STAT 448 HW #3

Tianqi Wu

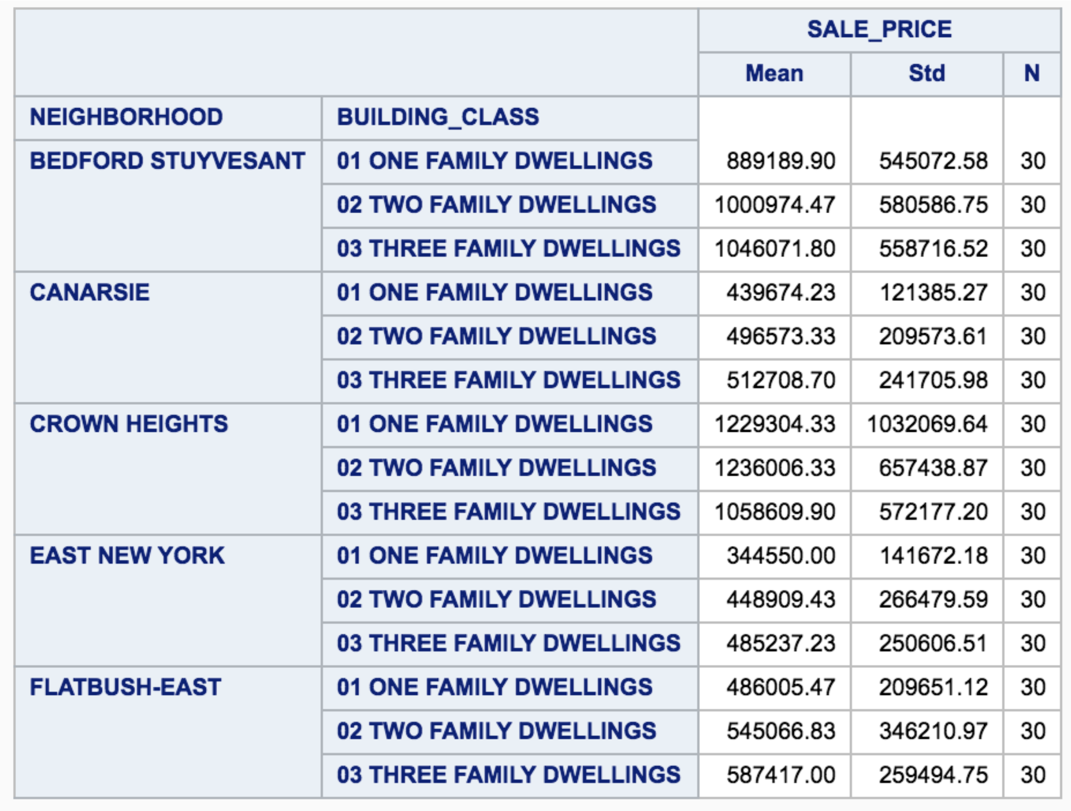
2018/10/03

Problem 1

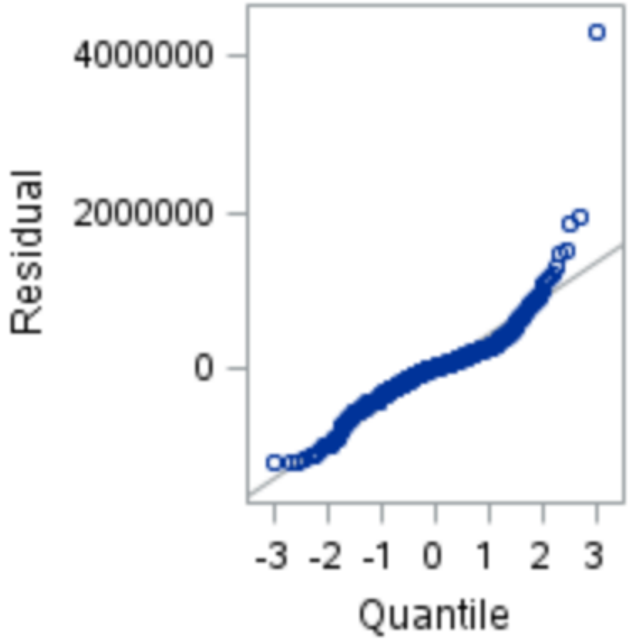
Here is the code that produces the table:

```
proc tabulate data = bkhomessmall;
 class NEIGHBORHOOD BUILDING_CLASS;
 var SALE_PRICE;
 table NEIGHBORHOOD*BUILDING_CLASS,
 SALE_PRICE*(mean std n);
run;

proc glm data = bkhomessmall plots= diagnostics;
 class NEIGHBORHOOD BUILDING_CLASS;
 model SALE_PRICE = NEIGHBORHOOD BUILDING_CLASS;
run;
```

From the tabulate, the data are balanced. From the QQ plot, the data does not depart much from a normal distribution since most points fall in a straight line. Hence, the assumptions of the general ANOVA are not violated.



		SALE_PRICE		
		Mean	Std	N
NEIGHBORHOOD	BUILDING_CLASS			
BEDFORD STUYVESANT	01 ONE FAMILY DWELLINGS	889189.90	545072.58	30
	02 TWO FAMILY DWELLINGS	1000974.47	580586.75	30
	03 THREE FAMILY DWELLINGS	1046071.80	558716.52	30
CANARSIE	01 ONE FAMILY DWELLINGS	439674.23	121385.27	30
	02 TWO FAMILY DWELLINGS	496573.33	209573.61	30
	03 THREE FAMILY DWELLINGS	512708.70	241705.98	30
CROWN HEIGHTS	01 ONE FAMILY DWELLINGS	1229304.33	1032069.64	30
	02 TWO FAMILY DWELLINGS	1236006.33	657438.87	30
	03 THREE FAMILY DWELLINGS	1058609.90	572177.20	30
EAST NEW YORK	01 ONE FAMILY DWELLINGS	344550.00	141672.18	30
	02 TWO FAMILY DWELLINGS	448909.43	266479.59	30
	03 THREE FAMILY DWELLINGS	485237.23	250606.51	30
FLATBUSH-EAST	01 ONE FAMILY DWELLINGS	486005.47	209651.12	30
	02 TWO FAMILY DWELLINGS	545066.83	346210.97	30
	03 THREE FAMILY DWELLINGS	587417.00	259494.75	30



