IDS462 – Statistical Software in Business

Project 2

Fall 2017

Name: Anjali Sodani

UIN:663678958

Introduction:

The original "Ames Home Sales" dataset contains information about the sale of individual residential property in Ames, Iowa, from 2006 to 2010 including 2,930 observations and 98 explanatory variables involved in assessing home values. In addition, new variables such as the natural log of the sale price of the homes are created.

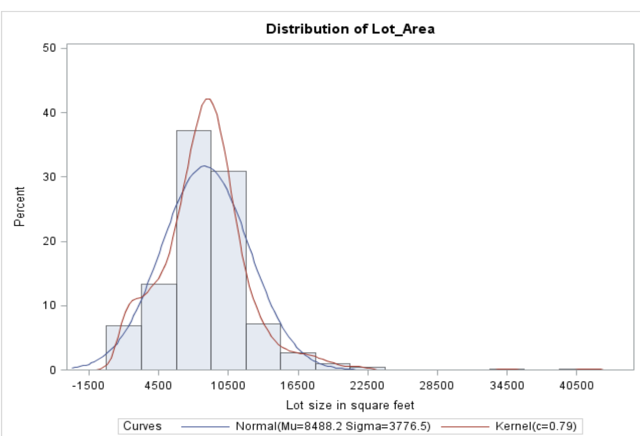
1. Use %LET statements to name the macro variables and set their values. The macro variables are referred to in the SAS code as &categorical and &interval, to distinguish those names from those of variables.

We created two macro variables called categorical for the categorical data and interval for the numerical data using %LET statement. We will run the proc contents and proc means to see if there are any missing values and see the type of variable whether it is numeric or character. Though, just separating the variables based on type is not the correct way so we will plot the graph and see if it is categorical or continuous. We will drop the variable PID as it is only an index and should not be used for calculation purpose, SalePrice will be removed from the interval macro for later questions as we will not want to use that for calculations. Based on our interpretation, we will form the following macros:

%let categorical =House_Style Overall_Qual Heating_QC Overall_Cond Central_Air Garage_Type_2 Foundation_2 Masonry_Veneer Lot_Shape_2 House_Style2 Bedroom_AbvGr Fireplaces Mo_Sold Yr_Sold Overall_Qual2 Overall_Cond2 Full_Bathroom Total_Bathroom
Half_Bathroom Season_Sold Bonus;

%let interval = /*SalePrice*/ Lot_Area
Gr_Liv_Area Garage_Area Basement_Area Deck_Porch_Area Age_Sold Log_Price Year_Built ;

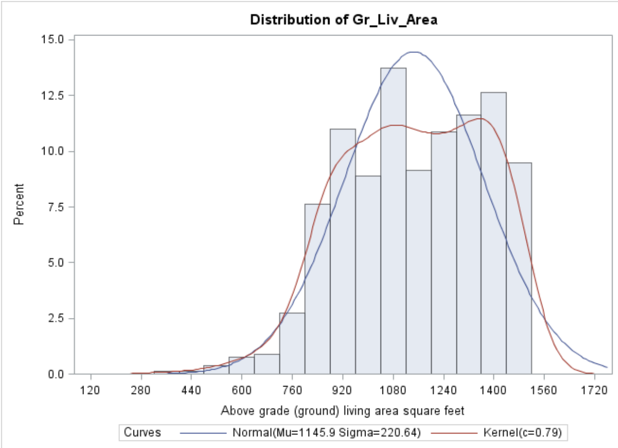
2.Use PROC UNIVARIATE to generate plots and descriptive statistics for continuous variables and PROC FREQ to generate plots and tables for categorical variables.



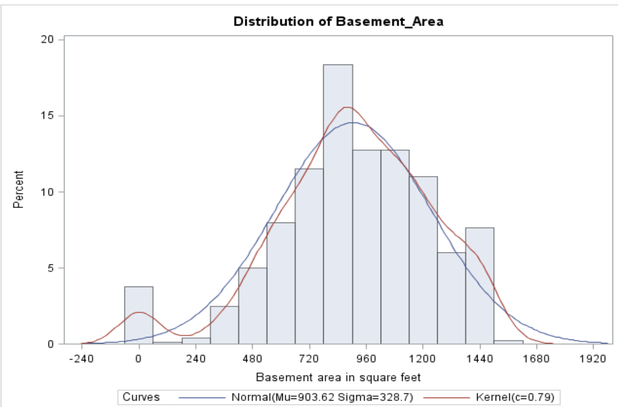
Continuous Variables:

The average lot area is 8488.2 square feet with a standard deviation of 3776.5 square feet.

