**PROJECT 2 REPORT**

**SIN\_SRE**

**JASBIR SINGH**

**MADHUVANTHI SREEDHAR**

**Problem1**

|  |  |
| --- | --- |
| Categorical | House\_Style Heating\_QC Central\_Air Bedroom\_AbvGr Fireplaces Mo\_Sold Yr\_Sold Full\_Bathroom Half\_Bathroom Total\_Bathroom Season\_Sold Garage\_Type\_2 Foundation\_2 Masonry\_Veneer Lot\_Shape\_2 House\_Style2 Overall\_Qual Overall\_Qual2 Overall\_Cond Overall\_Cond2 |
| Interval | Lot\_Area Year\_Built Gr\_Liv\_Area Garage\_Area SalePrice Basement\_Area Deck\_Porch\_Area Age\_Sold Log\_Price; |

**Problem 2**

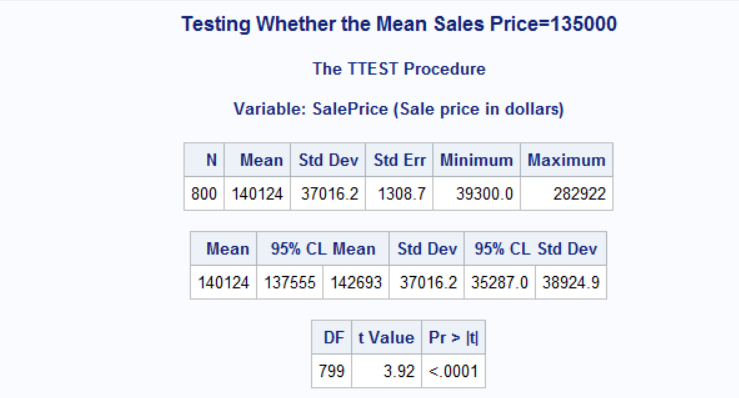
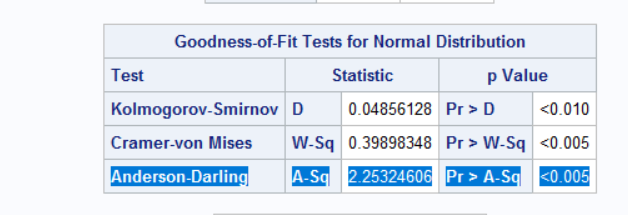
**Exploring continuous variables:**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Statistics** | **Plot** |
| Year\_Built |  |  |
| Gr\_Liv\_Area |  |  |
| Garage\_Area |  |  |
| SalePrice |  |  |
| Basement\_Area |  |  |
| Deck\_  Porch\_  Area |  |  |
| Age\_Sold |  |  |
| Log\_Price |  |  |

**Exploring categorical variables:**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Statistics** | **Plot** |
| Heating\_QC |  |  |
| House\_Style |  |  |
| Central\_Air |  |  |
| Bedroom\_  Abv\_Gr |  |  |
| Fireplaces |  |  |
| Mo\_Sold |  |  |
| Yr\_Sold |  |  |
| Full\_Bathroom |  |  |
| Half\_Bathroom |  |  |
| Total\_Bathroom |  |  |
| Season\_Sold |  |  |
| Garage\_Type\_2 |  |  |
| Foundation\_2 |  |  |
| Masonry\_Veneer |  |  |
| Lot\_Shape\_2 |  |  |
| House\_Style2 |  |  |
| Overall\_Qual |  |  |
| Overall\_Qual2 |  |  |
| Overall\_Cond |  |  |
| Overall\_Cond2 |  |  |

**Problem 3**



Variable should be normally distributed to perform TTESTS analysis , SalePrice Distribution is plotted to check the same.

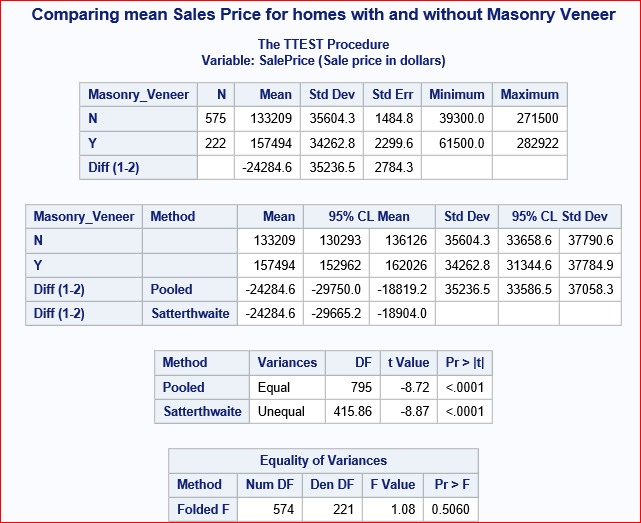
As data is not normally distributed as confirmed by **Anderson-Darling** statistics , we should not be performing TTEST in most of the cases , but the standard deviation of the population isn't known. Thus, we need to use the standard error of the mean, where sample standard deviation is used, which gets us to use the ttest to test the hypothesis.