Problem 2

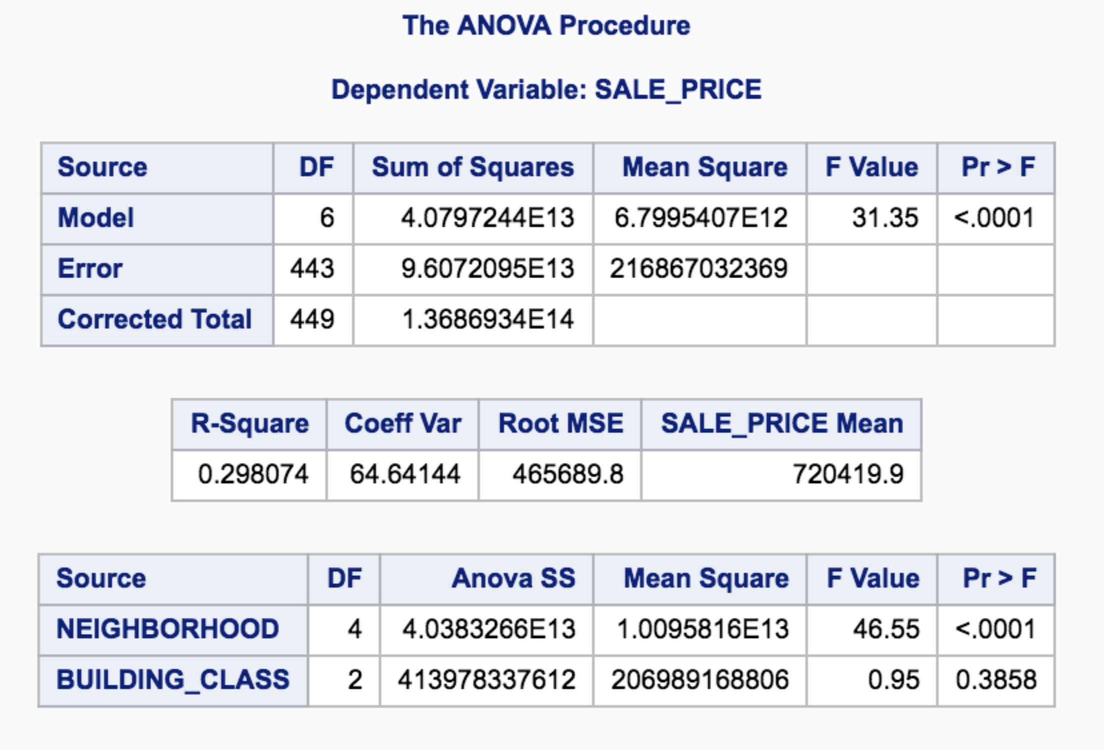
Here is the code that produces the ANOVA:

```
proc anova data = bkhomessmall;
 class NEIGHBORHOOD BUILDING_CLASS;
 model SALE_PRICE = NEIGHBORHOOD BUILDING_CLASS;
run;
```

From the ANOVA result, the overall model significance (p-value<0.0001) indicates that not all predictors' group means of sale prices are the same.

The effect neighborhood's significance (p-value<0.0001) indicates not all group means of this effect are equal. The effect building class's significance (p-value > 0.3858) indicates that there is not enough evidence to suggest a difference of response mean for all levels of this effect. Hence, main effect neighborhood is significant while building class is not.

From the Tukey test of neighborhood, East NewYork, Canarsie and Flatbush-East are not significantly different since they are covered by the same bar. Crown Heights and Bedford Stuyvesant are significantly different from other groups and each other.



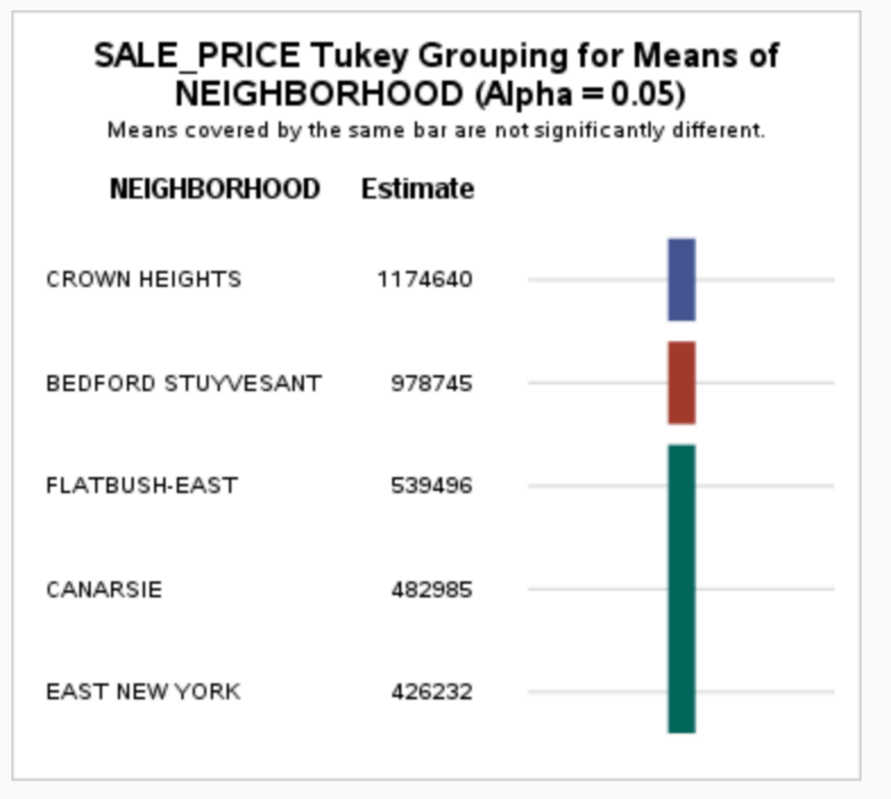
The ANOVA Procedure

Dependent Variable: SALE_PRICE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	4.0797244E13	6.7995407E12	31.35	<.0001
Error	443	9.6072095E13	216867032369		
Corrected Total	449	1.3686934E14			

R-Square	Coeff Var	Root MSE	SALE_PRICE Mean
0.298074	64.64144	465689.8	720419.9

Source	DF	Anova SS	Mean Square	F Value	Pr > F
NEIGHBORHOOD	4	4.0383266E13	1.0095816E13	46.55	<.0001
BUILDING_CLASS	2	413978337612	206989168806	0.95	0.3858



SALE_PRICE Tukey Grouping for Means of NEIGHBORHOOD (Alpha = 0.05)

Means covered by the same bar are not significantly different.

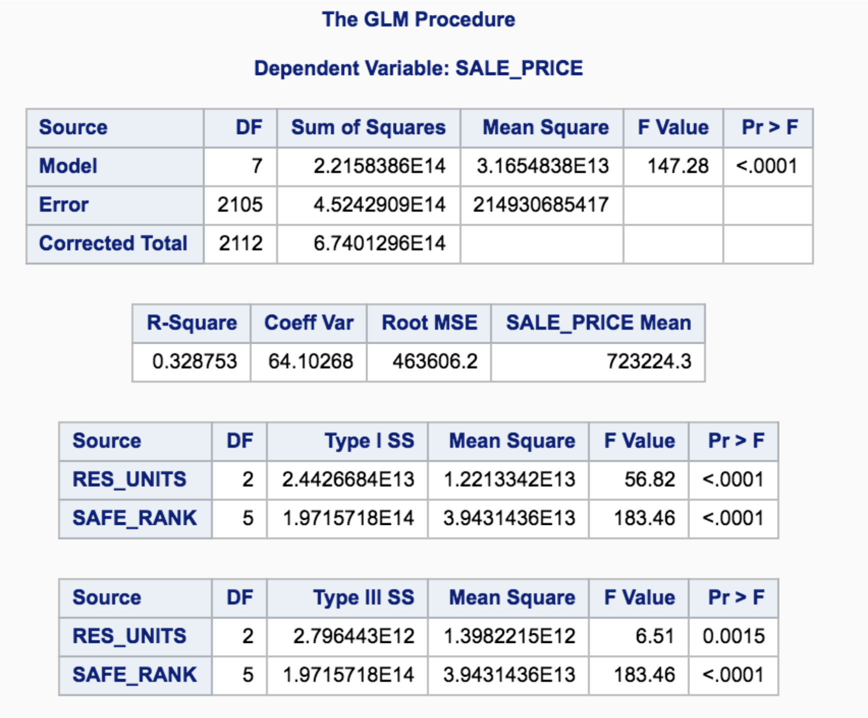
Problem 3

Here is the code that produces the ANOVA:

```
proc glm data = bkhomes;
 class RES_UNITS SAFE_RANK;
 model SALE_PRICE = RES_UNITS SAFE_RANK;
 lsmeans RES_UNITS SAFE_RANK/ adjust=tukey cl;
run;
```