The distribution is left-skewed indicating that most properties have a lot size lower than the average.

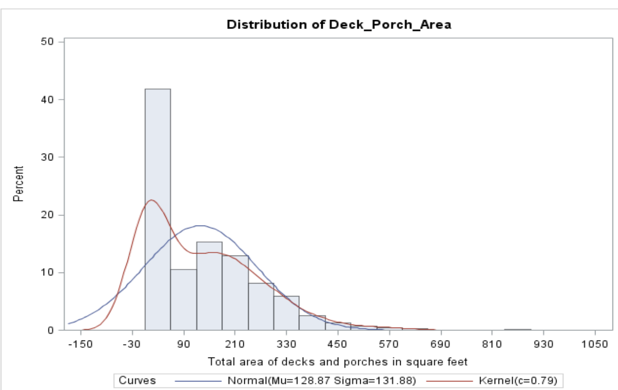


The average grade (ground) living area is 1145.9 square feet. The distribution appears bi-modal with peaks at 1080 and 1450 square feet.

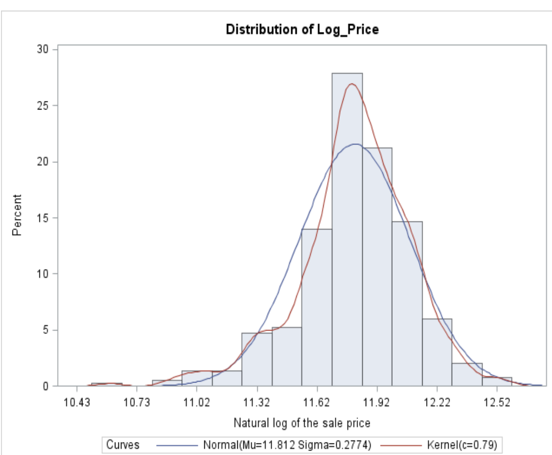


The distribution of basement area is nearly normal with a mean of 903.62 square feet.

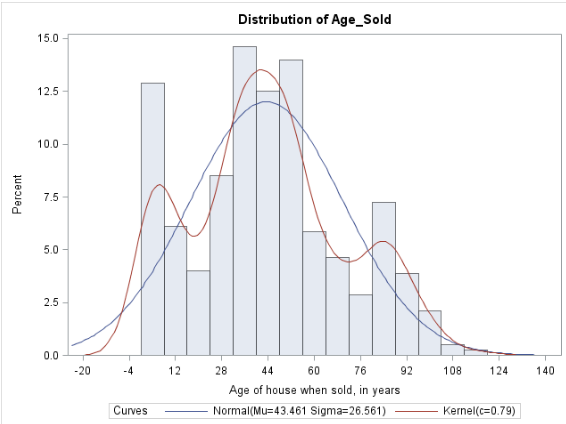
There is a dip between 0 and 240 square feet indicating that if a house does come with a basement, it is quite likely that the basement would be at least 240 square feet.



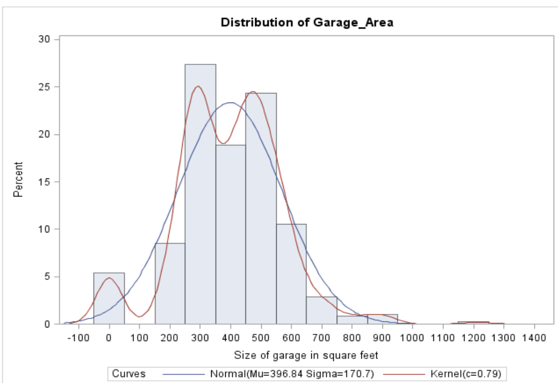
The porch area is skewed to the left, which is not very surprising.