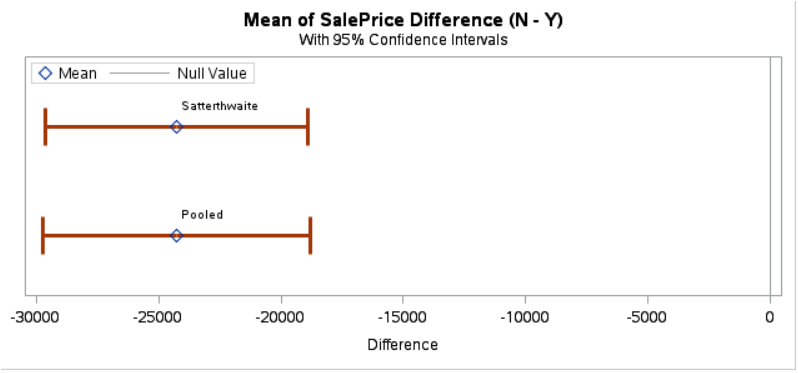
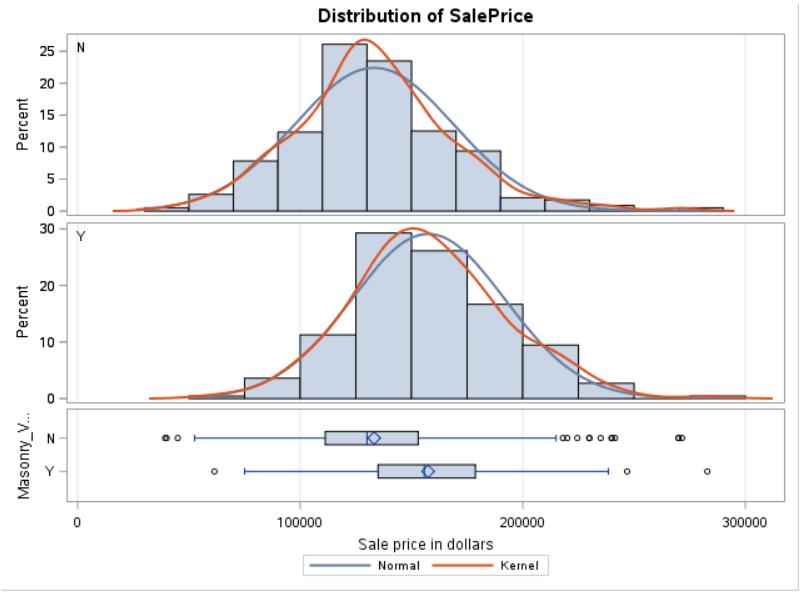
T-Test to check mean of SalePrice is $135,000 in the data set

We reject the null Hypothesis , Mean is not equal to 135000 at alpha value of 0.05

**Problem 4**

The masonry\_veneer variable has values N and Y indicating houses with and without Masonry Veneer.The Sale Price of these houses are compared using Two-sample T-test.



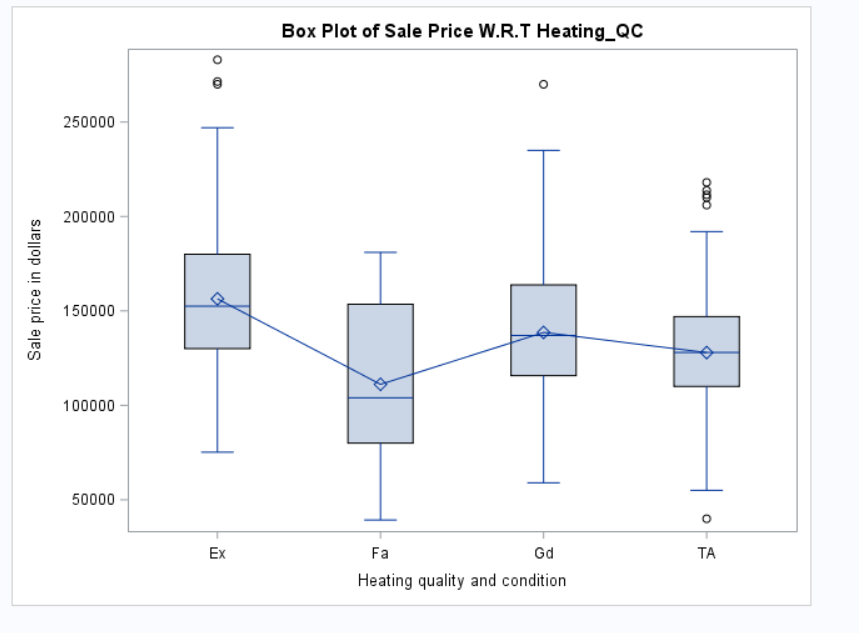
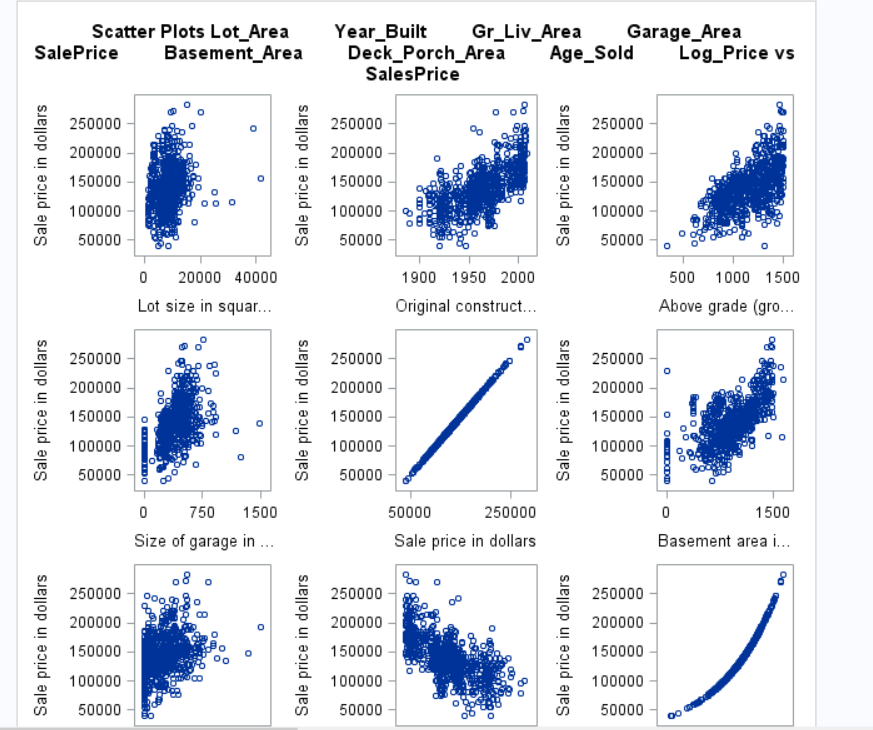




