS;
run;*/

/*Scatterplot the Saleprice to GrLivArea for the three requested neighborhoods*/
proc sgplot data=logneighborhoodSS;
scatter x=GrLivAreaper100 y=SalePrice / group=Neighborhood;
run;

/* Model for Salesprice to GrLivArea*/
proc glm data = logneighborhoodSS plots=all;
class Neighborhood;
model SalePrice = GrLivAreaper100 | Neighborhood / cli clm clparm;
run;

/*Filter out all homes greater than 40 100 sq.ft.*/
data logneighborhoodSSsmall;
    set logneighborhoodSS;
    where GrLivAreaper100 <= 40;
run;
```

```
/* Model for Salesprice to GrLivArea*/
proc glm data = logneighborhoodSSsmall plots=all;
class Neighborhood;
model SalePrice = GrLivAreaper100 | Neighborhood / cli clm clparm;
run;

/*Filter out all homes that costs more than $300,000*/
data logneighborhoodSScheap;
    set logneighborhoodSS;
    where SalePrice <= 300000;
run;

/* Model for Salesprice to GrLivArea*/
proc glm data = logneighborhoodSScheap plots=all;
class Neighborhood;
model SalePrice = GrLivAreaper100 | Neighborhood / cli clm clparm;
run;

/*Filter out all homes that costs more than $300,000 and smaller than 40 100 sq. ft.*/
data logneighborhoodSScheapsmall;
    set logneighborhoodSS;
    where SalePrice <= 300000 and GrLivAreaper100 <= 40;
run;

/* Model for Salesprice to GrLivArea*/
proc glm data = logneighborhoodSScheapsmall plots=all;
class Neighborhood;
model SalePrice = GrLivAreaper100 | Neighborhood / cli clm clparm solution;
run;

/* Model with log transformed SalePrice*/
proc glm data = logneighborhoodSS plots=all;
class Neighborhood;
model logSalePrice = GrLivAreaper100 | Neighborhood / cli clm clparm;
run;

/* Model with log transformed GrLivAreaper100*/
proc glm data = logneighborhoodSS plots=all;
class Neighborhood;
model SalePrice = logGrLivAreaper100 | Neighborhood / cli clm clparm;
run;

/* Model with log-log transformed SalePrice-GrLivAreaper100*/
proc glm data = logneighborhoodSS plots=all;
class Neighborhood;
model logSalePrice = logGrLivAreaper100 | Neighborhood / cli clm clparm;
run;

/* Model with log-log transformed SalePrice-GrLivAreaper100 on homes that are less than $300,000 and 40 100 sq. ft.*/
proc glm data = logneighborhoodSScheapsmall plots=all;
class Neighborhood;
model logSalePrice = logGrLivAreaper100 | Neighborhood / cli clm clparm; run;
```

Analysis 2

#R code for EDA

Appendix | Andy Ho; An Nguyen; Matthew Norton

#masterAN.csv is combination of train.csv and test.csv

```
trainDF <- read.csv("masterAN.csv", header = T, stringsAsFactors = F)
trainDF <- subset(trainDF[1:1470,])
trainDF$SalePrice <- as.integer(trainDF$SalePrice)
titles <- names(trainDF)

for (i in 1:length(trainDF)){
  if (is.character(trainDF[[i]])){
    boxplot(trainDF$SalePrice ~ trainDF[[i]], main = paste("Sale price vs ", titles[i], sep=""), ylab = titles[i], xlab = "Sale price", horizontal = T,
las = 1)
  }
  else if (is.integer(trainDF[[i]])){
    plot(x = trainDF[[i]], y = trainDF$SalePrice, main = paste("Sale price vs ", titles[i], sep=""), xlab = titles[i], ylab = "Sale price")
  }
}
```

/*SAS code for model selection (Forward, Backward, Stepwise, and custom user defined);*/

```
data train;
infile '/folders/myfolders/sasuser.v94/masterAN.csv' dlm=',' firstobs=2;
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 $ GarageYrBlt GarageFinish $ GarageCars GarageArea GarageQual $ GarageCond $ PavedDrive $ WoodDeckSF
      OpenPorchSF EnclosedPorch _3SsnPorch ScreenPorch PoolArea PoolQC $ Fence $ MiscFeature $ MiscVal MoSold YrSold
      SaleType $ SaleCondition $ SalePrice;
```

/*Forward selection*/

```
proc glmselect data=train seed=1;
  class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope
    Neighborhood 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 Fence MiscFeature SaleType SaleCondition PoolQC;
  model SalePrice=MSSubClass--SaleCondition / selection=forward (choose=CV
    stop=CV) cvmethod=split(10) CVdetails;
  output out=forward_results p=Predict;
run;
```

/* Subset only Test data with salesprice prediction*/

```
data forward_results2;
  set forward_results;

  if Predict > 0 then
    SalePrice=Predict;

  if Predict < 0 then
    SalePrice=10000;
  keep id SalePrice;
  where id > 1460;
run;
```

/* Export Kaggle Submission csv file*/

```
proc export data=forward_results2
```

```

        outfile='/folders/myfolders/sasuser.v94/submit_forward2.csv' dbms=csv replace;

run;

proc print data=forward_results2;
run;

/*Backward elimination*/
proc glmselect data=train seed=1;
    class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope
        Neighborhood 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 Fence MiscFeature SaleType SaleCondition PoolQC;
    model SalePrice=MSSubClass--SaleCondition / selection=backward(choose=CV
        stop=CV) cvmethod=split(10) CVdetails;
    output out=backward_results p=Predict;
run;

/* Subset only Test data with salesprice prediction*/
data backward_results2;
    set backward_results;

    if Predict > 0 then
        SalePrice=Predict;

    if Predict < 0 then
        SalePrice=10000;
    keep id SalePrice;
    where id > 1460;
run;

/* Export Kaggle Submission csv file*/
proc export data=backward_results2
    outfile='/folders/myfolders/sasuser.v94/submit_backwards.csv' dbms=csv replace;
run;

proc print data=backward_results2;
run;

/*Stepwise selection*/
proc glmselect data=train seed=1;
    class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope
        Neighborhood 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 Fence MiscFeature SaleType SaleCondition PoolQC;
    model SalePrice=MSSubClass--SaleCondition / selection=stepwise(choose=CV
        stop=CV) cvmethod=split(10) CVdetails;
    output out=stepwise_results p=Predict;
run;

/* Subset only Test data with salesprice prediction*/
data stepwise_results2;
    set stepwise_results;

    if Predict > 0 then
        SalePrice=Predict;

```

```

        if Predict < 0 then
            SalePrice=10000;
        keep id SalePrice;
        where id > 1460;
run;

/* Export Kaggle Submission csv file*/
proc export data=stepwise_results2
    outfile='/folders/myfolders/sasuser.v94/submit_stepwise.csv' dbms=csv replace;
run;

proc print data=stepwise_results2;
run;

/*Custom model*/
proc glmselect data=train2 plots=all;
    class KitchenQual GarageFinish BsmtQual ExterQual MasVnrType Neighborhood;
    model SalePrice=LotFrontage LotArea OverallQual OverallCond YearBuilt
        BsmtFinSF1 TotalBsmtSF _1stFlrSF _2ndFlrSF GrLivArea FullBath TotRmsAbvGrd
        GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF KitchenQual
        GarageFinish BsmtQual ExterQual MasVnrType Neighborhood / selection=none;
    output out=results p=Predict; run;

/* Subset only Test data with salesprice prediction*/
data results3;
    set results;

    if Predict > 0 then
        SalePrice=Predict;

    if Predict < 0 then
        SalePrice=10000;
    keep id SalePrice;
    where id > 1460;
run;

/* Export Kaggle Submission csv file*/
proc export data=results3 outfile='/folders/myfolders/sasuser.v94/submit_custom2.csv'
    dbms=csv replace;run;

