/\*\* Import an XLSX file.  \*\*/

PROC IMPORT DATAFILE="/folders/myfolders/sasuser.v94/house\_price\_test.xlsx"

   OUT=WORK.HOUSE\_PRICE\_TEST

   DBMS=XLSX

   REPLACE;

RUN;

/\* Run Means Procedure to check for Missing Values \*/

proc means data=House\_price\_test NMISS N;

RUN;

/\* after you've imported the training and test sets \*/

data test;

set house\_price\_test;

SalePrice = .;

run;

data train2;

set log\_House\_price\_train test;

run;

proc glm data = train2 plots = all;

class RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType;

model SalePrice = RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType LotArea BedroomAbvGr YearBuilt / cli;

output out = results p = Predict;

run;

/\* Can't have negative predictions because of RMSLE \*/

/\* Also must have only two columns with appropriate labels. \*/

data results2;

set results;

if Predict > 0 then Salerice = Predict;

if Predict < 0 then SalePrice = 100000;

keep id SalePrice;

where id> 1460;

;

proc means data = results2;

var SalePrice;

run;

NOTE: There were 24 observations out of the House\_price\_test dataset that SAS did not use. It didn’t give an error, just stated in the output “ \*Observation not used in this analysis “  -- I’m not sure if we should be concerned about this or not.









