**Project 2**

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.
2. Use PROC UNIVARIATE to generate plots and descriptive statistics for continuous variables and PROC FREQ to generate plots and tables for categorical variables.
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?
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.
5. Create scatter plots to show relationships between continuous predictors and **SalePrice** and comparative box plots to show relationships between categorical predictors and **SalePrice**.

For the comparative box plots to show relationships between categorical predictors and **SalePrice**, write a macro program to run the following procedure for each categorical variable automatically.

**proc** **sgplot** data=dataset\_name;

vbox response\_variable/ category=predictor1

connect=mean;

title "Sale Price Differences across predictor1";

**run**;

1. 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.
2. Use the LSmeans statement in proc glm to produce comparison information about   
   the mean sale prices of the different heating system quality ratings.
3. Examine the relationships between **SalePrice** and the continuous predictor variables in the data set. Use the CORR procedure.
4. 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?
5. Perform a two-way ANOVA of **SalePrice** with **Heating\_Q**C 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**.
6. Perform a two-way ANOVA of **SalePrice** with **Heating\_Q**C 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). Use the following proc step:

**proc** **plm** restore=dataset\_name plots=all;

slice interaction\_term / sliceby=predictor1 (or predictor2) adjust=tukey;

effectplot interaction(sliceby= predictor1 (or predictor2)) / clm;

**run**;

1. Perform a regression model of **SalePrice** with **Lot\_Area** and **Basement\_Area** as predictor variables.
2. 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?
3. 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.
4. Invoke PROC FREQ and create one‑way frequency tables for the variables **Bonus**, **Fireplaces**, and **Lot\_Shape\_2** and create two‑way frequency tables for the variables **Bonus** by **Fireplaces**, and **Bonus** by **Lot\_Shape\_2**. For the continuous variable, **Basement\_Area**, create histograms for each level of **Bonus**. Use a CLASS statement in PROC UNIVARIATE. Use the FORMAT procedure to format the values of **Bonus**.
5. Are there any unusual data values that could be due to coding errors for any of the categorical variables?
6. Examine the distribution of **Bonus** at each value of the predictors. What associations do you see?
7. Use the FREQ procedure to test for an association between the variables **Lot\_Shape\_2**   
   and **Bonus** as well as **Fireplaces** and **Bonus**. Generate the expected cell frequencies   
   and the cell’s contribution to the total chi‑square statistic.

* 1. Is there is evidence of an association between **Lot\_Shape\_2** and **Bonus**?
  2. Does the Odds Ratio and Relative Risk table show a measure of strength of association? Explain.
  3. What’s the Odds Ratio? How would you interpret that result?

1. Use PROC FREQ to test whether an ordinal association exists between **Bonus** and **Fireplaces**.
   1. DOES **Bonus** and **Fireplaces** have a significant ordinal association (use Mantel‑Haenszel chi‑square test)?
   2. For the Spearman correlation statistic, is the relationship is significant at the 0.05 significance level?
2. Fit a binary logistic regression model in PROC LOGISTIC. Select **Bonus** as the outcome variable and **VARIABLE assigned to your team** as the predictor variable. Use the EVENT= option to model the probability of being bonus eligible and request profile likelihood confidence intervals around the estimated odds ratios. Use the ALPHA=.10 option in the MODEL statement
3. The model should be based on the probability of being bonus eligible (Bonus=1)
4. The Testing Global Null Hypothesis: BETA=0 table. What does BETA=0 mean?
5. Is there any evidence for at least one of the regression coefficients for an explanatory variable is significantly different from 0? Explain.
6. What’s the logistic regression equation?
7. Is the p‑value for the VARIABLE significant at the 0.10 significance level?
8. What does the odds ratio for the VARIABLE indicate?
9. Interpret the value of the c (concordance) statistic.