```
proc datasets lib=work kill nolist memtype=data;
quit;

/* ...      read in training data set     ... */

FILENAME REFFILE '/folders/myfolders/stats_i/training_set_cleaned.csv';

PROC IMPORT DATAFILE = REFFILE
      DBMS = CSV
      OUT = home_prices;
      GETNAMES = yes;
RUN;

PROC CONTENTS DATA = home_prices; RUN;

/* ...      read in test data set   ... */

filename reffile '/folders/myfolders/stats_i/test_set_cleaned.csv';

proc import datafile = REFFILE
      DBMS = csv
      OUT = test_set;
      GETNAMES = yes;
RUN;

PROC CONTENTS DATA = test_set; RUN;

/* ...      combine train and test data sets         ... */

data train_test;
 set home_prices test_set;
run;

/* ...      scatter plots                            ... */
/* dependent response : log_saleprice    ... */

/*
proc sgscatter data = home_prices;
  matrix fullbath
            garagecars
            log_lotfrontage
            overallcond
            overallqual
            log_grlivarea
            totrmsabvgrd
            log_lotarea
            log_saleprice
            garagearea
            bsmtfinsf1
            x2ndflrsf
            totalbsmtsf
            x1stflrsf
            grlivarea
            yearbuilt
            yearremodadd
            / diagonal=(histogram normal);
run;
*/
```

```
/*******************************************************
                  second model with principal components
*******************************************************/

title 'PCR Using CrossValidation for Component Selection - all selected variables';
proc pls data = home_prices method = pcr cv = one cvtest (stat=press);
class housestyle
            garagetype
            masvnrtype
            neighborhood
            heatingqc
            bsmtqual
            exterqual
            kitchenqual
            bsmtfintype1
            fireplacequ
            foundation
            lotshape
            garagefinish
            mszoning
            electrical
            exterior1st
            exterior2nd
            saletype
            centralair;
model log_saleprice =
      /*          continuous variables    */
            fullbath
            garagecars
            log_lotfrontage
            overallcond
            overallqual
            log_grlivarea
            totrmsabvgrd
            log_lotarea
            garagearea
            bsmtfinsf1
            x2ndflrsf
            totalbsmtsf
            x1stflrsf
            grlivarea
            yearbuilt
            yearremodadd
      /*          categorical variables   */
            housestyle
            garagetype
            masvnrtype
            neighborhood
            heatingqc
            bsmtqual
            exterqual
            kitchenqual
            bsmtfintype1
            fireplacequ
            foundation
            lotshape
            garagefinish
            mszoning
```

```
            electrical
            exterior1st
            exterior2nd
            saletype
            centralair;
run;

title 'PCR Using Selected Factors';
proc pls data = train_test method = pcr nfact = 11;
class housestyle
            garagetype
            masvnrtype
            neighborhood
            heatingqc
            bsmtqual
            exterqual
            kitchenqual
            bsmtfintype1
            fireplacequ
            foundation
            lotshape
            garagefinish
            mszoning
            electrical
            exterior1st
            exterior2nd
            saletype
            centralair;
model log_saleprice =
      /*          continuous variables    */
            fullbath
            garagecars
            log_lotfrontage
            overallcond
            overallqual
            log_grlivarea
            totrmsabvgrd
            log_lotarea
            garagearea
            bsmtfinsf1
            x2ndflrsf
            totalbsmtsf
            x1stflrsf
            grlivarea
            yearbuilt
            yearremodadd
      /*          categorical variables   */
            housestyle
            garagetype
            masvnrtype
            neighborhood
            heatingqc
            bsmtqual
            exterqual
            kitchenqual
            bsmtfintype1
            fireplacequ
            foundation
            lotshape
```

```
            garagefinish
            mszoning
            electrical
            exterior1st
            exterior2nd
            saletype
            centralair;
            output out = result p = Predict;
run;

/* create kaggle submission file */
/* two columns with appropriate labels. */

proc means data = result Min Max;
run;

proc means data = result noprint;
      var Predict;
    output out = means mean(Predict) = mean_predict;
run;

data kaggle_submit;
set result;
SalePrice = exp(Predict);
if Predict = . then SalePrice = exp(12.018);
keep id SalePrice;
where id > 1460;
run;

proc export data = kaggle_submit replace
   outfile = '/folders/myfolders/stats_ii/kaggle_submit_pca.1.csv'
   dbms = csv;
run;
```