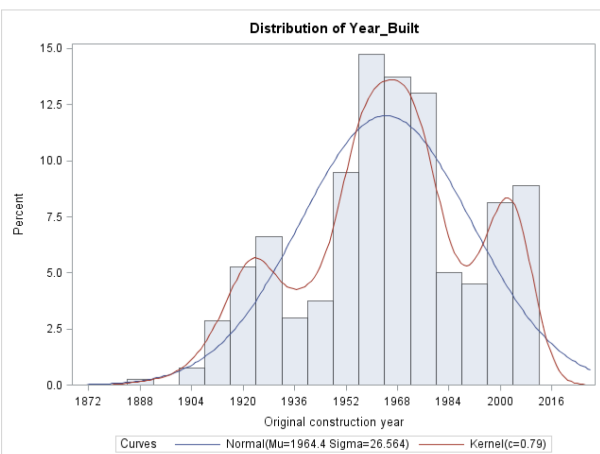
The rate of change of housing prices is nearly normally distributed. Since, it is a natural log of price itself, the normal distribution looks very accurate.



The age of properties at the time of sales appears multi-modal. The mean age is 43.46 years.



We observe that the garage area allocated to various properties starts at about 150 square feet. The mean value is 396.84 square feet. The distribution is bi-modal.



The original construction year is multi-modal. This is consistent with our observations of property ages at the time of sales.

Categorical Variables:

Style of dwelling				
House_Style	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
1.5Fin	81	10.13	81	10.13
1.5Unf	6	0.75	87	10.88
1Story	508	63.50	595	74.38
2.5Unf	1	0.13	596	74.50
2Story	116	14.50	712	89.00
SFoyer	37	4.63	749	93.63
SLvl	51	6.38	800	100.00

Overall material and finish of the house				
Overall_Qual	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
1	1	0.13	1	0.13
2	7	0.88	8	1.00
3	15	1.88	23	2.88
4	86	10.75	109	13.63
5	331	41.38	440	55.00
6	225	28.13	665	83.13
7	100	12.50	765	95.63
8	33	4.13	798	99.75
9	2	0.25	800	100.00

Heating quality and condition				
Heating_QC	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Ex	279	34.88	279	34.88
Fa	31	3.88	310	38.75
Gd	167	20.88	477	59.63
TA	323	40.38	800	100.00

Overall condition of the house				
Overall_Cond	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
1	1	0.13	1	0.13
2	2	0.25	3	0.38
3	17	2.13	20	2.50
4	29	3.63	49	6.13
5	353	44.13	402	50.25
6	177	22.13	579	72.38
7	151	18.88	730	91.25
8	63	7.88	793	99.13
9	7	0.88	800	100.00

Presence of central air conditioning				
Central_Air	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
N	63	7.88	63	7.88
Y	737	92.13	800	100.00

Garage attached or detached				
Garage_Type_2	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Attached	463	58.68	463	58.68
Detached	283	35.87	746	94.55
NA	43	5.45	789	100.00
Frequency Missing = 11				

Foundation Type				
Foundation_2	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Brick/Tile/Stone	97	12.13	97	12.13
Cinder Block	456	57.00	553	69.13
Concrete/Slab	247	30.88	800	100.00

Masonry veneer or not				
Masonry_Veneer	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
N	565	70.80	565	70.80

Masonry veneer or not				
Masonry_Veneer	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Y	233	29.20	798	100.00
Frequency Missing = 2				

Regular or irregular lot shape				
Lot_Shape_2	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Irregular	236	29.57	236	29.57
Regular	562	70.43	798	100.00
Frequency Missing = 2				

Style of dwelling				
House_Style2	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
1.5Fin	81	10.13	81	10.13
1Story	514	64.25	595	74.38
2Story	117	14.63	712	89.00
SFoyer	37	4.63	749	93.63
SLvl	51	6.38	800	100.00

Bedrooms above grade				
Bedroom_AbvGr	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	1	0.13	1	0.13
1	44	5.50	45	5.63
2	279	34.88	324	40.50
3	464	58.00	788	98.50
4	12	1.50	800	100.00