From the ANOVA result, the overall model significance (p-value<0.0001) indicates that not all predictors' group means of sale prices are the same. The effect residential units' significance (p-value<0.0001) indicates that not all group means of this effect are equal. The effect safety ranking's significance (p-value<0.0001) indicates that not all group means of this effect are equal. Hence, main effects residential units and safety ranking are significant and both of them should be kept. The coefficient of determination indicates that the model explains 32.875% of variation in the response.

For the main effect residential units, group of 1 unit appears to be different from other groups. For the main effect safety ranking, safe rank of 21 is different from all other groups. Safe rank of 20 is different from all other groups except safety rank of 13. Safe rank of 18 is different from all other groups except safety rank of 6 and 11. Safe rank of 13 is different from all other groups except safety rank of 20. Safe rank of 11 is different from all other groups except safety rank of 6 and 18. Safe rank of 6 is different from all other groups except safety rank of 11 and 18.



The GLM Procedure

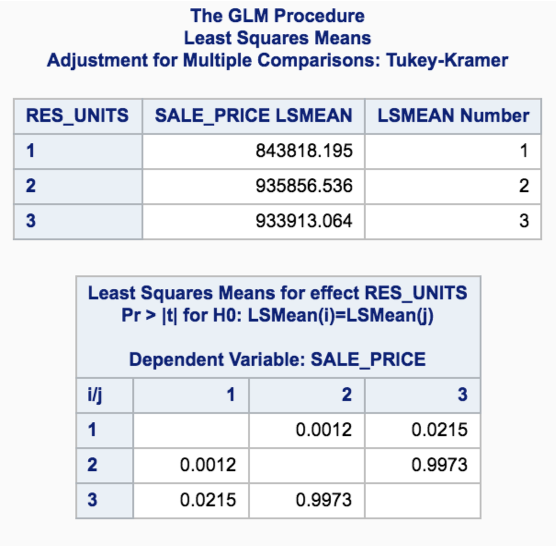
Dependent Variable: SALE_PRICE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	2.2158386E14	3.1654838E13	147.28	<.0001
Error	2105	4.5242909E14	214930685417		
Corrected Total	2112	6.7401296E14			

R-Square	Coeff Var	Root MSE	SALE_PRICE Mean
0.328753	64.10268	463606.2	723224.3

Source	DF	Type I SS	Mean Square	F Value	Pr > F
RES_UNITS	2	2.4426684E13	1.2213342E13	56.82	<.0001
SAFE_RANK	5	1.9715718E14	3.9431436E13	183.46	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
RES_UNITS	2	2.796443E12	1.3982215E12	6.51	0.0015
SAFE_RANK	5	1.9715718E14	3.9431436E13	183.46	<.0001



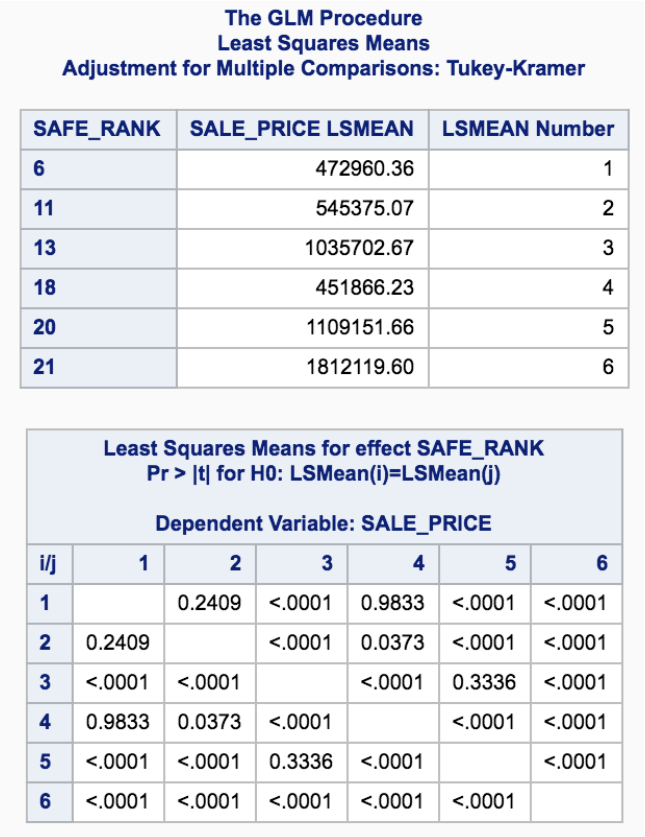
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

RES_UNITS	SALE_PRICE LSMEAN	LSMEAN Number
1	843818.195	1
2	935856.536	2
3	933913.064	3

Least Squares Means for effect RES_UNITS
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: SALE_PRICE

i/j	1	2	3
1		0.0012	0.0215
2	0.0012		0.9973
3	0.0215	0.9973	



The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

SAFE_RANK	SALE_PRICE LSMEAN	LSMEAN Number
6	472960.36	1
11	545375.07	2
13	1035702.67	3
18	451866.23	4
20	1109151.66	5
21	1812119.60	6

Least Squares Means for effect SAFE_RANK
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: SALE_PRICE

i/j	1	2	3	4	5	6
1		0.2409	<.0001	0.9833	<.0001	<.0001
2	0.2409		<.0001	0.0373	<.0001	<.0001
3	<.0001	<.0001		<.0001	0.3336	<.0001
4	0.9833	0.0373	<.0001		<.0001	<.0001
5	<.0001	<.0001	0.3336	<.0001		<.0001
6	<.0001	<.0001	<.0001	<.0001	<.0001	