/*Custom model, use for obtaining Cook's D, Leverage, and Studentized residuals;*/
proc glm data=train plots=all;
    class KitchenQual GarageFinish BsmtQual ExterQual MasVnrType Neighborhood;
    model SalePrice=LotFrontage LotArea OverallQual OverallCond YearBuilt
        BsmtFinSF1 TotalBsmtSF _1stFlrSF _2ndFlrSF GrLivArea FullBath TotRmsAbvGrd
        GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF KitchenQual
        GarageFinish BsmtQual ExterQual MasVnrType Neighborhood / solution;
    output out=results student=res cookd = cookd h = lev;run;

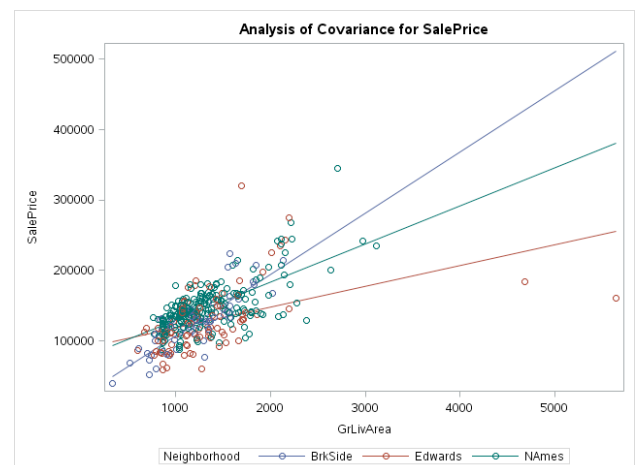
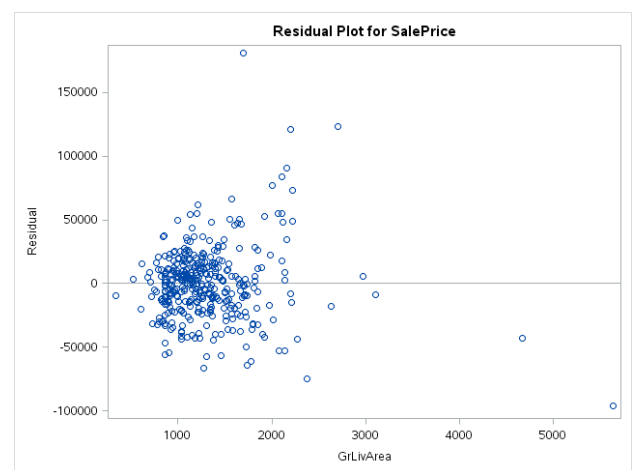
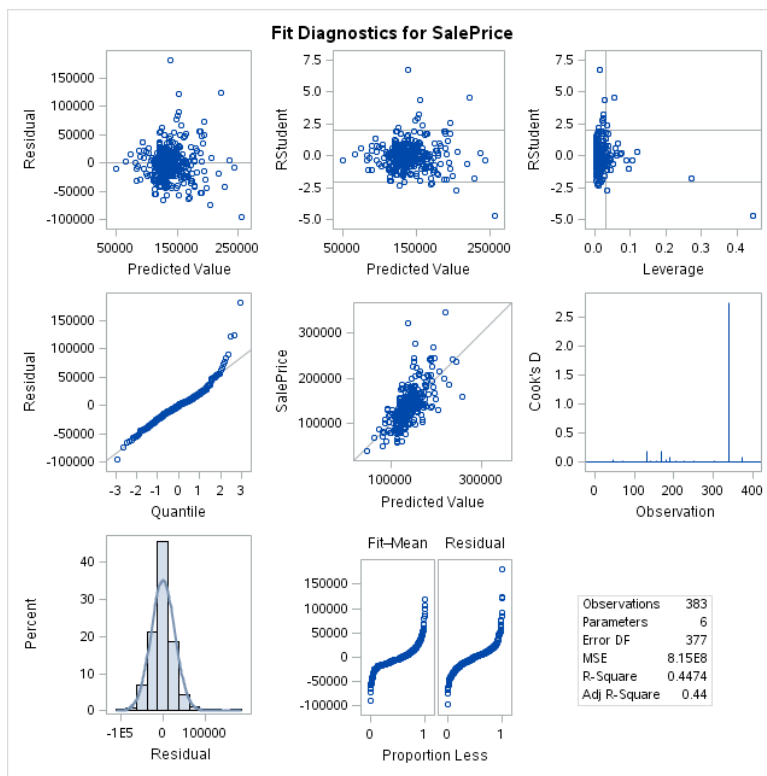
/*Display individuals with high Cook's D, large absolute studentized residuals or high leverage;*/
data results2;
    set results;
    where id <= 1460 and (cookd > 1 or res > 10 or res < -10 or lev > 1);
run;
proc print data = results2;run;

```

Appendix 1: Vanilla Regression – Raw Data | Model Original

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	74676.40154	B	6337.89399	11.78	<.0001
GrLivArea	54.31586	B	4.61364	11.77	<.0001
Neighborhood BrkSide	-54704.88774	B	13882.33364	-3.94	<.0001
Neighborhood Edwards	13676.70324	B	9097.57465	1.50	0.1336
Neighborhood NAmes	0.00000	B	.	.	.
GrLivArea*Neighborhood BrkSide	32.84667	B	10.81538	3.04	0.0026
GrLivArea*Neighborhood Edwards	-24.56556	B	6.36139	-3.86	0.0001
GrLivArea*Neighborhood NAmes	0.00000	B	.	.	.

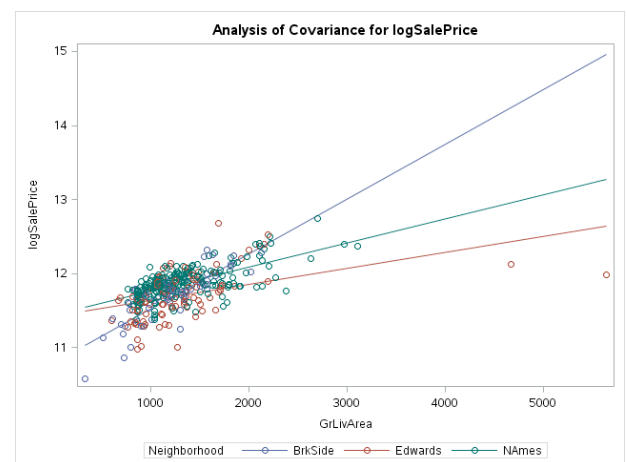
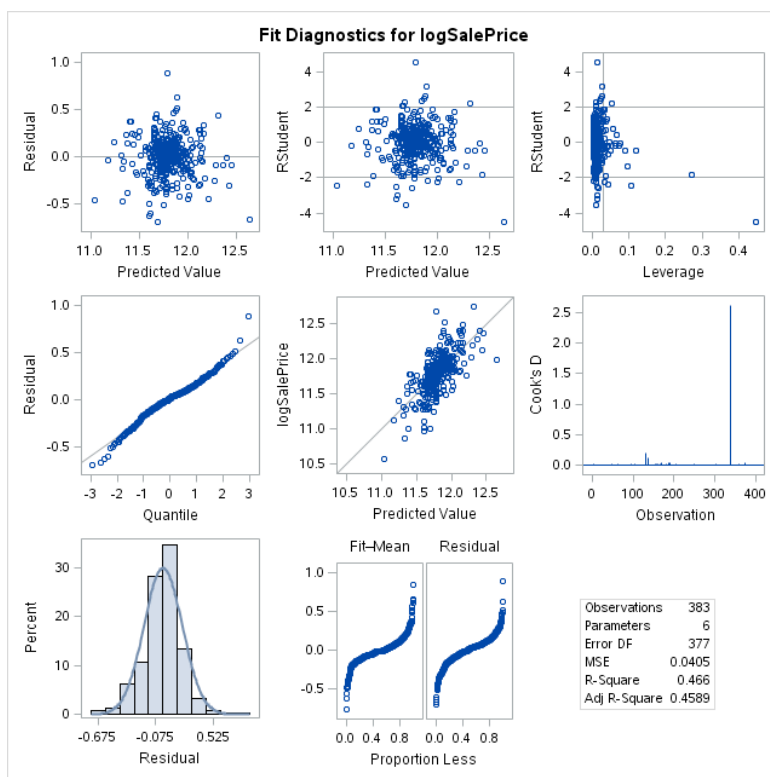
R-Square	Coeff Var	Root MSE	SalePrice Mean
0.447376	20.68070	28552.30	138062.5



Appendix 2: Vanilla Regression | Model: Log-Linear

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	11.44334070	B	0.04465161	256.28	<.0001
GrLivArea	0.00032412	B	0.00003250	9.97	<.0001
Neighborhood BrkSide	-0.65174673	B	0.09780355	-6.66	<.0001
Neighborhood Edwards	-0.02139976	B	0.06409406	-0.33	0.7387
Neighborhood NAmes	0.00000000	B	.	.	.
GrLivArea*Neighborhood BrkSide	0.00041410	B	0.00007620	5.43	<.0001
GrLivArea*Neighborhood Edwards	-0.00010744	B	0.00004482	-2.40	0.0170
GrLivArea*Neighborhood NAmes	0.00000000	B	.	.	.

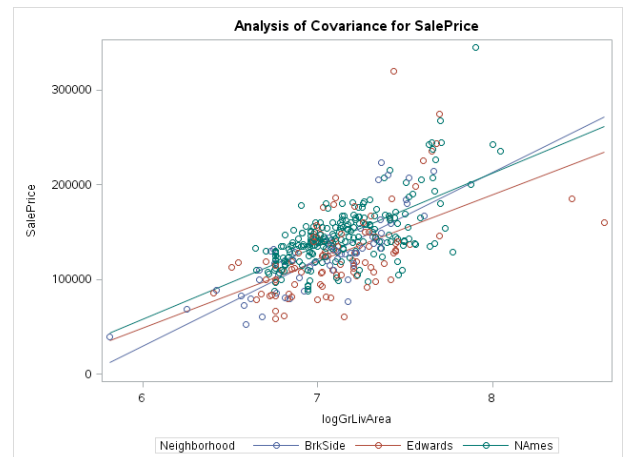
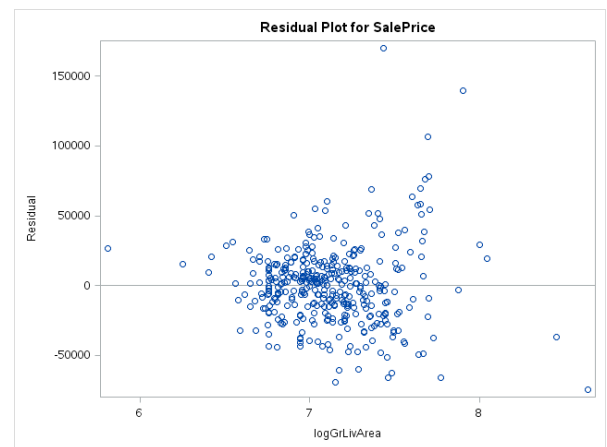
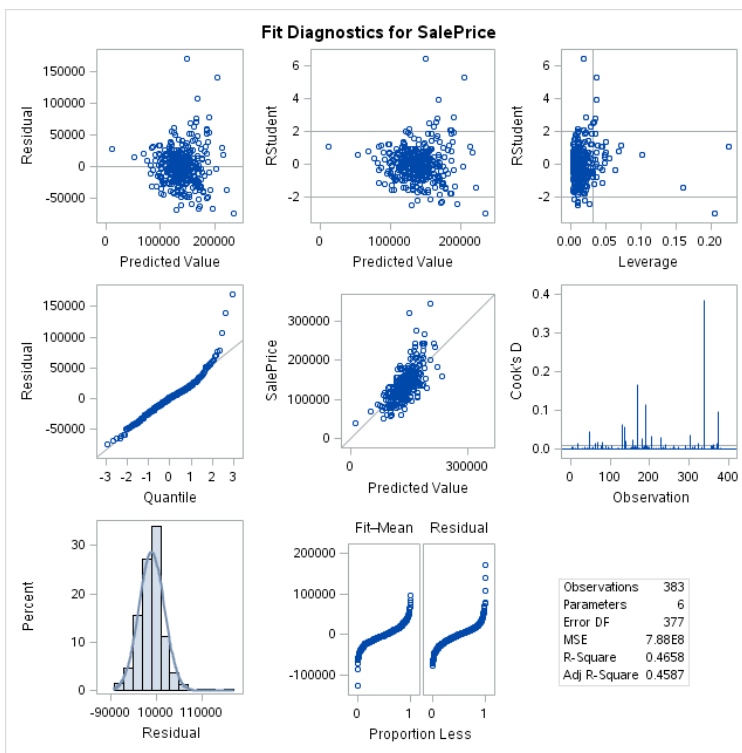
R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.465985	1.704877	0.201156	11.79887



Appendix 3: Vanilla Regression | Model: Linear - Log

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	-405474.3769	B	47364.43769	-8.56	<.0001
logGrLivArea	77263.2789	B	6632.56245	11.65	<.0001
Neighborhood BrkSide	-115363.9795	B	87581.82712	-1.32	0.1886
Neighborhood Edwards	29405.8761	B	75555.48597	0.39	0.6974
Neighborhood NAMES	0.0000	B	.	.	.
logGrLivA*Neighborhood BrkSide	14507.4723	B	12383.61185	1.17	0.2421
logGrLivA*Neighborhood Edwards	-6546.6413	B	10581.99330	-0.62	0.5365
logGrLivA*Neighborhood NAMES	0.0000	B	.	.	.

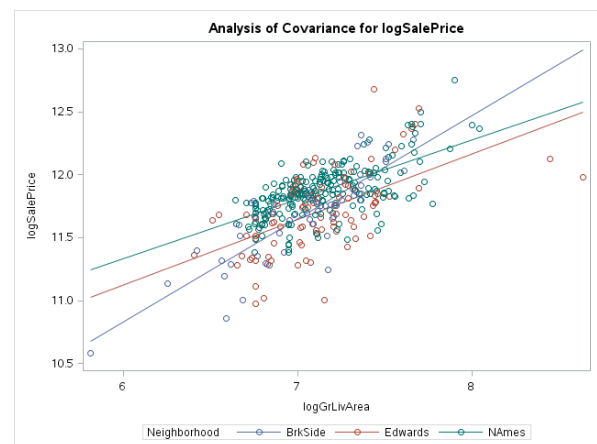