Number of fireplaces				
Fireplaces	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	514	64.25	514	64.25
1	251	31.38	765	95.63
2	34	4.25	799	99.88
3	1	0.13	800	100.00

Month Sold (MM)				
Mo_Sold	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
1	25	3.13	25	3.13
2	35	4.38	60	7.50
3	68	8.50	128	16.00
4	90	11.25	218	27.25
5	131	16.38	349	43.63
6	138	17.25	487	60.88
7	124	15.50	611	76.38
8	65	8.13	676	84.50
9	34	4.25	710	88.75
10	39	4.88	749	93.63
11	31	3.88	780	97.50
12	20	2.50	800	100.00

Year Sold (YYYY)				
Yr_Sold	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
2006	150	18.75	150	18.75
2007	191	23.88	341	42.63
2008	171	21.38	512	64.00
2009	172	21.50	684	85.50
2010	116	14.50	800	100.00

Overall material and finish of the house				
Overall_Qual2	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
4	109	13.63	109	13.63
5	331	41.38	440	55.00
6	360	45.00	800	100.00

Overall condition of the house				
Overall_Cond2	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
4	49	6.13	49	6.13
5	353	44.13	402	50.25
6	398	49.75	800	100.00

Number of full bathrooms				
Full_Bathroom	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
1	322	40.25	322	40.25
2	401	50.13	723	90.38
3	76	9.50	799	99.88
4	1	0.13	800	100.00

Number of half bathrooms				
Half_Bathroom	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	572	71.50	572	71.50
1	222	27.75	794	99.25
2	6	0.75	800	100.00

Season when house sold				
Season_Sold	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
1	80	10.00	80	10.00
2	289	36.13	369	46.13
3	327	40.88	696	87.00
4	104	13.00	800	100.00

Sale Price > $175,000				
Bonus	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	673	84.13	673	84.13
1	127	15.88	800	100.00

3. Use the TTEST procedure to test whether the mean of SalePrice is $135,000 in the data set. Is the mean value in the sample statistically significantly different from $135,000 at an alpha level of 0.05?

- On performing the one sample T-test, we obtain a t-statistic = 3.71.The resulting p-value is less than 0.05.
- Thus, we reject the null hypothesis. We can conclude that the mean sales price is not equal to $135,000. It could be less than or greater than this value, as we have conducted a two-tail test.

4. Use the TTEST procedure to test whether the mean of SalePrice is the same for homes with masonry veneer and those without. Provide your insights.

- From the F-fold Test, we fail to reject the null hypothesis as the f value is 0.8387 which is greater than the alpha. Thus, we can conclude that both groups (homes with masonry veneer and those without) have equal variance.
- The t-statistic associated with pooled standard deviation is -8.72 and the p-value is much lower than our significance level (0.05). Thus, we can conclude that the mean of sales price is different for homes with and without masonry veneer. The mean sales prices of houses with masonry veneer is 156687 and the mean sales prices of houses without masonry veneer is 132759

5. Create scatter plots to show relationships between continuous predictors and SalePrice and comparative box plots to show relationships between categorical predictors and SalePrice.

- We created scatter plot for all the continuous variables against the SalePrice and see some strong correlation between the SalePrice and variables like year of construction, basement area, grade living area etc.
- We created the Box Plots for all the categorical variables against the Saleprice by creating a new macro function called Boxplot and a new local variable called 'y' and 'variable' and it is used to create a loop to run the report automatically for generating the boxplots for all the categorical variables.

6. Run an analysis of variance with SalePrice as the response variable and Heating_QC as the categorical predictor variable. Output diagnostic plots and look at Levene's test of homogeneity of variances.

First, we calculate the descriptive statistics of the SalePrice based on heating quality and condition. We observe that the residuals are normally distributed. The overall F-statistic (from the analysis of variance table) is associated with a p value which is significantly lower than our threshold value of 0.05.

Provided our model meets all the assumptions of ANOVA, we conclude that at least one Heating_QC group has a mean sales price which is significantly different from other groups.

7. Use the LSMEANS statement in PROC GLM to produce comparison information about the mean sale prices of the different heating system quality ratings.