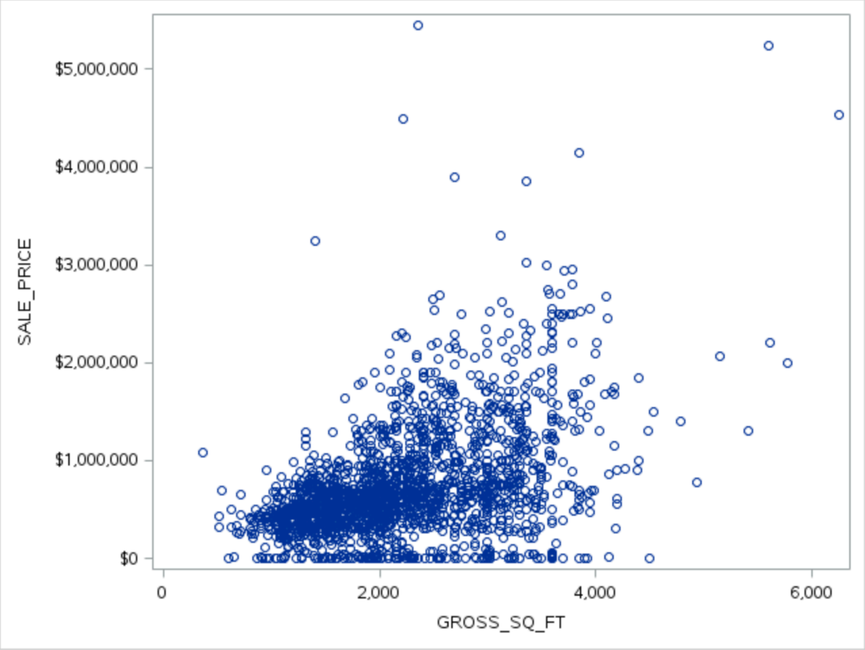
Problem 4

Here is the code producing the linear regression:

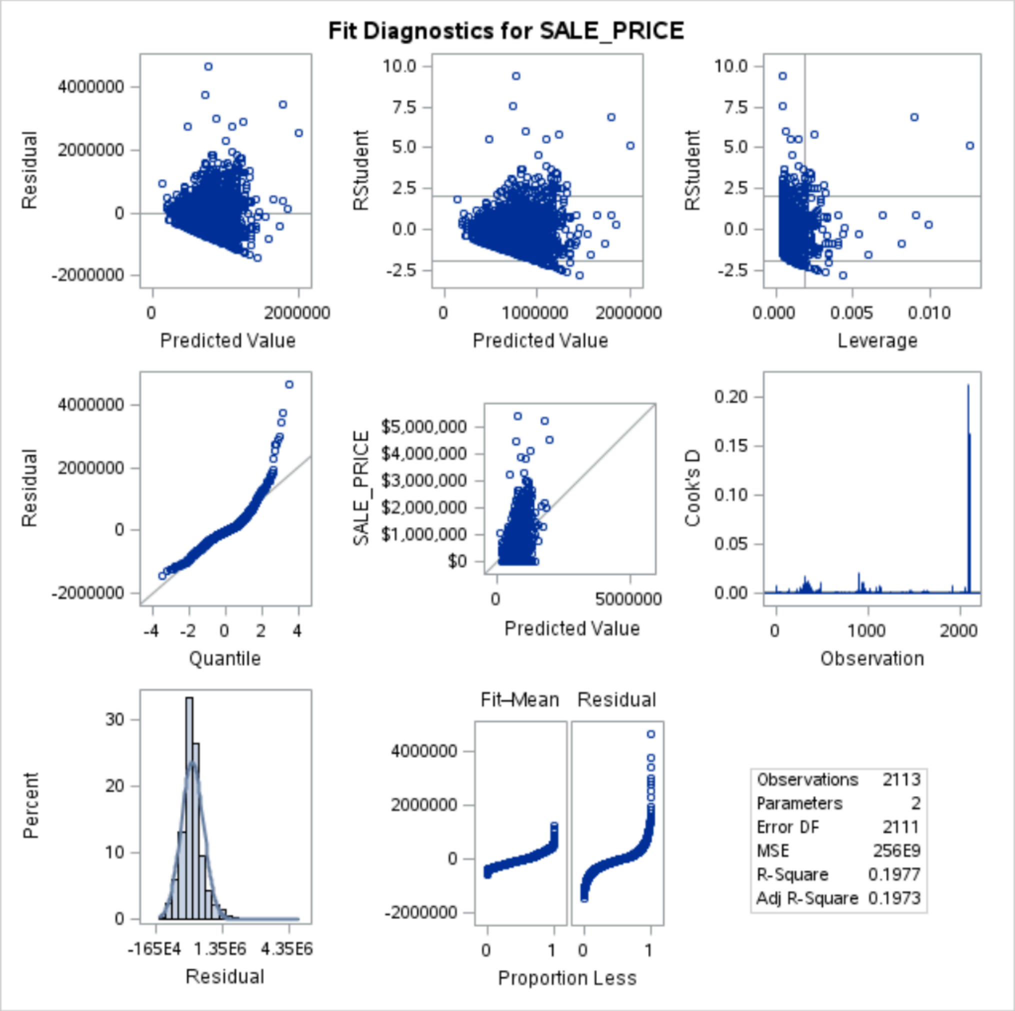
```
proc sgplot data=bkhomes;
 scatter y = SALE_PRICE x = GROSS_SQ_FT;
run;

proc reg data=bkhomes plots=diagnostics;
 model SALE_PRICE = GROSS_SQ_FT;
run;
```

From the scatter plot, there might be a moderate positive linear relationship between sale price and gross square footage.

From the linear regression result, the coefficient of determination indicates that the model captures 19.77% of variation in the response. From the QQ plot, the data does not depart much from a normal distribution since most points fall in a straight line. Looking at the histogram, it seems that the distribution is reasonably bell-shaped but the tails appear to deviate from normality. For the residual vs. prediction values plot, there is no strong evidence to reject the homogeneity. For Cook’s distances, there is no point having Cook’s distance greater than 1. Hence, there is no need to worry about strongly influential points.



The REG Procedure
Model: MODEL1
Dependent Variable: SALE_PRICE



Observations	2113
Parameters	2
Error DF	2111
MSE	256E9
R-Square	0.1977
Adj R-Square	0.1973