**Null hypothesis**: The mean Sale price for homes with and without masonry\_veneer is equal. **Alternate hypothesis:** The mean Sale price for homes with and without masonry\_veneer is not equal.

Based on the the p-value for the Equality of Variances (0.5060), which is greater than the alpha level of 0.05, we would not reject the null hypothesis. This conclusion supports the assumption of equal variance (the null hypothesis being tested here). That is, there is not enough strong evidence to say conclusively that the mean sale price of homes with masonry\_veneer is different from those without.

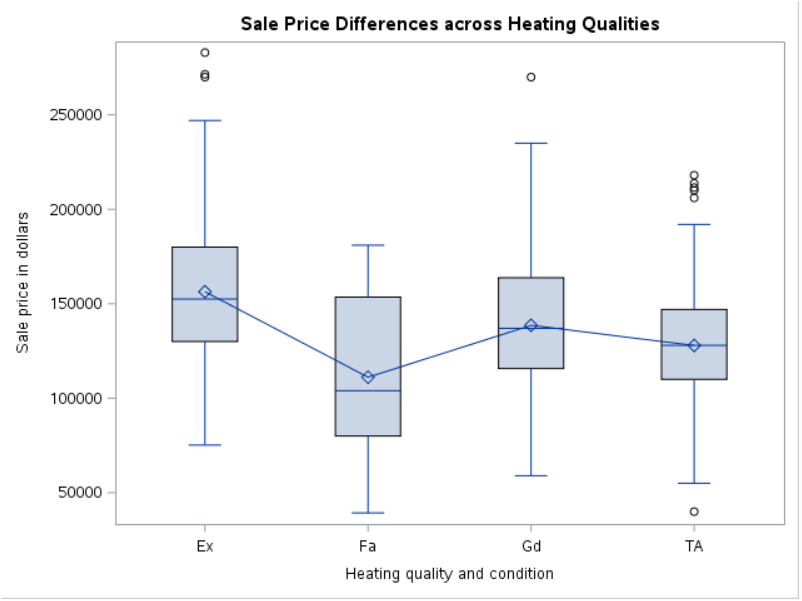
**Problem 5**

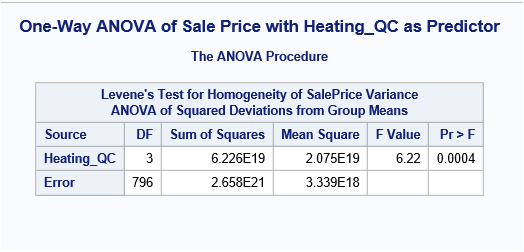
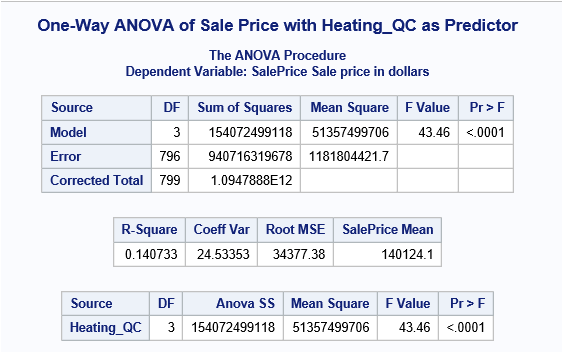


Macro has been created to create Box plots for all Categorical predictors with SalesPrice (Run Macro code from Appendix to get all the Boxplots).

Screen shot of one of the box plots is shown below along with Scatter Plots for Interval variables.

**Problem 6**

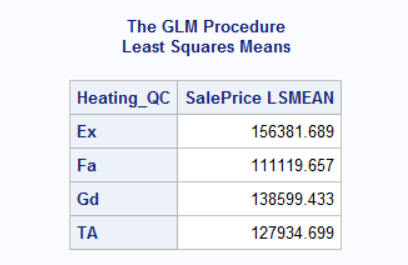
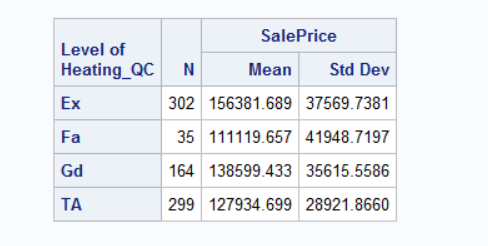




The overall F value from the analysis of variance table is associated with a p value= <0.0001. Presuming that all assumptions of model are valid, we know that at least one heating quality is different from other heating types.