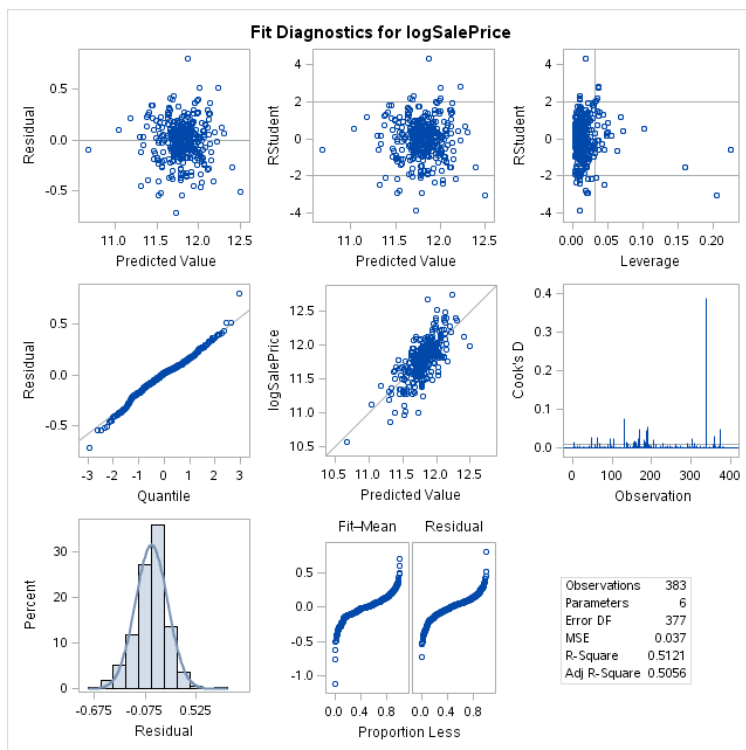
R-Square	Coeff Var	Root MSE	SalePrice Mean
0.465809	20.33287	28072.06	138062.5



Appendix 4: Vanilla Regression | Model: Log-Log

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	8.492727641	B	0.32441709	26.18	<.0001
logGrLivArea	0.473023602	B	0.04542895	10.41	<.0001
Neighborhood BrkSide	-2.579806905	B	0.59988132	-4.30	<.0001
Neighborhood Edwards	-0.486220461	B	0.51750833	-0.94	0.3481
Neighborhood NAmes	0.000000000	B	-	-	-
logGrLivA*Neighborhood BrkSide	0.346624454	B	0.08482008	4.09	<.0001
logGrLivA*Neighborhood Edwards	0.046643642	B	0.07248011	0.64	0.5203
logGrLivA*Neighborhood NAmes	0.000000000	B	-	-	-

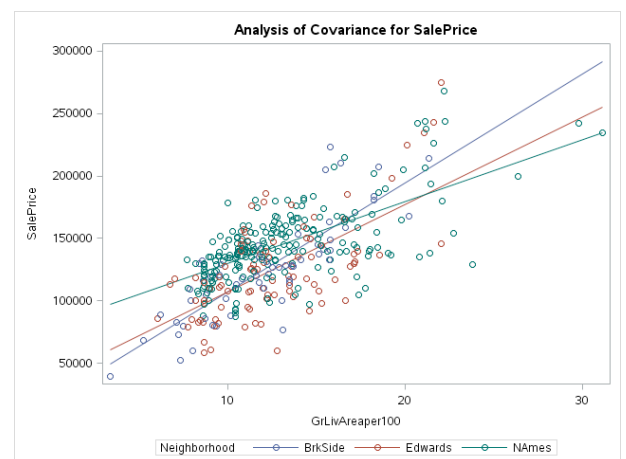
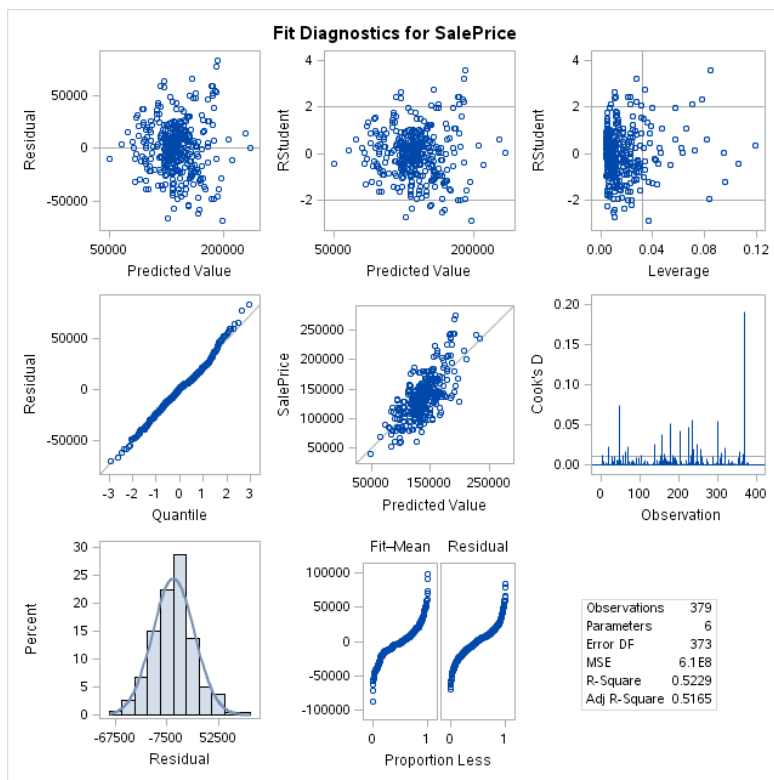
R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.512092	1.629617	0.192276	11.79887



Appendix 5: Vanilla Regression | Model: Outliers Addressed*

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	80325.71230	B	5592.03832	14.36	<.0001
GrLivAreaaper100	4956.12477	B	409.70671	12.10	<.0001
Neighborhood BrkSide	-60354.19850	B	12060.03479	-5.00	<.0001
Neighborhood Edwards	-43225.29073	B	10837.81644	-3.99	<.0001
Neighborhood NAMES	0.00000	B	.	.	.
GrLivArea*Neighborhood BrkSide	3760.12849	B	940.21789	4.00	<.0001
GrLivArea*Neighborhood Edwards	2059.71212	B	820.38610	2.51	0.0125
GrLivArea*Neighborhood NAMES	0.00000	B	.	.	.

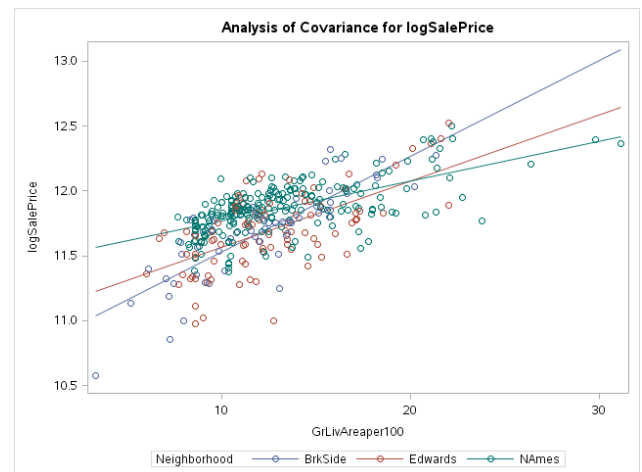
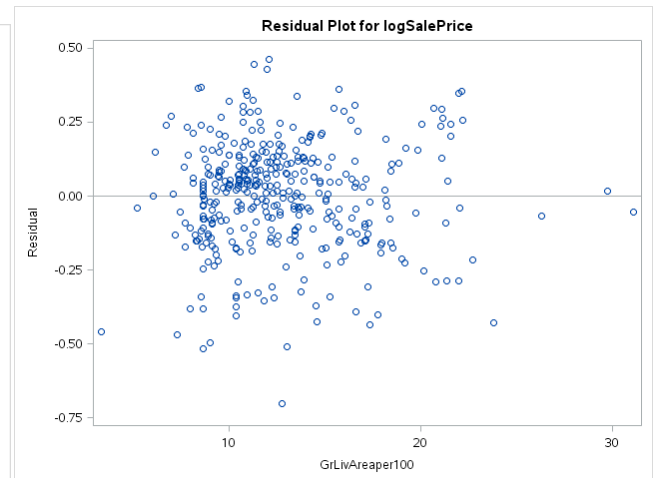
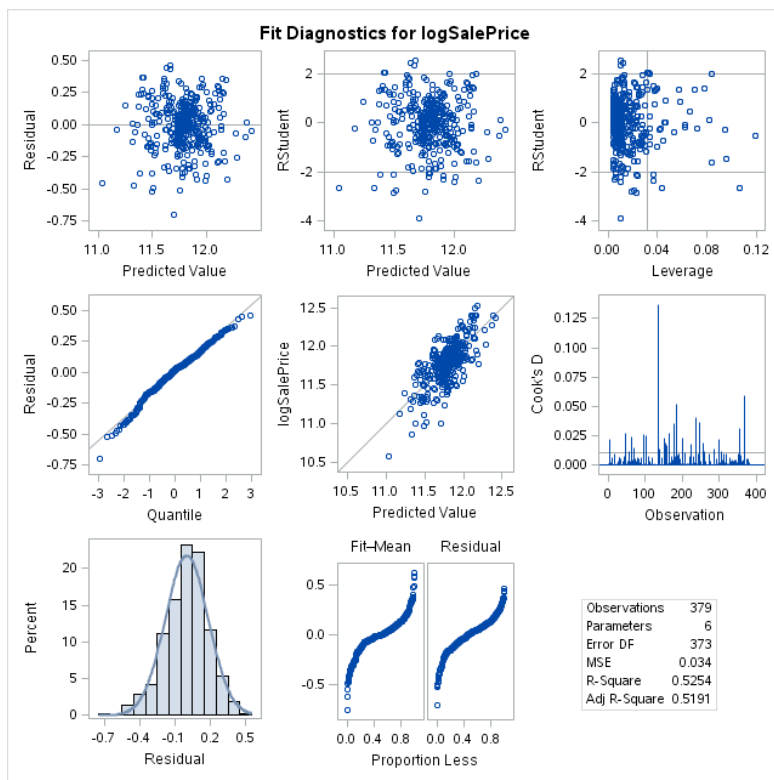
R-Square	Coeff Var	Root MSE	SalePrice Mean
0.522897	18.04909	24701.16	136855.4



Appendix 6: Vanilla Regression | Model: Log-Linear (Outliers Addressed)

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	11.46308762	B	0.04175292	274.55	<.0001
GrLivAreaper100	0.03075050	B	0.00305907	10.05	<.0001
Neighborhood BrkSide	-0.67149365	B	0.09004618	-7.46	<.0001
Neighborhood Edwards	-0.41144266	B	0.08092049	-5.08	<.0001
Neighborhood NAMES	0.00000000	B			
GrLivArea*Neighborhood BrkSide	0.04307178	B	0.00702013	6.14	<.0001
GrLivArea*Neighborhood Edwards	0.02043149	B	0.00612541	3.34	0.0009
GrLivArea*Neighborhood NAMES	0.00000000	B			

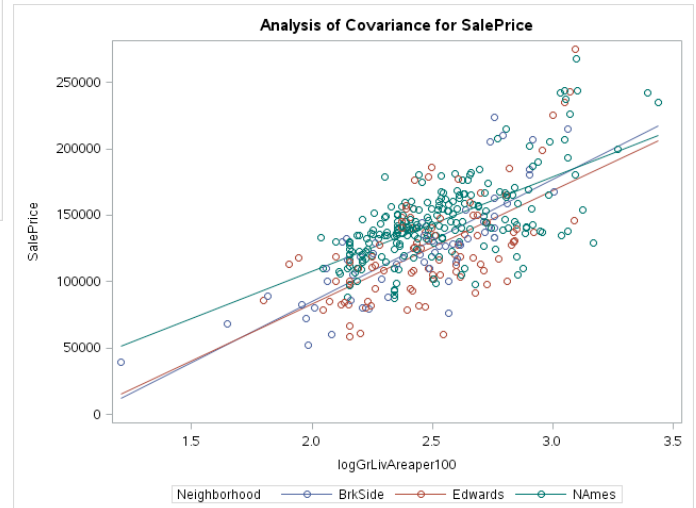
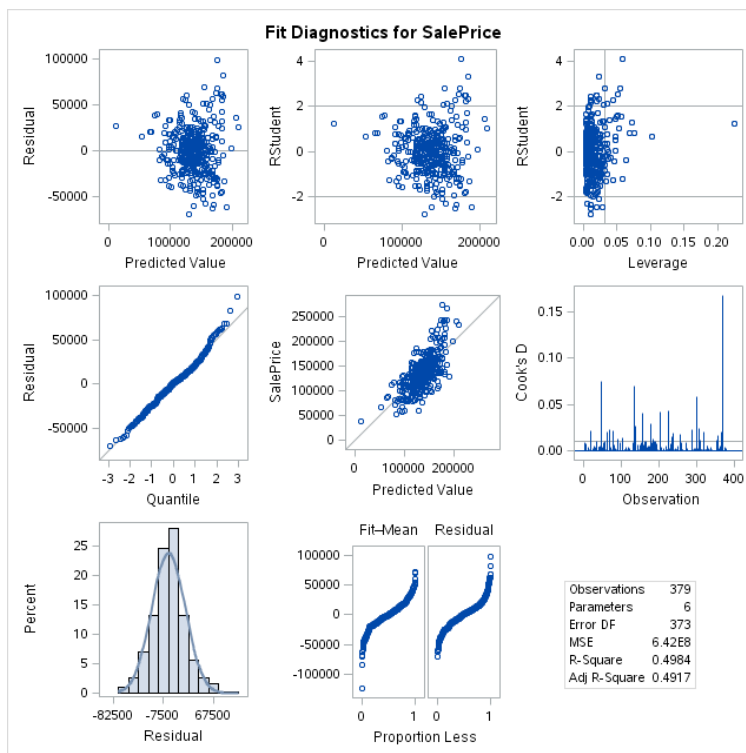
R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.525413	1.563944	0.184431	11.79269



Appendix 7: Vanilla Regression | Model: Linear-Log (Outliers Addressed)

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	-34567.92221	B	15470.76545	-2.23	0.0260
logGrLivAreaper100	71042.34430	B	6085.38391	11.67	<.0001
Neighborhood BrkSide	-63650.50684	B	27862.76048	-2.28	0.0229
Neighborhood Edwards	-52636.52333	B	28037.21033	-1.88	0.0612
Neighborhood NAmes	0.00000	B	.	.	.
logGrLivA*Neighborho BrkSide	20728.40689	B	11227.58949	1.85	0.0657
logGrLivA*Neighborho Edwards	14188.73230	B	11130.71521	1.27	0.2032
logGrLivA*Neighborho NAmes	0.00000	B	.	.	.

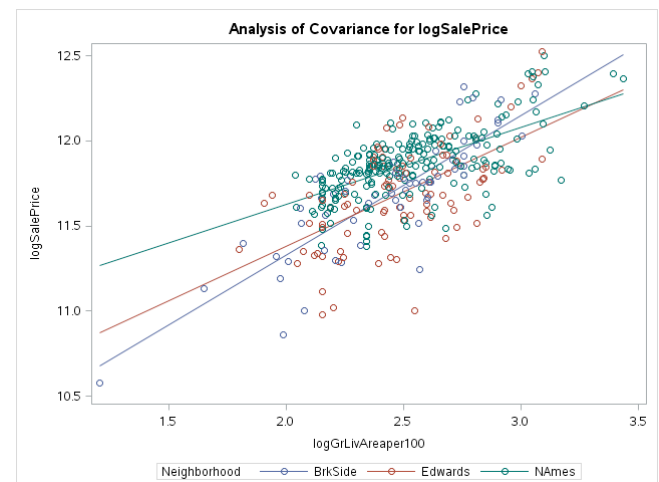
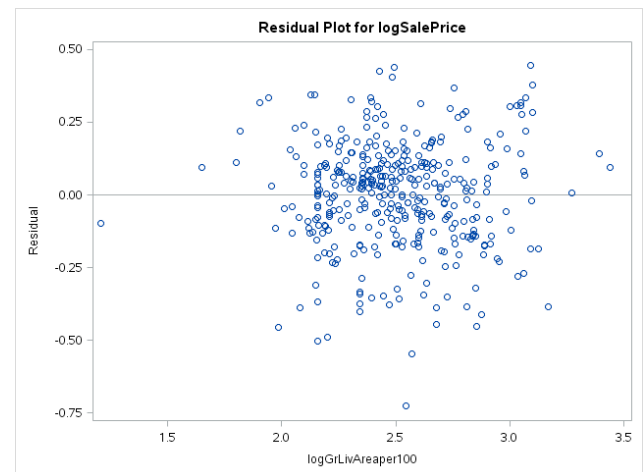
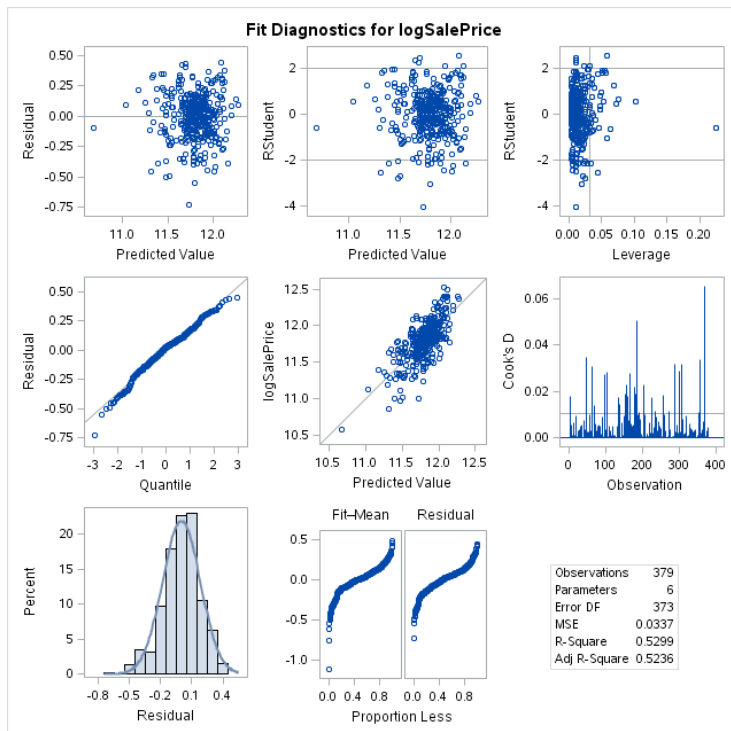