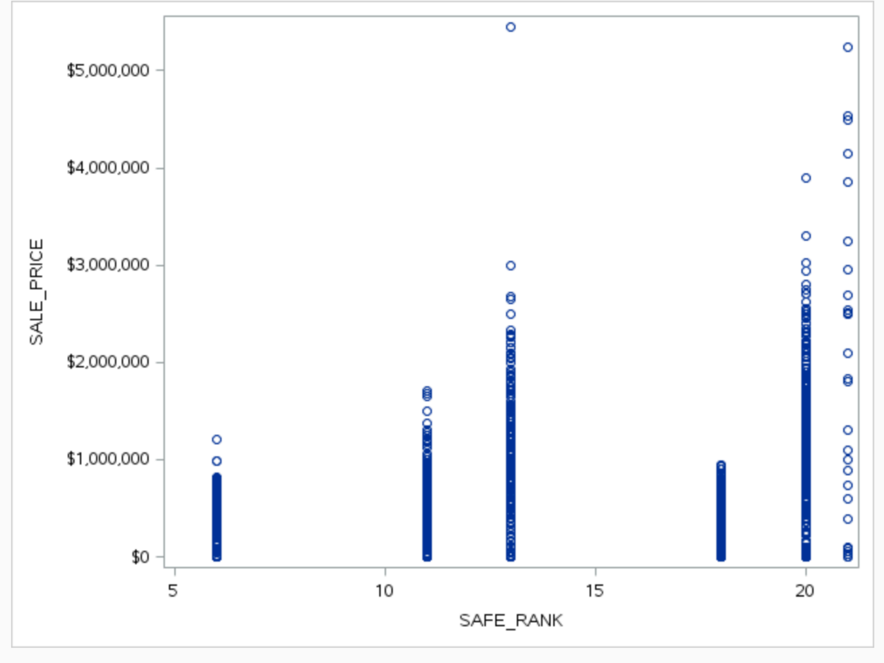
Problem 5

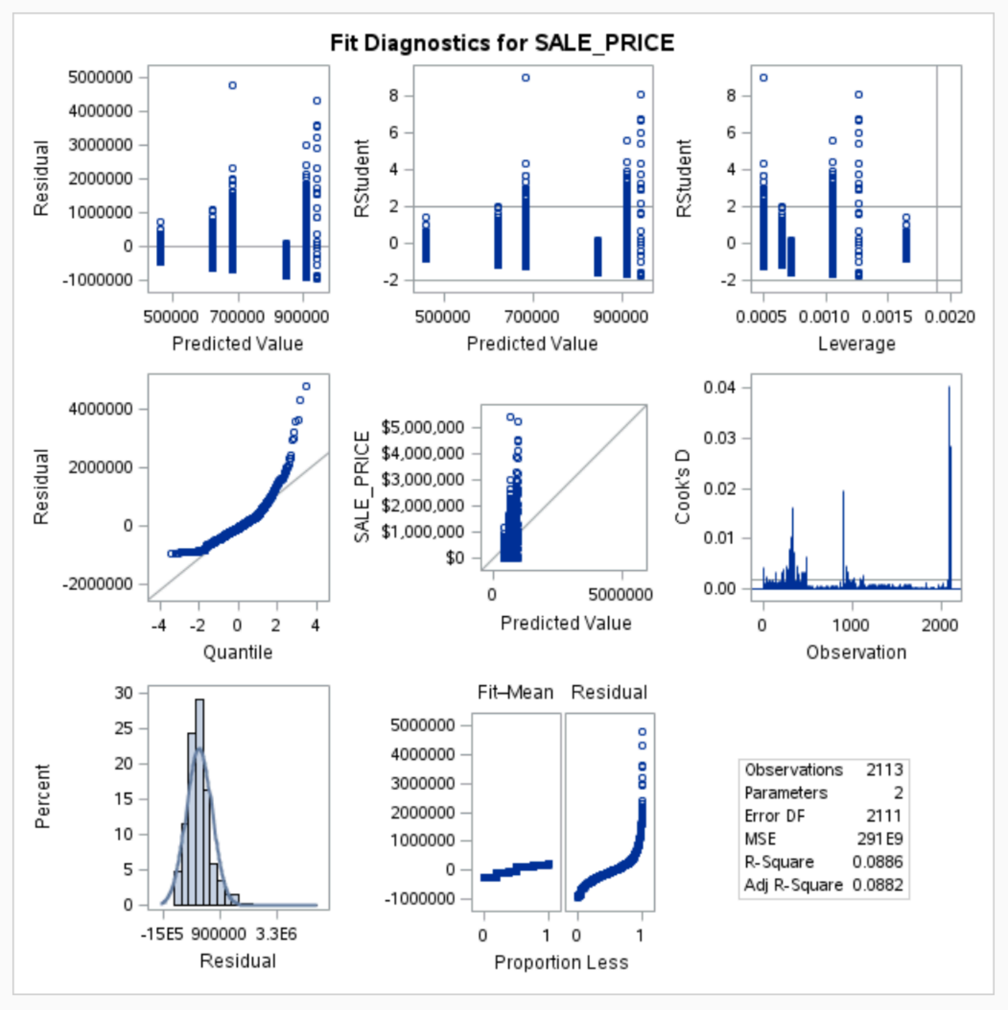
Here is the code producing the linear regression:

```
proc sgplot data=bkhomes;
 scatter y = SALE_PRICE x = SAFE_RANK;
run;

proc reg data=bkhomes plots=diagnostics;
 model SALE_PRICE = SAFE_RANK;
run;
```

From the scatter plot, sale price seems to have positive relationship with gross square footage since homes with greater safety rank have greater sale price.

From the linear regression result, the coefficient of determination indicates that the model captures 8.86% of variation in the response. From the QQ plot, the data does not depart much from a normal distribution since most points fall in a straight line. Looking at the histogram, it seems that the distribution is reasonably bell-shaped but the tails appear to deviate from normality. For the residual vs. prediction values plot, there is no strong evidence to reject the homogeneity. For Cook's distances, there is no point having Cook's distance greater than 1. Hence, there is no need to worry about strongly influential points.





Observations	2113
Parameters	2
Error DF	2111
MSE	291E9
R-Square	0.0886
Adj R-Square	0.0882

Problem 6

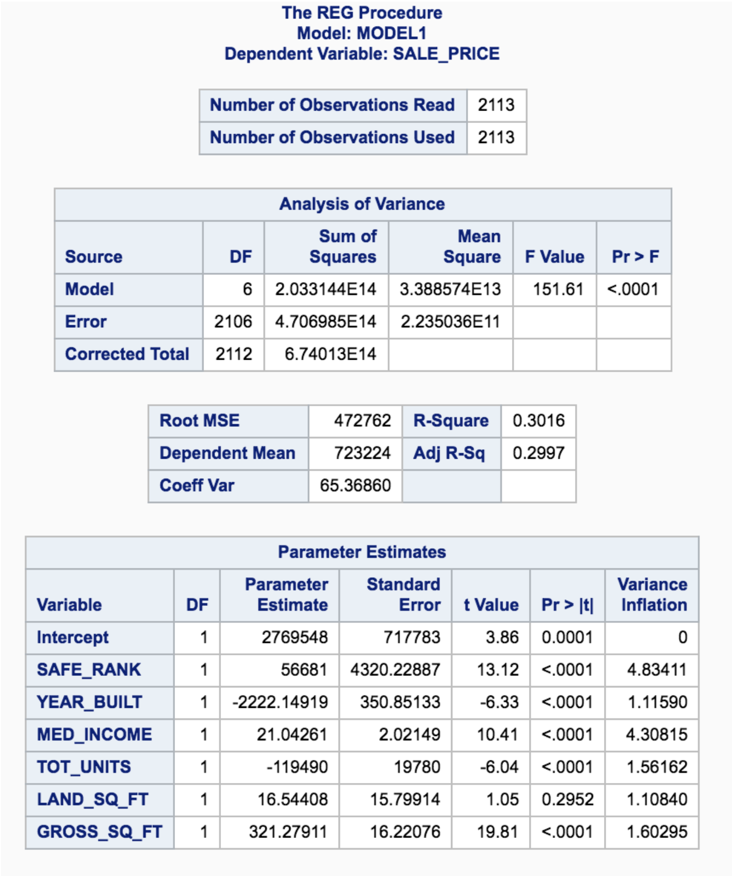
Based on the results above, linear regression model for the sale price as a function of gross square footage is better than linear regression model for the sale price as a function of safety ranking since the first model captures greater variation (19.77% vs. 8.86%) of the response with better assumptions of normality and homogeneity..