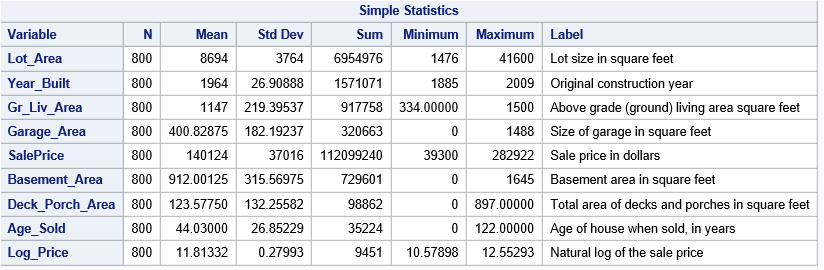
The Levene's Test for Homogeneity of Variance shows a p-value(0.0004) lesser than alpha(0.05). Therefore, we reject hypothesis of homogeneity of variances (equal variances across Heating Quality types). We can conclusively say that the mean sale price is different amongst various heating quality and conditions.

**Problem 7**



Accordingly, for the unbalanced two-way design, there is no discrepancy reflected in the SalesPrice arithmetic means and SalesPrice LS-means for any of the category of Heating\_QC.

**Problem 8**

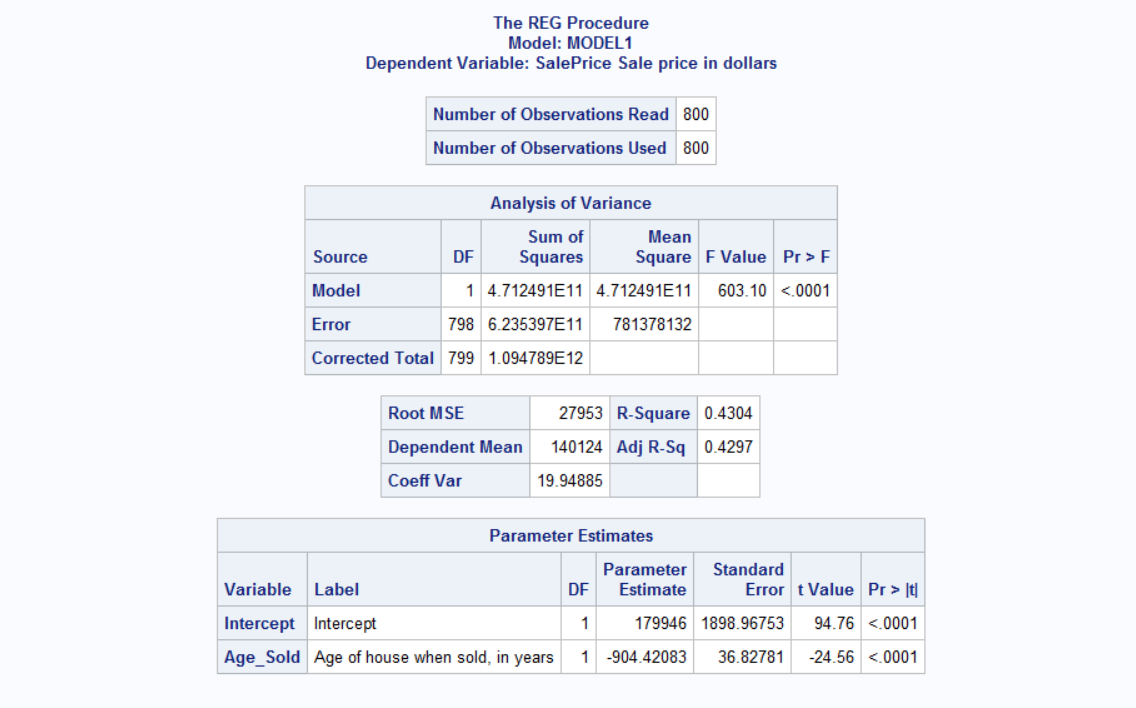




The corr procedure shows the relationship of all the continuous variables with sale price. The results show that the relation is positive for all the variables except for age of the house when sold which has

negative correlation. Keeping 60% as a threshold, it can be observed that lot area, garage area and deck porch area have lesser correlation and the rest have stronger correlation