R-Square	Coeff Var	Root MSE	SalePrice Mean
0.498378	18.50708	25327.93	136855.4



Appendix 8: Vanilla Regression | Model: Log-Log (Outliers Addressed)

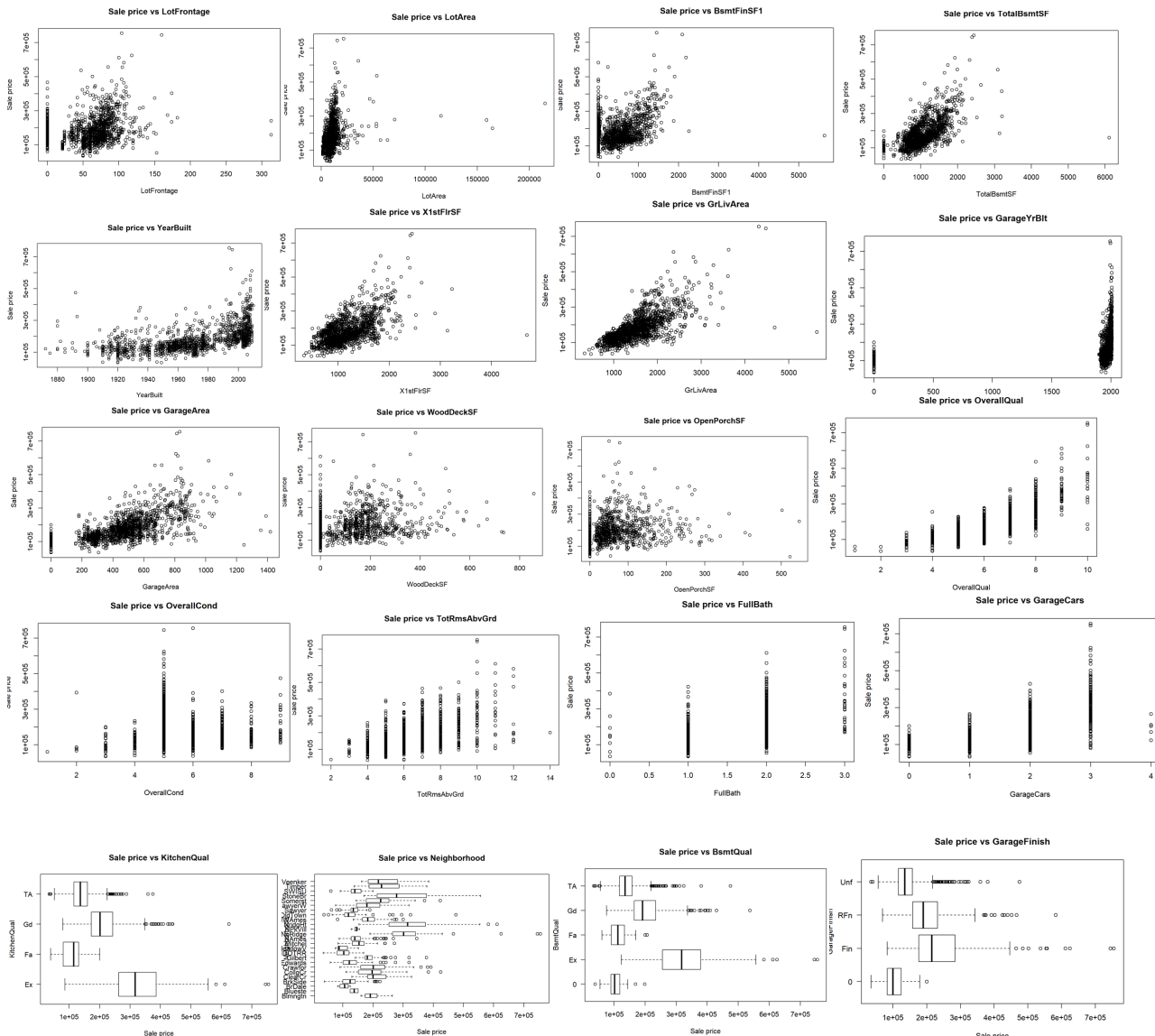
Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	10.72724642	B	0.11211841	95.68	<.0001
logGrLivAreaper100	0.44987850	B	0.04410147	10.20	<.0001
Neighborhood BrkSide	-1.03970690	B	0.20192461	-5.15	<.0001
Neighborhood Edwards	-0.62585626	B	0.20318887	-3.08	0.0022
Neighborhood NAmes	0.00000000	B	.	.	.
logGrLivA*Neighborhood BrkSide	0.36976955	B	0.08136763	4.54	<.0001
logGrLivA*Neighborhood Edwards	0.18931410	B	0.08066557	2.35	0.0195
logGrLivA*Neighborhood NAmes	0.00000000	B	.	.	.

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.529914	1.556511	0.183554	11.79269



Appendix 9: Analysis 2 | Checking Assumptions (Pairwise Variables with Linearity)

We disregarded the values of '0' as it was originally 'NA'. For a few of the plots, if '0' was not included in judging the plots, there would be sufficient evidence of correlation between the pairwise variables.



Appendix 10 : Forward Selection SAS Results

The GLMSELECT Procedure

Data Set	WORK.TRAIN4
Dependent Variable	SalePrice
Selection Method	Forward
Select Criterion	SBC
Stop Criterion	Cross Validation
Choose Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	10
Effect Hierarchy Enforced	None

Number of Observations Read	2919
Number of Observations Used	1453

The GLMSELECT Procedure
Selected Model

The selected model, based on Cross Validation, is the model at Step 7.

Effects: Intercept MSSubClass Neighborhood OverallQual ExterQual BsmtExposure BsmtFinSF1 GrLivArea

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	35	7.706526E12	2.201865E11	267.65
Error	1417	1.165735E12	822677995	
Corrected Total	1452	8.872261E12		

Root MSE	28682
Dependent Mean	180405
R-Square	0.8686
Adj R-Sq	0.8654
AIC	31318
AICC	31320
SBC	30053
CV PRESS	1.284085E12

Cross Validation Details

Index	Observations		CV PRESS
	Fitted	Left Out	
1	1307	146	9.66431E10
2	1307	146	9.4771E10
3	1307	146	2.47223E11
4	1308	145	1.12533E11
5	1308	145	1.37267E11
6	1308	145	1.8053E11
7	1308	145	1.35943E11
8	1308	145	9.83451E10
9	1308	145	1.02265E11
10	1308	145	7.85748E10
Total			1.28409E12

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	23083	10518	2.19
MSSubClass	1	-275.366217	21.280080	-12.94
Neighborhood Blmngtn	1	-7602.992318	11296	-0.67
Neighborhood Blueste	1	-13329	22296	-0.60
Neighborhood BrDale	1	-28498	11717	-2.43
Neighborhood BrkSide	1	-34635	9732.085281	-3.56
Neighborhood ClearCr	1	-22912	10333	-2.22
Neighborhood CollgCr	1	-20846	9085.523663	-2.29
Neighborhood Crawfor	1	-11832	9735.585870	-1.22
Neighborhood Edwards	1	-44564	9384.758640	-4.75
Neighborhood Gilbert	1	-20714	9389.515096	-2.21
Neighborhood IDOTRR	1	-49267	10218	-4.82
Neighborhood MeadowV	1	-21426	11605	-1.85
Neighborhood Mitchel	1	-31967	9776.667107	-3.27
Neighborhood NAmes	1	-37252	9109.777197	-4.09