Problem 7

Here is the code producing the regression model:

```
proc reg data=bkhomes plots=diagnostics;
 model SALE_PRICE = SAFE_RANK YEAR_BUILT MED_INCOME
TOT_UNITS LAND_SQ_FT GROSS_SQ_FT;
run;
```

From the model result, The effect land square footage's significance (p-value > 0.2952) indicates that there is not enough evidence to suggest a difference of response mean for all levels of this effect. All other effects are significant with p-value < 0.0001. Also, the variance inflation factors are less than 10 for all the effects. It indicates that there is no multicollinearity existing. Hence, land square footage should be removed from the model. Other than that, there is no strong evidence violating the assumptions of the model.



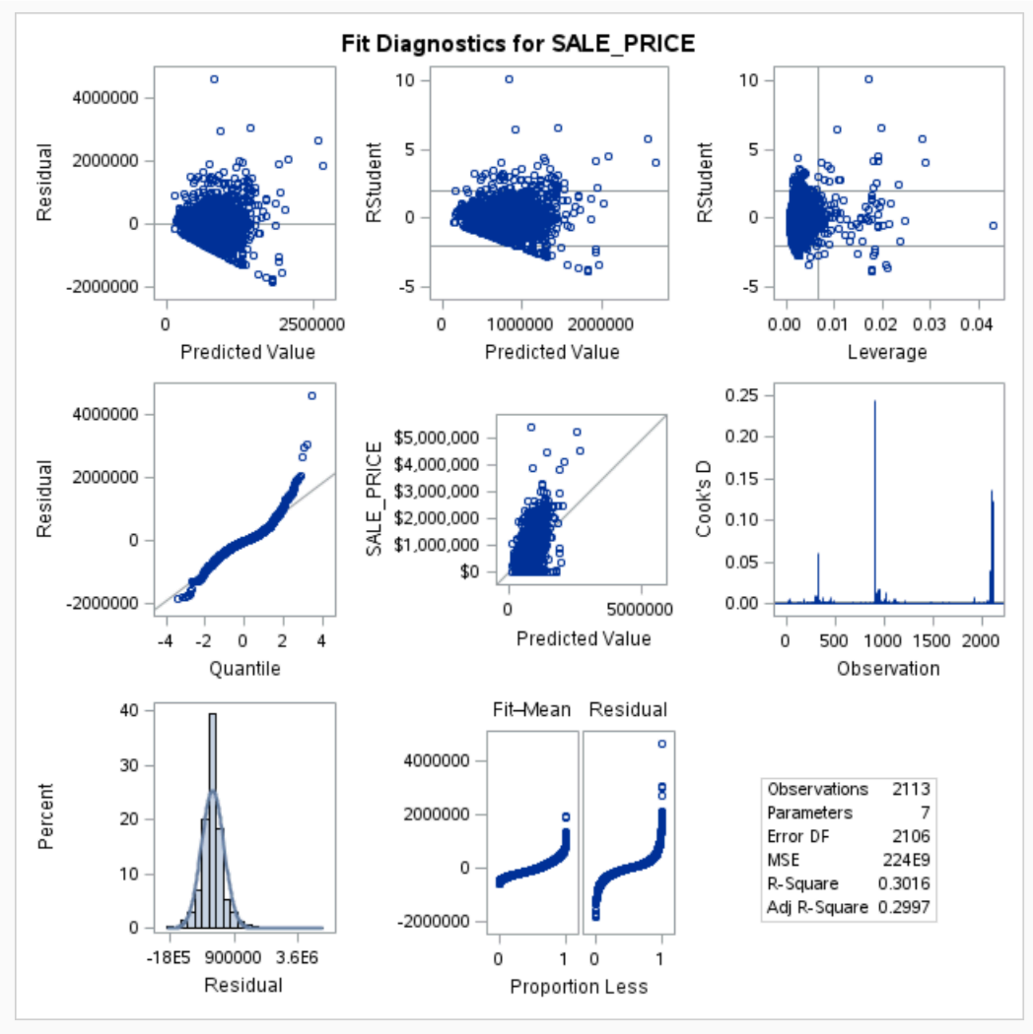
The REG Procedure
Model: MODEL1
Dependent Variable: SALE_PRICE

Number of Observations Read	2113
Number of Observations Used	2113

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	2.033144E14	3.388574E13	151.61	<.0001
Error	2106	4.706985E14	2.235036E11		
Corrected Total	2112	6.74013E14			

Root MSE	472762	R-Square	0.3016
Dependent Mean	723224	Adj R-Sq	0.2997
Coeff Var	65.36860		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	2769548	717783	3.86	0.0001	0
SAFE_RANK	1	56681	4320.22887	13.12	<.0001	4.83411
YEAR_BUILT	1	-2222.14919	350.85133	-6.33	<.0001	1.11590
MED_INCOME	1	21.04261	2.02149	10.41	<.0001	4.30815
TOT_UNITS	1	-119490	19780	-6.04	<.0001	1.56162
LAND_SQ_FT	1	16.54408	15.79914	1.05	0.2952	1.10840
GROSS_SQ_FT	1	321.27911	16.22076	19.81	<.0001	1.60295



Problem 8

The model captures 30.16% of the variance in the response variable and it indicates that the model may still need improvement. Since the effect land square footage is not significant and not enough variation is described, we would not want to keep the model. 1 unit increase of predictors leads to following change to the sale price (denoted in bracket): total units (-119490), year built (-2222), safe rank (56681), median household income (21) and gross square footage (321). Total units and year built are negatively related to sale price while safe rank, median household income and gross square footage are positively related to sale price.The order of effects' significance to the response sale price is as follows: total units > safe rank > year built > gross square footage > median household income.

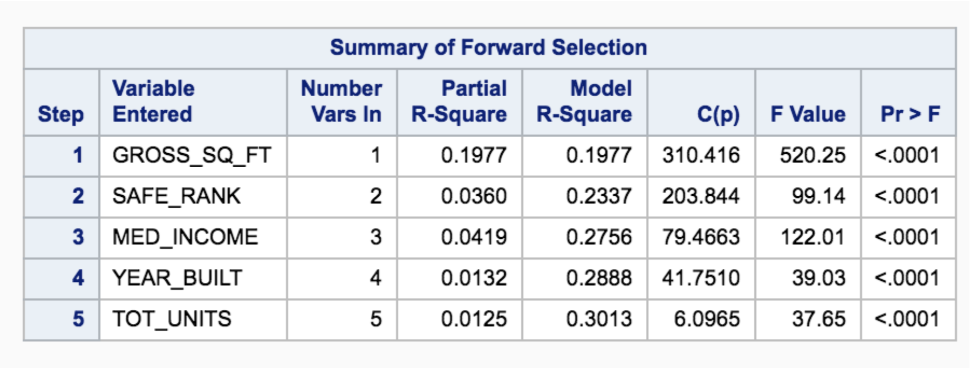
Problem 9

Here is the code producing automatic procedure and information criterion measure:

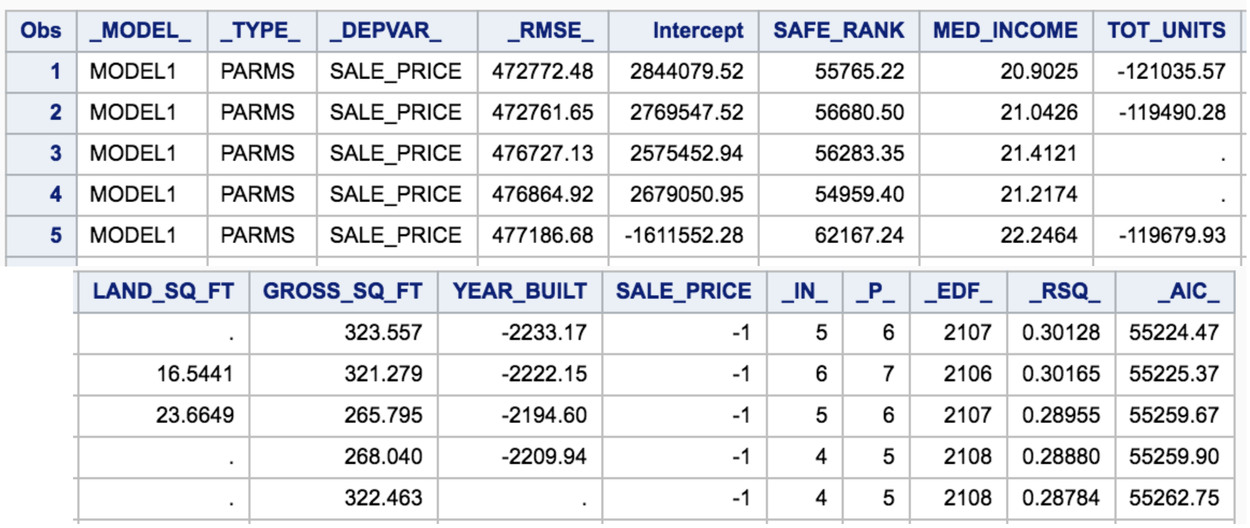
```
proc reg data = bkhomes;
 model SALE_PRICE = SAFE_RANK MED_INCOME TOT_UNITS--
YEAR_BUILT / selection = f sle=0.05;
run;
proc reg data = bkhomes outest= bkhomes_2;
 model SALE_PRICE = SAFE_RANK MED_INCOME TOT_UNITS--
YEAR_BUILT / selection = adjrsq aic;
run;
proc sort data = bkhomes_2;
 by _aic_ _rsq_;
run;
proc print data = bkhomes_2;
run;
proc reg data=bkhomes plots=diagnostics;
 model SALE_PRICE = SAFE_RANK MED_INCOME TOT_UNITS
GROSS_SQ_FT YEAR_BUILT;
run;
```

The result of forward automatic selection procedure agrees with the result of AIC criterion measure that safety ranking, year built, median household income, total units and gross square footage should be kept in the model.

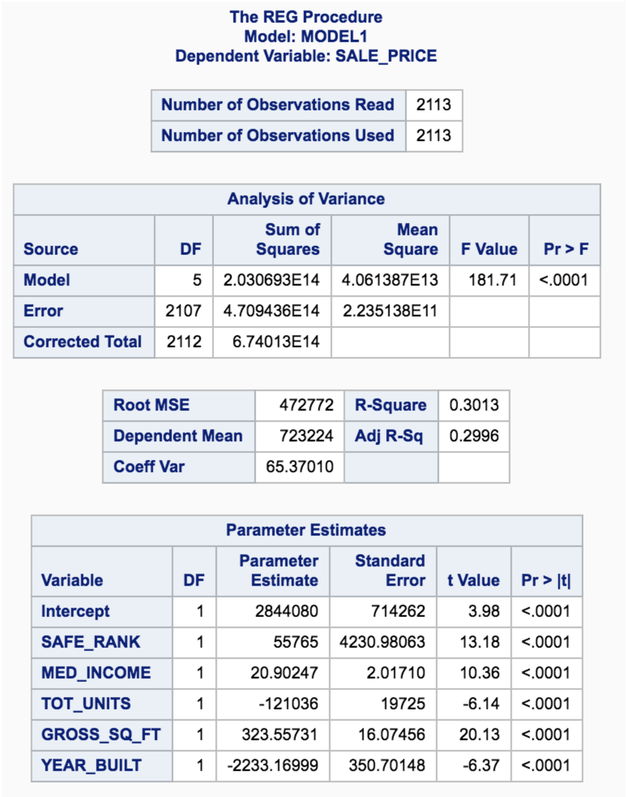
From the QQ plot of the linear regression result, the data does not depart much from a normal distribution since most points fall in a straight line. Looking at the histogram, it seems that the distribution is reasonably bell-shaped but the tails appear to deviate from normality. For the residual vs. prediction values plot, there is no strong evidence to reject the homogeneity. For Cook's distances, there is no point having Cook's distance greater than 1. Hence, there is no need to worry about strongly influential points.



Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	GROSS_SQ_FT	1	0.1977	0.1977	310.416	520.25	<.0001
2	SAFE_RANK	2	0.0360	0.2337	203.844	99.14	<.0001
3	MED_INCOME	3	0.0419	0.2756	79.4663	122.01	<.0001
4	YEAR_BUILT	4	0.0132	0.2888	41.7510	39.03	<.0001
5	TOT_UNITS	5	0.0125	0.3013	6.0965	37.65	<.0001



Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	SAFE_RANK	MED_INCOME	TOT_UNITS	LAND_SQ_FT	GROSS_SQ_FT	YEAR_BUILT	SALE_PRICE	_IN_	_P_	_EDF_	_RSQ_	_AIC_
1	MODEL1	PARMS	SALE_PRICE	472772.48	2844079.52	55765.22	20.9025	-121035.57	.	323.557	-2233.17	-1	5	6	2107	0.30128	55224.47
2	MODEL1	PARMS	SALE_PRICE	472761.65	2769547.52	56680.50	21.0426	-119490.28	16.5441	321.279	-2222.15	-1	6	7	2106	0.30165	55225.37
3	MODEL1	PARMS	SALE_PRICE	476727.13	2575452.94	56283.35	21.4121	.	23.6649	265.795	-2194.60	-1	5	6	2107	0.28955	55259.67
4	MODEL1	PARMS	SALE_PRICE	476864.92	2679050.95	54959.40	21.2174	.	.	268.040	-2209.94	-1	4	5	2108	0.28880	55259.90
5	MODEL1	PARMS	SALE_PRICE	477186.68	-1611552.28	62167.24	22.2464	-119679.93	.	322.463	.	-1	4	5	2108	0.28784	55262.75