The GLM Procedure
Least Squares Means

Heating_QC	SalePrice LSMEAN
Ex	139834.394
Fa	139834.394
Gd	139834.394
TA	139834.394

8. Examine the relationships between SalePrice and the continuous predictor variables in the data set. Use the CORR procedure.

- Since PROC CORR only shows the relationship of the first 10 variables in the scatter plot, we use best=4 to show top 4 relationships with the sales price.
- We observe a strong positive correlation of sales price with year of construction, basement area and grade living area.
- Since year of construction and age of property are inversely proportional, sales price has a strong negative correlation with property age. These correlations indicate the presence of a linear relationship of sales price with these variables.

9. Perform a simple linear regression analysis with SalePrice as the response variable, and one of the significant predictors. Explain why you have chosen that variable. What's the prediction equation

From the correlation matrix, we see the correlation values of sales price against the different variables and we find that the correlation value for the log_price is the highest. This is obvious since log_price is natural log of sales price, so it is directly related. Therefore, we use Overall_Qual because it's p value<0.005 and correlation is 0.72272 the value of f statistics is 872.58(p<0.001), therefore null hypothesis is rejected;

The predicted equation is:

Prediction Equation:

SalePrice = 8305.62015 +23963*Overall_Qual

10. Perform a two-way ANOVA of SalePrice with Heating_QC and Season_Sold as predictor variables. Before conducting an analysis of variance, you should explore the data. To further explore the numerous treatments, examine the means graphically. You can use the GLM procedure to discover the effects of both Season_Sold and Heating_QC.

From the output it appears that the season_sold affects the sales price, but the heating_QC doesn't show the same kind of pattern across all levels.

For Fair and Good quality, the sales price shows significant change but for the other two categories (i.e. Ex, Ta) there is not significant change in the response variable.

a. The global F test indicates the difference across different groups because there is an interaction in the model for all the possible combinations of the Season_sold*Heating_QC against all other combinations.

b. The value of R-square implies that 19% variation in the sales price is explained by the explanatory variables.

c. The sliced table shows the effects on sales price at each level. The effect is significant for Fa and Gd.

11. Perform a two-way ANOVA of SalePrice with Heating_QC and Season_Sold as predictor variables. Include the interaction between the two explanatory variables. Store the output to a dataset and adjust p-values using PROC PLM (explain why you would need to do that).

The GLM Procedure
Least Squares Means

Season_So*Heating_QC Effect Sliced by Heating_QC for SalePrice					
Heating_QC	DF	Sum of Squares	Mean Square	F Value	Pr > F
Ex	3	2779199876	926399959	0.83	0.4771
Fa	3	11109778636	3703259545	3.32	0.0194
Gd	3	10261200946	3420400315	3.07	0.0273
TA	3	2570969564	856989855	0.77	0.5118

The global F test indicates the significnt different in different groups because there is interaction in the model this tests for all the possible combinations of the Season_sold*Heating_QC against all other combinations

With the help of PROC PLM, we can produce a source item store. The use of item stores and PROC PLM enables us to separate common postprocessing tasks, such as predicting new set of values with the existing model or testing for the treatment differences. A numerically expensive model fitting technique can be applied once to produce a source item store. The PLM procedure can then be called multiple times and the results of the fitted model analyzed without incurring the model fitting expenditure again.

Since, in our case we need to adjust p -values we will not have to build the model again and again and instead we can produce source item store and do the analysis by calling PROC PLM multiple times.

12. Perform a regression model of SalePrice with Lot_Area and Basement_Area as predictor variables.

Prediction Equation:
SalePrice = 70892+1.07517*Lot_Area +66.19581*Basement_Area

The p-value is less than 0.05, therefore the explanatory variables are significant

13. Write a macro to invoke PROC GLMSELECT five times on the SalePrice variable regressing on the interval variables. For each, request STEPWISE selection with the SELECTION= option and include DETAILS=STEPS to obtain step information and the selection summary table. Use 0.05 as the significance level for entry into and staying in the model. Call to macro to run SELECT for the options SL, AIC, BIC, AICC, and SBC and compare the selected models from the output. Does the significance level for entry into and staying in the model have any impact when you use options other

than SL? Which variables stay in the model for each 5 options? Which selection methods and criteria would you recommend?