Neighborhood NPKvIII	1	-10921	13243	-0.82
Neighborhood NWAmes	1	-30911	9542.657568	-3.24
Neighborhood NoRidge	1	18607	9985.134713	1.86
Neighborhood NridgHt	1	24169	9538.973472	2.53
Neighborhood OldTown	1	-47004	9380.734580	-5.01
Neighborhood SWISU	1	-52040	10704	-4.86
Neighborhood Sawyer	1	-33888	9558.483583	-3.55
Neighborhood SawyerW	1	-27646	9529.227186	-2.90
Neighborhood Somerst	1	1083.589725	9375.122303	0.12
Neighborhood StoneBr	1	33689	10549	3.19
Neighborhood Timber	1	-17183	9873.938952	-1.74
Neighborhood Veenker	0	0	.	.
OverallQual	1	13747	1010.396593	13.61
ExterQual Ex	1	61646	5435.303177	11.34
ExterQual Fa	1	-4045.246585	8079.321547	-0.50
ExterQual Gd	1	9922.462958	2536.197499	3.91
ExterQual TA	0	0	.	.
BsmtExposure 0	1	-789.054205	4985.392953	-0.16
BsmtExposure Av	1	11100	2363.318582	4.70
BsmtExposure Gd	1	33250	3062.316378	10.86
BsmtExposure Mn	1	8970.798814	2889.601355	3.10
BsmtExposure No	0	0	.	.
BsmtFinSF1	1	24.467273	2.048623	11.94
GrLivArea	1	60.931671	2.057459	29.62

Appendix 11 : Analysis 2 | Backwards Elimination

The GLMSELECT Procedure

Data Set	WORK.TRAIN4
Dependent Variable	SalePrice
Selection Method	Backward
Select Criterion	SBC
Stop Criterion	Cross Validation
Choose Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	10
Effect Hierarchy Enforced	None

Number of Observations Read	2919
Number of Observations Used	1453

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	231	8.273593E12	35816421120	73.05
Error	1221	5.986678E11	490309403	
Corrected Total	1452	8.872261E12		

Root MSE	22143
Dependent Mean	180405
R-Square	0.9325
Adj R-Sq	0.9198
AIC	30742
AICC	30831
SBC	30512
CV PRESS	1.2082E12

Cross Validation Details				
		Observations		CV PRESS
Index	Fitted	Left Out		
1	1307	146	6.29004E10	
2	1307	146	7.0841E10	
3	1307	146	2.43882E11	
4	1308	145	9.98807E10	
5	1308	145	1.03041E11	
6	1308	145	1.38758E11	
7	1308	145	9.06695E10	
8	1308	145	7.13204E10	
9	1308	145	8.98909E10	
10	1308	145	2.40959E11	
Total			1.2082E12	

Parameter Estimates					Neighborhood Sawyer					RoofMat WdShngl					BmtFinType1 ALQ				
Parameter	DF	Estimate	Standard Error	t Value															
Intercept	1	455801	1019977	0.45															
MSSubClass	1	-47.601185	80.414938	-0.59															
MSZoning C (all)	1	-21515	9341.518072	-2.30															
MSZoning FV	1	11452	7418.640168	1.54															
MSZoning RH	1	585.002446	7373.814431	0.08															
MSZoning RL	1	3593.740453	3061.969781	0.08															
MSZoning RM	0	0	0	-															
LotFrontage	1	11.804637	22.370992	0.53															
Li Area	1	0.730481	0.104492	6.99															
St et Grvl	1	-22557	11902	-2.74															
St et Pave	0	0	0	-															
Alley 0	1	529.320919	4612.54680	0.11															
Alley Grvl	1	852.399977	5888.060953	0.14															
Alley Pave	0	0	0	-															
LotShape IR1	1	2005.359904	1588.572325	-1.26															
LotShape IR2	1	1984.673674	4221.582565	0.47															
LotShape IR3	1	3717.705375	8637.540302	0.43															