The REG Procedure
Model: MODEL1
Dependent Variable: SALE_PRICE

Number of Observations Read	2113
Number of Observations Used	2113

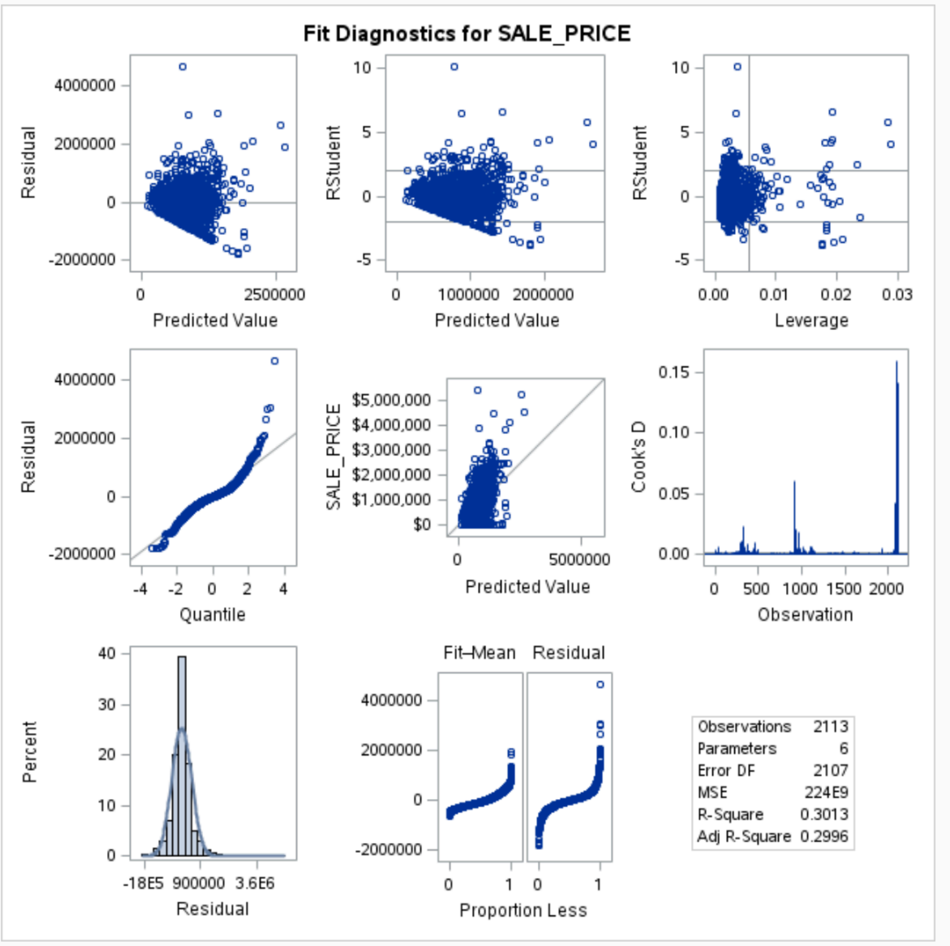
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2.030693E14	4.061387E13	181.71	<.0001
Error	2107	4.709436E14	2.235138E11		
Corrected Total	2112	6.74013E14			

Root MSE	472772	R-Square	0.3013
Dependent Mean	723224	Adj R-Sq	0.2996
Coeff Var	65.37010		

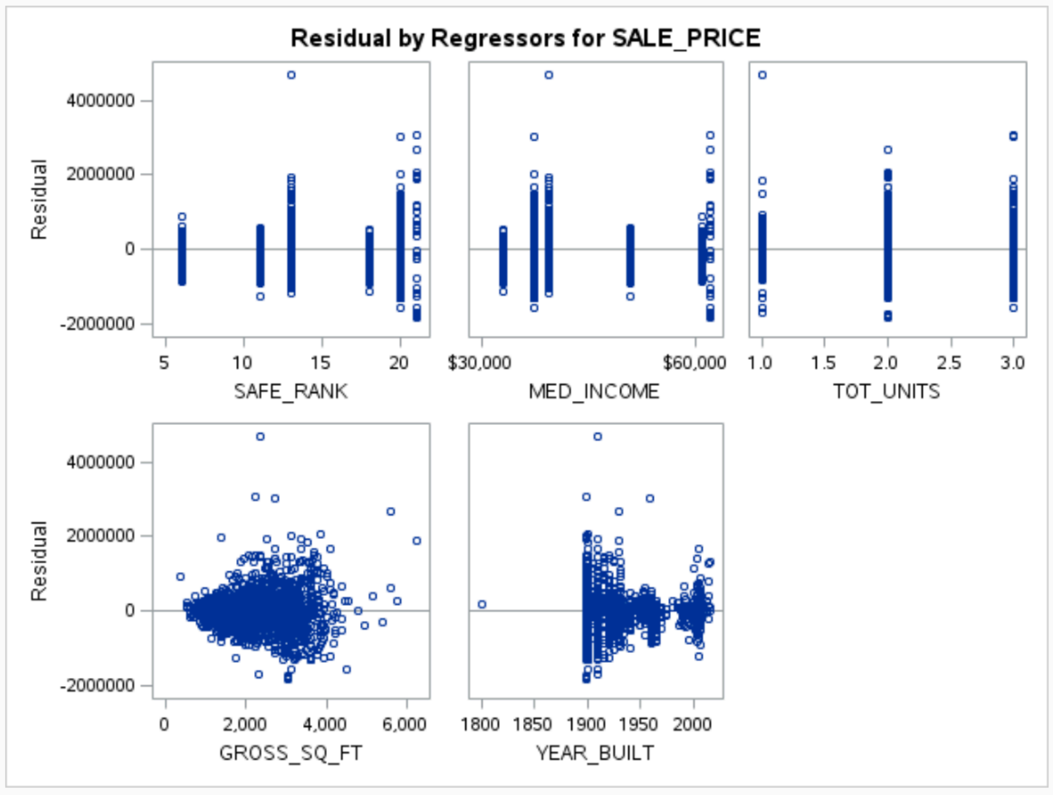
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2844080	714262	3.98	<.0001
SAFE_RANK	1	55765	4230.98063	13.18	<.0001
MED_INCOME	1	20.90247	2.01710	10.36	<.0001
TOT_UNITS	1	-121036	19725	-6.14	<.0001
GROSS_SQ_FT	1	323.55731	16.07456	20.13	<.0001
YEAR_BUILT	1	-2233.16999	350.70148	-6.37	<.0001



Fit Diagnostics for SALE_PRICE

Observations	2113
Parameters	6
Error DF	2107
MSE	224E9
R-Square	0.3013
Adj R-Square	0.2996



Problem 10

From the regression result, the overall model significance (p-value<0.0001) indicates that not all parameter estimates of the effects are zero. The coefficient of determination indicates that the model captures 30.13% of variation in the response. As the previous problem stated, there is no lingering diagnostic issues looking at the residual plots. The overall mean of sale price is 2844080. 1 unit increase of predictors leads to following change to the sale price (denoted in bracket): total units (-121036), year built (-2233), safe rank (55765), median household income (21) and gross square footage (323). Total units and year built are negatively related to sale price while safe rank, median household income and gross square footage are positively related to sale price.The order of effects' significance to the response sale price is as follows: total units > safe rank > year built > gross square footage > median household income.