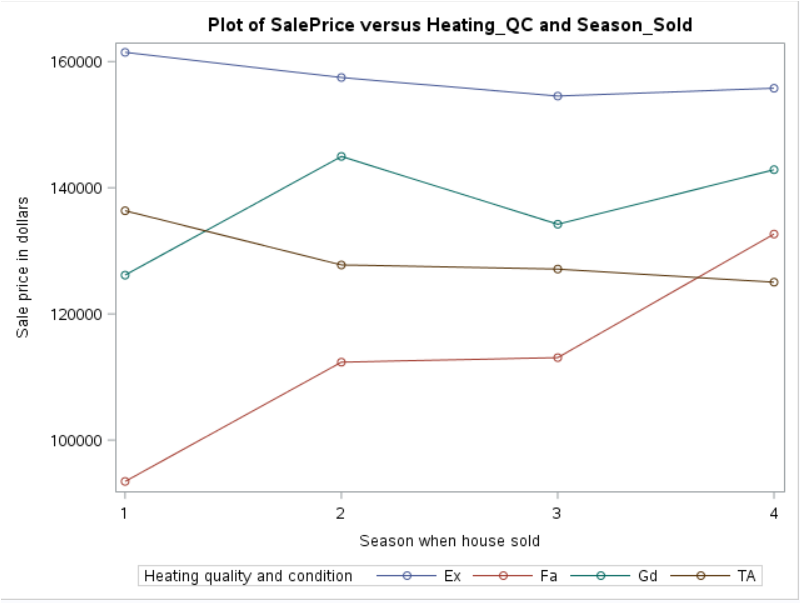
**Problem 9**



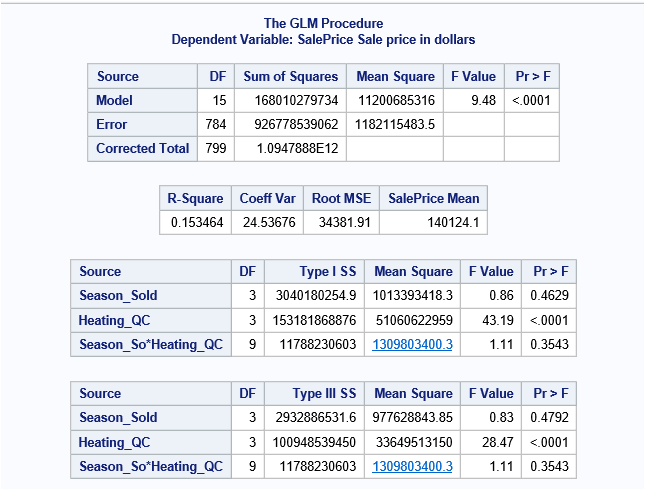
**Model Equations** : SalePrice=179946 -904\*Age\_Sold

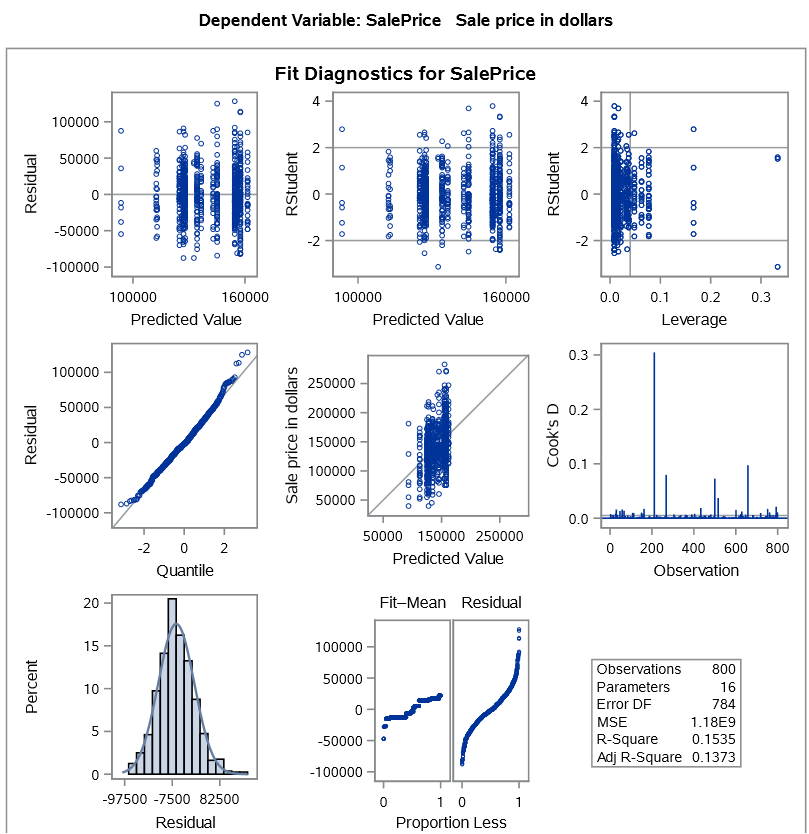
**Age\_sold** has highest correlation with SalePrice that’s why it being used as a most significant predictor

**Problem 10**



It appears that heating quality affects the sale price of homes. However, the effect is not consistent across the seasons sold. The price of the Fa heating type gradually increases over seasons, for Ex and TA types gradually decreases over seasons. The Gd type, the price increases and decreases consistently.