LotShape Reg	0	0	0	-															
LandContour Bnk	1	5012.242086	3589.675608	-1.40															
LandContour HLS	1	2831.838534	3972.579777	0.71															
LandContour Low	1	-15611	5518.449937	-2.83															
LandContour Lvl	0	0	0	-															
Utilities AllPub	1	59700	26792	1.54															
Utilities NoSeWa	0	0	0	-															
LotConfig Corner	1	1463.381554	1711.234849	0.86															
LotConfig CulDSa	1	10029	2936.802256	3.42															
LotConfig FR2	1	-5680.332208	3672.918171	-1.54															
LotConfig FR3	1	-14168	12199	-1.16															
LotConfig Inside	0	0	0	-															
LandSlope Gtl	1	43224	11116	3.89															
LandSlope Mod	1	50495	11048	4.57															
LandSlope Sev	0	0	0	-															
Neighborhood Blmngtn	1	578.779303	10214	0.06															
Neighborhood Blueste	1	7496.030490	19000	0.39															
Neighborhood BrDale	1	1303.307897	11245	-0.12															
Neighborhood BrkSide	1	5699.361008	9189.583066	-0.62															
Neighborhood ClearCr	1	-14707	9240.047869	-1.59															
Neighborhood CollgCr	1	9906.376941	8037.591543	-1.23															
Neighborhood Crawford	1	9583.684201	8662.312413	1.11															
Neighborhood Edwards	1	20526	9318.071740	2.47															
Neighborhood Gilbert	1	-10245	8402.251023	-1.22															
Neighborhood IDOTRR	1	15648	10378	1.52															
Neighborhood MeadowV	1	-8845.749553	11375	-0.60															
Neighborhood Mitchel	1	-20826	8499.582839	-2.45															
Neighborhood NAmes	1	-18562	7950.336877	-2.08															
Neighborhood NPkVill	1	13214	11373	1.01															
Neighborhood NWAmes	1	-18994	8182.603346	-2.08															
Neighborhood NoRidge	1	22709	8825.154154	2.58															
Neighborhood NridgHt	1	17620	8572.122604	2.06															
Neighborhood OldTown	1	-15622	9343.086187	-1.67															
Neighborhood SWISU	1	-10675	9871.034455	-1.08															

Appendix 12 : Analysis 2 | Stepwise Selection

The GLMSELECT Procedure

Data Set	WORK.TRAIN4
Dependent Variable	SalePrice
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Choose Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	10
Effect Hierarchy Enforced	None

Number of Observations Read	2919
Number of Observations Used	1453

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	35	7.706526E12	2.201865E11	267.65
Error	1417	1.165735E12	822677995	
Corrected Total	1452	8.872261E12		

Root MSE	28682
Dependent Mean	180405
R-Square	0.8686
Adj R-Sq	0.8654
AIC	31318
AICC	31320
SBC	30053
CV PRESS	1.284085E12

Cross Validation Details

Index	Observations		CV PRESS
	Fitted	Left Out	
1	1307	146	9.66431E10
2	1307	146	9.4771E10
3	1307	146	2.47223E11
4	1308	145	1.12533E11
5	1308	145	1.37267E11
6	1308	145	1.8053E11
7	1308	145	1.35943E11
8	1308	145	9.83451E10
9	1308	145	1.02255E11
10	1308	145	7.85748E10
Total			1.28409E12

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	23083	10518	2.19
MSSubClass	1	-275.366217	21.280080	-12.94