14. Invoke PROC REG with the plots option using rsquare adjrsq cp to produce a regression of SalePrice on all the other interval variables in the data set. Use selection = rsquare adjrsq cp. Which model you would suggest, and why? Hint: compare the options rsquare adjrsq cp.

APPENDIX

```
libname project2 "C:\Users\smisri2\Documents\My SAS Files\9.4\project_2";

data project2.dataset;
 set project2.team14;
run;

proc contents data = project2.dataset varnum;
   title "Contents of the given file";
run;

proc means data = project2.dataset ;
   title" Summary of the given file";
run;
proc means data = project2.dataset n nmiss;
   title" To see if there are any missing value";
run;

proc format;
```

```
 value $missfmt ' '='Missing' other='Not Missing';
 value  missfmt  . ='Missing' other='Not Missing';
run;

proc freq data=project2.dataset;
format _CHAR_ $missfmt.; /* apply format for the duration of this PROC */
tables _CHAR_ / missing missprint nocum nopercent;
format _NUMERIC_ missfmt.;
tables _NUMERIC_ / missing missprint nocum nopercent;
run;

/*Problem #1*/
%let interval = /*SalePrice*/
Lot_Area
Gr_Liv_Area
Garage_Area
Basement_Area
Total_Bathroom
Deck_Porch_Area
Age_Sold
Log_Price
Year_Built
;

%let categorical =House_Style
Overall_Qual
Heating_QC
Overall_Cond
Central_Air
Garage_Type_2
Foundation_2
Masonry_Veneer
Lot_Shape_2
House_Style2
Bedroom_AbvGr
Fireplaces
Mo_Sold
Yr_Sold
Overall_Qual2
Overall_Cond2
Full_Bathroom
Half_Bathroom
Season_Sold
Bonus
;

/*Problem #2*/
title 'Exploratory data analysis of continuous variables';
ods graphics on;
proc univariate data= project2.dataset noprint;
        histogram &interval/ normal kernel;
        var &interval;
run;
ods graphics off;

/*To check the variables by their plots to see which are categorical and which are
```

```
numeric*/
proc univariate data =project2.dataset plot;

title 'Exploratory data analysis of categorical variables';
Proc FREQ Data = project2.dataset;
tables &categorical / plots=freqplot(type=bar);
run;

/*Problem #3*/
proc ttest data=project2.dataset h0=135000
       plots(only shownull)= interval;
  var SalePrice;
  title 'Testing Whether the Mean SalePrice 135000';
run;

/*Problem #4*/
proc ttest data=project2.dataset
plots(shownull)=interval;
  class Masonry_Veneer    ;
  var SalePrice;
  title "SalePrice the same for homes , Comparing masonry veneer or not";
run;
title;

/*Problem #5*/
*scatter plots to show relationships between
continuous predictors and SalePrice;

ods graphics on;
PROC sgscatter  DATA=project2.dataset;
  PLOT SalePrice*(&interval) ;
  RUN;

*To plot vbox for categorical variables;

%macro Boxplot(index);
%local y variable;
%do y=1 %to %sysfunc(countw(&&index));
  %let variable = %scan(&&index, &y);
  proc sgplot data=project2.dataset;
  vbox SalePrice/ category=&variable
                 connect=mean;
  title "BoxPlot of Sale Price with &variable";
run;
%end;
%mend Boxplot;

%Boxplot(&categorical);

/*Problem #6*/
* First we will see the SalePrice based on heating quality and condition;
proc means data= project2.dataset;
  var SalePrice;
  class Heating_QC ;
  title 'Descriptive Statistics of SalePrice by Heating_QC ';
run;
```

```
proc sgplot data=project2.dataset;
    vbox SalePrice / category=Heating_QC connect=mean;
    title "SalePrice Differences across Heating_QC";
run;

ods graphics;
proc  glm  data=project2.dataset plots (only)=diagnostics;
   class Heating_QC;
   model SalePrice=Heating_QC;