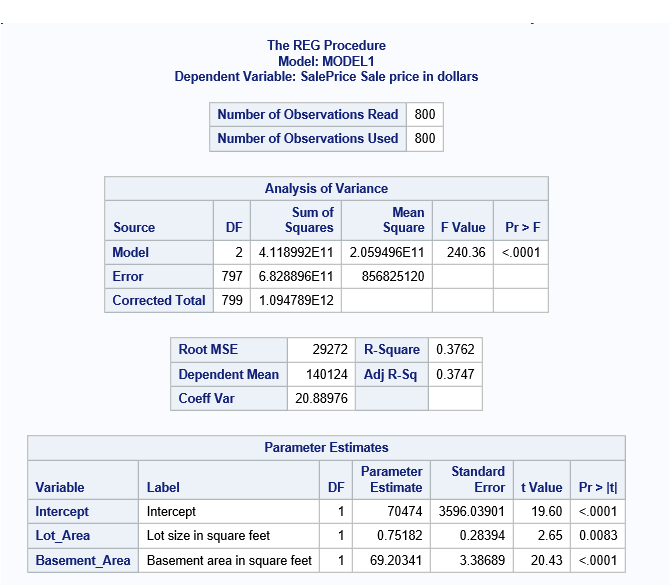
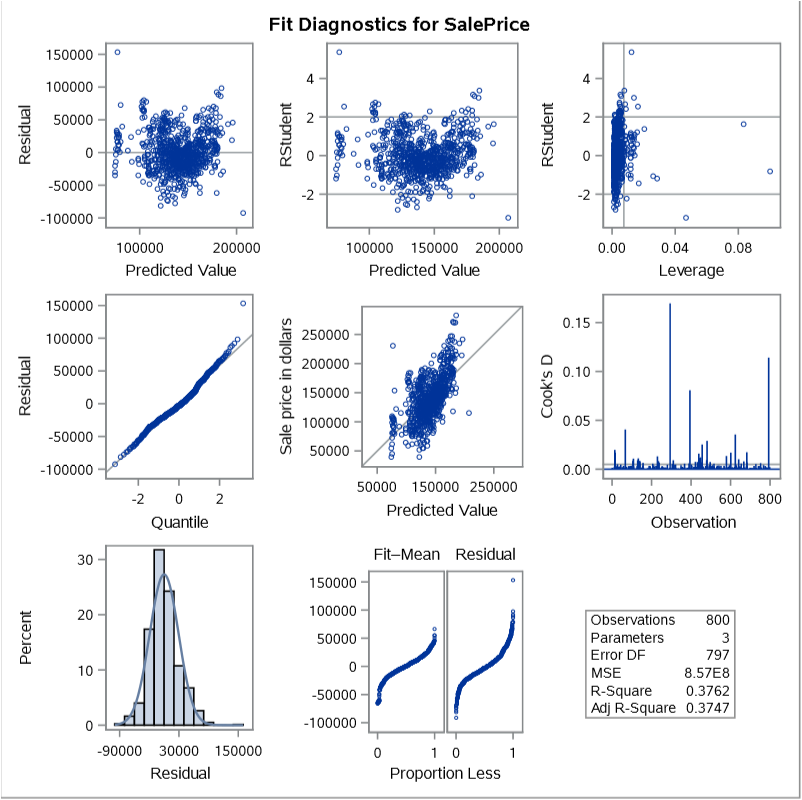


Across all sum of squares types, it can be observed that the sale price is significantly different across all heating\_QC types whereas among the seasons sold and heating\_QC by season sold, the prices are not statistically significant.

**Problem 11**

We adjusted P- square value because it is a sample data not a population dataset , so statistics obtained are not the same as that of population

**Problem 12**

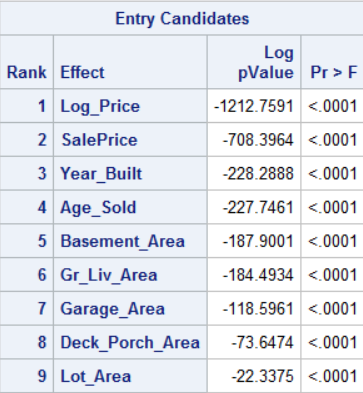
 **12\*/**

The reg model results show that both lot area and basement area are significant variables.

The equation for sales price is , SalePrice = 70474 + 0.75182 Lot\_Area + 69.20341 Basement\_Area

**Problem 13**

Running GLSELECT for all the interval variables and keeping SELECT option has SL with significance level of 0.05 following are the variables selected for the model



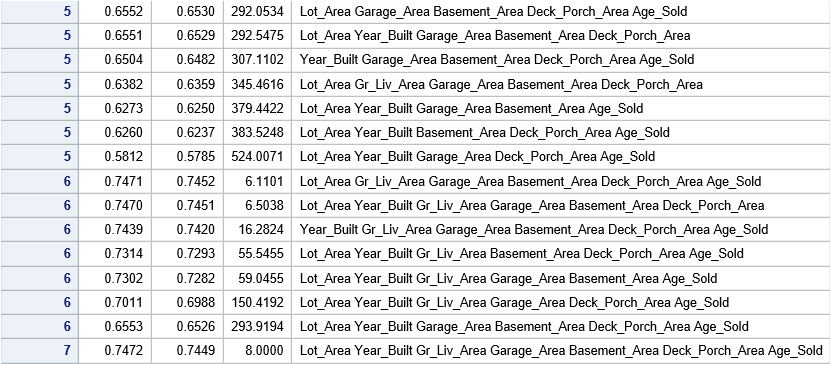
As clear from the below table that there is no significant difference , when changing selection method for GLSELECT , same nine variables are kept in the final model .

Variables remain in the model are : Log\_Price, Year\_Built,Age\_Sold,Basement\_Area, Gr\_Liv\_Area, Garage\_Area, Deck\_Porch\_Area, lot\_Area.