Neighborhood Blmngtn	1	-7602.992318	11296	-0.67
Neighborhood Blueste	1	-13329	22296	-0.60
Neighborhood BrDale	1	-28498	11717	-2.43
Neighborhood BrkSide	1	-34635	9732.085281	-3.56
Neighborhood ClearCr	1	-22912	10333	-2.22
Neighborhood CollgCr	1	-20846	9085.523663	-2.29
Neighborhood Crawfor	1	-11832	9735.585870	-1.22
Neighborhood Edwards	1	-44564	9384.758640	-4.75
Neighborhood Gilbert	1	-20714	9389.515096	-2.21
Neighborhood IDOTRR	1	-49267	10218	-4.82
Neighborhood MeadowV	1	-21426	11605	-1.85
Neighborhood Mitchel	1	-31967	9776.667107	-3.27
Neighborhood NAmes	1	-37252	9109.777197	-4.09
Neighborhood NPKvill	1	-10921	13243	-0.82
Neighborhood NWAmes	1	-30911	9542.657568	-3.24
Neighborhood NoRidge	1	18607	9985.134713	1.86
Neighborhood NridgHt	1	24169	9538.973472	2.53
Neighborhood OldTown	1	-47004	9380.734580	-5.01
Neighborhood SWISU	1	-52040	10704	-4.86
Neighborhood Sawyer	1	-33888	9558.483583	-3.55
Neighborhood SawyerW	1	-27646	9529.227186	-2.90
Neighborhood Somerst	1	1083.589725	9375.122303	0.12
Neighborhood StoneBr	1	33689	10549	3.19
Neighborhood Timber	1	-17183	9873.938952	-1.74
Neighborhood Veenker	0	0	.	.
OverallQual	1	13747	1010.396593	13.61
ExterQual Ex	1	61646	5435.303177	11.34
ExterQual Fa	1	-4045.246585	8079.321547	-0.50
ExterQual Gd	1	9922.462958	2536.197499	3.91
ExterQual TA	0	0	.	.
BsmtExposure 0	1	-789.054205	4985.392953	-0.16
BsmtExposure Av	1	11100	2363.318582	4.70
BsmtExposure Gd	1	33250	3062.316378	10.86
BsmtExposure Mn	1	8970.798814	2889.601355	3.10
BsmtExposure No	0	0	.	.
BsmtFinSF1	1	24.467273	2.048623	11.94
GrLivArea	1	60.931671	2.057459	29.62

Appendix 13 : Analysis 2 | Custom Selection

The GLMSELECT Procedure	
Data Set	WORK.TRAIN4
Dependent Variable	SalePrice
Selection Method	None
Number of Observations Read	2919
Number of Observations Used	1460

The GLMSELECT Procedure Least Squares Model (No Selection)					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	58	7.879044E12	1.358456E11	143.22	<.0001
Error	1401	1.328867E12	948513132		
Corrected Total	1459	9.207911E12			
Root MSE		30798			
Dependent Mean		180921			
R-Square		0.8557			
Adj R-Sq		0.8497			
AIC		31699			
AICC		31704			
SBC		30548			

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	-427134	165878	-2.57	0.0101
LotFrontage	1	-20.289697	25.933358	-0.78	0.4341
LotArea	1	0.545618	0.096061	5.68	<.0001
OverallQual	1	10486	1163.841790	9.01	<.0001
OverallCond	1	6775.936796	881.410022	7.69	<.0001
YearBuilt	1	298.450330	76.443384	3.90	<.0001
BsmtFinSF1	1	11.857029	2.310627	5.13	<.0001
TotalBsmtSF	1	3.199482	4.882422	0.66	0.5124
A1stFlrSF	1	27.624703	18.947882	1.46	0.1451
A2ndFlrSF	1	18.648615	18.375416	1.01	0.3103
GrLivArea	1	19.129055	18.322085	1.04	0.2966
FullBath	1	2430.130029	2385.894354	1.02	0.3086
TotRmsAbvGrd	1	1163.576980	973.104965	1.20	0.2320
GarageYrBlt	1	-95.689282	68.537013	-1.40	0.1629
GarageCars	1	14015	2772.172788	5.06	<.0001
GarageArea	1	1.331760	9.463098	0.14	0.8881
WoodDeckSF	1	23.713846	7.144184	3.32	0.0009