 means Heating_QC /  hovtest=levene;

   title "One-Way ANOVA with Heating_QC as Predictor";
run;
quit;

/*Problem #6 continued*/
proc  glm  data=project2.dataset plots (only)=diagnostics;
   class Heating_QC;
   model SalePrice=Heating_QC;
   means Heating_QC /  hovtest=levene;
   title "One-Way ANOVA with Heating_QC as Predictor";
run;
quit;

/*Problem #7*/
proc glm data=project2.dataset plots (only)=intplot;
  class Heating_QC;
  model SalePrice =Heating_QC Heating_QC*SalePrice ;
  lsmeans Heating_QC;
run;
quit;

/*Problem #8*/
proc corr data =project2.dataset
 plots(only)=scatter(nvar=all ellipse=none)
 nosimple best =4 ;
   var &interval;
   with SalePrice;
   title "Correlations with SalePrice";
run;

*Since proc corr only shows the relationship of the first 10 variable in the
scatter plot we use the best=4 to show top 4 relationships with the SalePrice;

/*Problem #9*/
proc corr data=project2.dataset  nomiss
         /*nosimple
         best=5   */
         out=project2.pearson;
   title "Correlations of Predictors";
run;
%let big=0.7;
data project2.bigcorr;
   set project2.pearson;
   array vars{*} &interval;
```

```
  do i=1 to dim(vars);
    if abs(vars{i})<&big then vars{i}=.;
  end;
  if type="CORR";
  drop i type;
run;
proc print data=project2.bigcorr;
  format &interval 5.2;
run;


ods graphics on;
Proc reg data=project2.dataset;
  model SalePrice= Overall_Qual ;
  title "Regression of SalePrice on Overall_Qual";
run;

/*Problem #10*/
*To explore the data ------;
proc corr data=project2.dataset nomiss
        plots=scatter(nvar=all ellipse=none);
  var &interval;
  with SalePrice;
  title "Correlations and Scatter Plots";
run;

proc sgplot data=project2.dataset;
  vline Season_Sold / group=Heating_QC
                stat=mean
                response=SalePrice
                markers;
run;

proc glm data=project2.dataset plots(only)=intplot;
  class Season_Sold Heating_QC;
  model SalePrice=Season_Sold|Heating_QC;
  lsmeans Season_Sold*Heating_QC / slice=Heating_QC;
run;
quit;


/*Problem #11*/
proc glm data=project2.dataset plots(only)=intplot;
  class Season_Sold Heating_QC;
  model SalePrice=Season_Sold|Heating_QC;
  lsmeans Season_Sold*Heating_QC / slice=Heating_QC;
  store storeddata;
run;
quit;
proc plm restore=storeddata plots=all;
  slice Season_sold*Heating_QC/ sliceby=Heating_QC  adjust=tukey;
  effectplot interaction (sliceby= Heating_QC )/ clm;
run;

/*Problem #12*/
/*Regression Model*/
```

```
proc reg data=project2.dataset;
  model SalePrice= Lot_Area Basement_Area;
  title 'Regression of SalePrice on Lot_Area and Basement_Area';
run;

/*Problem #13*/

%macro mod_sel(mod);
proc glmselect data=project2.dataset plots=all;
  model salePrice = &interval / SELECTION= stepwise SELECT=&mod details=steps;
run;
quit;
%mend mod_sel;
Title 'Select= SL with salePrice';
%mod_sel(SL)
Title 'Select= AIC with salePrice';
%mod_sel(AIC)
Title 'Select= BIC with salePrice';
%mod_sel(BIC)
Title 'Select= AICC with salePrice';
%mod_sel(AICC)
Title 'Select= SBC with salePrice';
%mod_sel(SBC)


/*Problem #14*/
ods graphics on;
%macro mod_sel(mod);
Proc reg data=Project2.dataset plots=(&mod);
  model SalePrice= &interval/ SELECTION= &mod;
  title "Regression of SalePrice ";
run;
quit;

%mend mod_sel;
Title 'Regression model with selection=rsquare';
%mod_sel(rsquare);
Title 'Regression model with selection=adjrsq';
%mod_sel(adjrsq);
Title 'Regression model with selection=cp';
%mod_sel(cp)
```