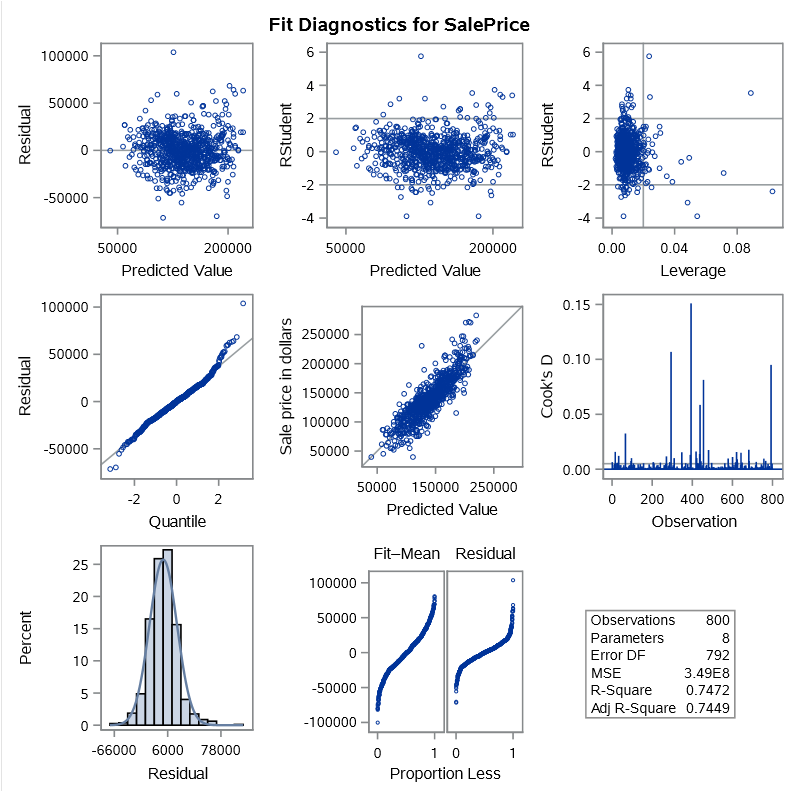
Any one of the selection criteria can be used except BIC , no variable are selected for the model , it is independent of all the predictors

|  |  |  |
| --- | --- | --- |
| **Select** | **Variables** | **Outcome** |
| AIC |  | Same as SL |
| BIC | No variables |  |
| AICC |  | Same as SL |
| SBC |  | Same as SL |

**Problem 14**







The above highlighted output snippet shows the best regression model. This model has the highest percentage of variability (74.52%) of dependent variable that is explained by the variation of the independent variables. The complexity parameter(cp) value of this model is 6.1101 and comprises of 6 of the 7 independent variables. The predictor variables of this model are Lot\_Area, Gr\_Liv\_Area, Garage\_Area,, Basement\_Area, Deck\_Porch\_Area and Age\_Sold.

**APPENDIX**

Libname project2 "C:\1\_Jasbir\UIC\_Local\Spring 2017\SAS\SAS\_Lab\Project2";

**data** project2.housing;

set project2.team24;

**run**;

/\*Problem 1\*/

%let categorical = House\_Style

Heating\_QC

Central\_Air

Bedroom\_AbvGr

Fireplaces

Mo\_Sold

Yr\_Sold

Full\_Bathroom

Half\_Bathroom

Total\_Bathroom

Season\_Sold

Garage\_Type\_2

Foundation\_2

Masonry\_Veneer

Lot\_Shape\_2

House\_Style2

Overall\_Qual

Overall\_Qual2

Overall\_Cond

Overall\_Cond2 ;

%let interval = Lot\_Area

Year\_Built

Gr\_Liv\_Area

Garage\_Area

SalePrice

Basement\_Area

Deck\_Porch\_Area

Age\_Sold

Log\_Price;

%let interval\_new = Lot\_Area

Year\_Built

Gr\_Liv\_Area

Garage\_Area

Basement\_Area

Deck\_Porch\_Area

Age\_Sold;

**proc** **print** data = project2.housing;

title 'List of attributes involved in assessing home values';

**run**;

/\*Problem 2\*/

**proc** **univariate** data=project2.housing;

var &interval;

histogram;

title 'Univariate analysis of continuos variables';

**run**;

**proc** **freq** data = project2.housing;

tables &categorical/plots = freq plot ;

title 'Frequency measures of categorical variables';

**run**;

/\*Problem 3\*/

/\*SalesPrice Distribution Analysis\*/

ods graphics;

ods select histogram;

**proc** **univariate** data=project2.housing noprint;

var salePrice;

histogram salePrice / normal kernel;

inset n mean std / position=ne;

title "SalesPrice Distribution Analysis";

**run**;

/\* T-Test to check mean of SalePrice is $135,000 in the data set\*/

/\*We reject the null Hypothesis , Mean is not equal to 135000 at alpha value of 0.05\*/

**proc** **ttest** data=project2.housing H0=**135000** ALPHA=**0.05**

plots(only shownull)=interval;

var salePrice ;

title 'Testing Whether the Mean Sales Price=135000';

**run**;

title;

/\*Problem 4\*/

**proc** **ttest** data=project2.housing plots(shownull)=interval;

class Masonry\_Veneer;

var SalePrice;

title "Comparing mean Sales Price for homes with and without Masonry Veneer";

**run**;

/\*

\* Null hypothesis: The mean Sale price for homes with and without masonry\_veneer is same

\* Alternate hypothesis: The mean Sale price for homes with and without masonry\_veneer is not same

\* Because the p-value for the Equality of Variances test is greater than the alpha level of 0.05, we would not reject the null hypothesis.

This conclusion supports the assumption of equal variance (the null hypothesis being tested here).