OpenPorchSF	1	5.858968	13.737009	0.43	0.6698
KitchenQual Ex	1	30562	4820.201639	6.34	<.0001
KitchenQual Fa	1	673.600459	5631.775551	0.12	0.9048
KitchenQual Gd	1	4211.597007	2560.005327	1.65	0.1002
KitchenQual TA	0	0			
GarageFinish 0	1	-172383	133386	-1.29	0.1964
GarageFinish Fin	1	6447.686766	2862.799050	2.25	0.0245
GarageFinish RFn	1	2305.719977	2470.049925	0.93	0.3507
GarageFinish Unf	0	0			
BsmtQual 0	1	-9626.708062	7309.116629	-1.32	0.1880
BsmtQual Ex	1	30387	5025.225166	6.05	<.0001
BsmtQual Fa	1	-1825.588841	5688.341768	-0.32	0.7483
BsmtQual Gd	1	-2523.238208	2957.712595	-0.85	0.3937
BsmtQual TA	0	0			
ExterQual Ex	1	19183	6503.615057	2.95	0.0032
ExterQual Fa	1	2890.634534	9229.274923	0.31	0.7542
ExterQual Gd	1	1655.229412	3047.200080	0.54	0.5871

BsmtQual TA	0	0	.	.	.
ExterQual Ex	1	19183	6503.615057	2.95	0.0032
ExterQual Fa	1	2890.634534	9229.274923	0.31	0.7542
ExterQual Gd	1	1655.229412	3047.200080	0.54	0.5871
ExterQual TA	0	0	.	.	.
MasVnrType 0	1	-727.980970	11377	-0.06	0.9490
MasVnrType BrkCm	1	-13212	8855.522649	-1.49	0.1359
MasVnrType BrkFa	1	-1912.709008	3508.983949	-0.55	0.5858
MasVnrType None	1	470.556562	3628.403230	0.13	0.8968
MasVnrType Stone	0	0	.	.	.
Neighborhood Blmngtn	1	-30558	12415	-2.46	0.0140
Neighborhood Blueste	1	-50121	24004	-2.09	0.0370
Neighborhood BrDale	1	-45077	12846	-3.51	0.0005
Neighborhood BrkSide	1	-24276	10848	-2.24	0.0254
Neighborhood ClearCr	1	-13600	11278	-1.21	0.2281
Neighborhood CollgCr	1	-15807	9902.689966	-1.60	0.1107
Neighborhood Crawfor	1	-1818.648751	10774	-0.17	0.8680
Neighborhood Edwards	1	-39691	10274	-3.86	0.0001
Neighborhood Gilbert	1	-22840	10355	-2.21	0.0276
Neighborhood IDOTRR	1	-35346	11470	-3.08	0.0021
Neighborhood MeadowV	1	-44209	12376	-3.57	0.0004
Neighborhood Mitchel	1	-36508	10566	-3.46	0.0006
Neighborhood NAmes	1	-28200	9988.832172	-2.82	0.0048
Neighborhood NPKvill	1	-38212	14239	-2.68	0.0074
Neighborhood NWAmes	1	-31493	10326	-3.05	0.0023
Neighborhood NoRidge	1	38068	10852	3.51	0.0005
Neighborhood NridgHt	1	8794.347363	10577	0.83	0.4059
Neighborhood OldTown	1	-40255	10637	-3.78	0.0002
Neighborhood SWISU	1	-33778	12118	-2.79	0.0054
Neighborhood Sawyer	1	-29777	10381	-2.87	0.0041
Neighborhood SawyerW	1	-22278	10356	-2.15	0.0316
Neighborhood Somerst	1	-8774.462064	10342	-0.85	0.3963
Neighborhood StoneBr	1	25030	11373	2.20	0.0279
Neighborhood Timber	1	-18263	10808	-1.69	0.0913
Neighborhood Veenker	0	0	.	.	.

APPENDIX 14 : Kaggle Submission Score

Submission and Description	Public Score	Use for Final Score
submit_stepwise3.csv 11 minutes ago by Andy H add submission details	0.15937	<input checked="" type="checkbox"/>
submit_forward2.csv a day ago by Andy H add submission details	0.15937	<input checked="" type="checkbox"/>
submit_custom2.csv a day ago by Andy H add submission details	0.18171	<input checked="" type="checkbox"/>
submit_backwards.csv a day ago by Andy H add submission details	0.17309	<input checked="" type="checkbox"/>