SAS Assignment: Linear Regression (ML and Non-ML Approaches)

Dataset: sas\_linear\_regression\_dataset.csv

Objective: Predict house prices using both classical and machine learning-based linear regression approaches in SAS.

## 1. Create the dataset in SAS using the DATA step or import a `.csv` file.

Use the following code to import the CSV:  
  
PROC IMPORT DATAFILE='/your-path/sas\_linear\_regression\_dataset.csv' OUT=house\_data DBMS=CSV REPLACE;  
GETNAMES=YES;  
RUN;

## 2. Check for missing values and outliers using PROC MEANS, PROC UNIVARIATE, and PROC FREQ.

PROC MEANS DATA=house\_data N NMISS MIN MAX MEAN STD;  
RUN;  
  
PROC UNIVARIATE DATA=house\_data;  
VAR Price;  
RUN;  
  
PROC FREQ DATA=house\_data;  
TABLES Has\_Garage;  
RUN;

## 3. Visualize relationships between `Price` and other predictors using PROC SGPLOT.

PROC SGPLOT DATA=house\_data;  
SCATTER X=Area\_sqft Y=Price;  
RUN;

## 4. Use PROC CORR to explore how each predictor is correlated with Price.

PROC CORR DATA=house\_data;  
VAR Area\_sqft Num\_Bedrooms Num\_Bathrooms Num\_Floors Has\_Garage Distance\_to\_City House\_Age Price;  
RUN;

## 5. Regress `Price` on `Area\_sqft` using PROC REG. Interpret the coefficients.

PROC REG DATA=house\_data;  
MODEL Price = Area\_sqft;  
RUN;  
  
Interpretation: Coefficient of Area\_sqft indicates the average increase in price per sqft.

## 6. Multiple regression with all predictors. Report R² and p-values.

PROC REG DATA=house\_data;  
MODEL Price = Area\_sqft Num\_Bedrooms Num\_Bathrooms Num\_Floors Has\_Garage Distance\_to\_City House\_Age;  
RUN;

## 7. Use VIF to check multicollinearity.

PROC REG DATA=house\_data;  
MODEL Price = Area\_sqft Num\_Bedrooms Num\_Bathrooms Num\_Floors Has\_Garage Distance\_to\_City House\_Age / VIF;  
RUN;

## 8. Remove insignificant predictors and refit the model.

Based on p-values, drop variables with p > 0.05 and refit using PROC REG.

## 9. Use residual plots to verify assumptions.

PROC REG DATA=house\_data;  
MODEL Price = Area\_sqft Num\_Bedrooms ... / PLOTS=ALL;  
RUN;

## 10. Add interaction term between Area\_sqft and Num\_Bedrooms.

PROC GLM DATA=house\_data;  
MODEL Price = Area\_sqft|Num\_Bedrooms;  
RUN;

## 11. Use PROC GLMSELECT for forward selection.

PROC GLMSELECT DATA=house\_data;  
MODEL Price = Area\_sqft Num\_Bedrooms Num\_Bathrooms Num\_Floors Has\_Garage Distance\_to\_City House\_Age / SELECTION=FORWARD;  
RUN;

## 12. Use PROC SURVEYSELECT to split into train/test.

PROC SURVEYSELECT DATA=house\_data OUT=train\_test SPLIT=0.7 OUTALL METHOD=SRS;  
RUN;

## 13. Train model using PROC GLMSELECT or PROC HPREG on training data.

PROC GLMSELECT DATA=train\_test;  
WHERE Selected=1;  
MODEL Price = Area\_sqft Num\_Bedrooms ... ;  
RUN;

## 14. Predict on test data and compute RMSE.

Use PROC SCORE or manually calculate RMSE using:  
DATA results;  
SET test\_predictions;  
Error = Price - Predicted;  
SquaredError = Error\*\*2;  
RUN;  
  
PROC MEANS DATA=results MEAN;  
VAR SquaredError;  
RUN;

## 15. Compare performance of ML and Non-ML approaches.

Compare R² and RMSE from PROC REG vs. PROC GLMSELECT. Note which model performs better and interpret why.