That is, there is not enough strong evidence to say conclusively that the mean sale price of homes with masonary

veneer is different from those without \*/

/\*Problem 5\*/

/\*Scatter Plote between all continous variables and SalePrice\*/

**proc** **sgscatter** data=project2.housing;

plot salePrice \* (&interval);

title "Scatter Plots &interval vs SalesPrice";

**run**;

/\* Box Plot for All categorocal variables against SalePrice\*/

**%macro** ***BoxPlot***;

%local i next\_name;

%do i=**1** %to %sysfunc(countw(&categorical));

%let next\_name = %scan(&categorical, &i);

proc sgplot data=project2.housing;

vbox SalePrice/ category=&next\_name

connect=mean;

title "Box Plot of Sale Price W.R.T &next\_name";

run;

%end;

**%mend** BoxPlot;

/\*Run Macro Variable for Box Plot \*/

%***BoxPlot***

/\*Problem 6\*/

**proc** **sgplot** data=project2.housing;

vbox SalePrice / category=Heating\_QC connect=mean;

title "Sale Price Differences across Heating Qualities";

**run**;

ods graphics;

**proc** **anova** data=project2.housing ;

class Heating\_QC;

model SalePrice=Heating\_QC;

means Heating\_QC / hovtest=levene ;

title "One-Way ANOVA of Sale Price with Heating\_QC as Predictor";

**run**;

/\*The overall F value from the analysis of variance table is associated with a p value= <.0001. Presuming that all

assumptions of model are valid, we know that at least one heating Quality is different from other heating types.

The Levene's Test for Homogeneity of Variance shows a p-value lesser than alpha(0.05).

Therefore, we reject hypothesis of homogenity of variances (equal variances across Heating Quality types) \*/

/\*Problem 7 \*/

%Let categorical\_1 =Heating\_QC ;

**Proc** **GLM** data=project2.housing ;

class &categorical\_1;

model SalePrice=&categorical;

means &categorical\_1;

lsmeans &categorical;

title "Two way Anova Heating\_QC as Predictor";

**run**;

/\*Problem 8\*/

**PROC** **CORR** DATA=project2.housing ;

var SalePrice;

WITH &interval;

**RUN**;

/\*The corr procedure shows the relationship of all the continuous ariables with sale price. The results show that the

relation is positive for all the variables except for age of the house when sold which has

negative correlation. Keeping 60% as a threshold, it can be observed that lot area, garage area and deck porch area

have lesser correlation and the rest have stronger correlation \*/

/\*Problem 9\*/

/\*Linear Regression Model\*/

/\*Model Equations : SalePrice=179946 -904\*Age\_Sold \*/

/\*Age\_sold has highest correlation with SalePrice\*/

%Let interval = Age\_sold;

**proc** **reg** data=project2.housing ;

model SalePrice=&interval;

title "Linear Regreession SalesPrice & Age\_Sold";

**run**;

/\*Problem 10\*/

**proc** **sgplot** data=project2.housing;

title 'Plot of SalePrice versus Heating\_QC and Season\_Sold ';

vline Season\_Sold / group=Heating\_QC

stat=mean

response=SalePrice

markers;

**run**;

/\*It appears that heating quality affects the sale price of homes. However the effect is not consistent across

the seasons sold.

The price of the Fa heating type gradually increases over seasons , for Ex and TA types gradually decreases

over seasons. The Gd type , the price increases and decreases consistently. \*/

/\* Problem 11\*/

/\*Two way Anova\*/

**proc** **glm** data=project2.housing plots(only)=intplot;

class Heating\_qc season\_sold;

model SalePrice=Heating\_qc|season\_sold;

lsmeans season\_sold\*Heating\_qc / slice=Heating\_qc;

store team24;

title "Two way Anova";

**run**;

**proc** **plm** restore= team24 plots=all;

slice season\_sold\*Heating\_qc / sliceby=Heating\_qc adjust=tukey;

effectplot interaction(sliceby= Heating\_qc) / clm;

**run**;

**proc** **glm** data=project2.housing plots(only)=diagnostics;

class Season\_Sold Heating\_QC;

model SalePrice=Season\_Sold|Heating\_QC;

**run**;

**quit**;

/\*Across all sum of squares types, it can be observed that the sale price is significally different

across all heating\_QC types where as among the season sold and heating\_QC by season sold are not statistically

significant. \*/

/\*Problem 12\*/

**proc** **reg** data= project2.housing;

model SalePrice = Lot\_Area Basement\_Area;

TITLE 'Regression model of Sales Price with Lot Area and Basement Area as predictor variables';

**run**;

**quit**;

/\*The reg model results show that both lot area and basement area are significant variables.

The equantion for sales price is , SalePrice = 70474 + 0.75182 Lot\_Area + 69.20341 Basement\_Area \*/

/\*Problem 13\*/

**%macro** GlmSelect(type);

proc glmselect data=project2.housing plots=all;

model SalePrice= &interval / SELECTION=STEPWISE DETAILS=STEPS

SELECT=&type

SLSTAY= **0.05**;

title ''&type' selection';

run;

quit;

**%mend** modsel;

\*SL, AIC, BIC, AICC, and SBC ;

%***GlmSelect***(Sl);

%***GlmSelect***(AIC);

%***GlmSelect***(BIC);

%***GlmSelect***(AICC);

%***GlmSelect***(SBC);

/\*Prroblem 14\*/

%let options = rsquare adjrsq cp ;

**proc** **reg** data=project2.housing;

model SalePrice=&interval\_new/ P R

selection = &options;

/\*plot RESIDUAL. \* &interval\_new; \*/

**run**;

/\*6

0.7471

0.7452

6.1101

Lot\_Area Gr\_Liv\_Area Garage\_Area Basement\_Area Deck\_Porch\_Area Age\_Sold\*/