**SAS Script: Create Synthetic Housing Dataset**

data house\_data;

call streaminit(42); /\* for reproducibility \*/

do i = 1 to 100;

House\_ID = catt("H", put(i, z3.));

Area\_sqft = rand("integer", 800, 3500);

Num\_Bedrooms = rand("integer", 1, 5);

Num\_Bathrooms = rand("integer", 1, 3);

Num\_Floors = rand("integer", 1, 3);

Has\_Garage = rand("integer", 0, 1);

Distance\_to\_City = round(rand("uniform") \* 29 + 1, 0.01);

House\_Age = rand("integer", 0, 49);

/\* Price formula with noise \*/

Price = round(

50 +

0.05 \* Area\_sqft +

10 \* Num\_Bedrooms +

8 \* Num\_Bathrooms +

5 \* Num\_Floors +

15 \* Has\_Garage -

1.5 \* Distance\_to\_City -

0.3 \* House\_Age +

rand("normal", 0, 10)

, 0.01);

output;

end;

drop i;

run;

**Answer the below questions. In case you need help from ChatGPT feel free to use it. But make sure you understand the AI generated SAS script.**

**Section A: Data Exploration and Preparation**

1. **Create the dataset** in SAS using the DATA step or import a .csv file.
2. **Check for missing values** and outliers in the dataset using PROC MEANS, PROC UNIVARIATE, and PROC FREQ.
3. **Visualize relationships** between Price and other predictors using PROC SGPLOT (e.g., scatter plots, histograms).
4. **Correlation Analysis:** Use PROC CORR to explore how each predictor is correlated with Price.

**Section B: Non-ML Approach (Classical Regression using PROC REG / PROC GLM)**

1. **Simple Linear Regression:** Use PROC REG to regress Price on Area\_sqft. Interpret the coefficients.
2. **Multiple Linear Regression:** Regress Price on all numeric predictors. Report R², Adjusted R², and p-values.
3. **Multicollinearity Detection:** Use VIF in PROC REG to check for multicollinearity.
4. **Model Refinement:** Remove insignificant predictors and refit the model.
5. **Assumption Checks:** Use residual plots to verify linearity, normality, and homoscedasticity (PLOT option in PROC REG).
6. **Interaction Terms:** Add interaction between Area\_sqft and Num\_Bedrooms using PROC GLM and interpret the effect.

**Section C: ML-Based Approach (Using PROC GLMSELECT or PROC HPREG)**

1. **Feature Selection:** Use PROC GLMSELECT to apply forward selection or stepwise regression to identify best predictors.
2. **Train-Test Split:** Use PROC SURVEYSELECT to split data into 70% training and 30% testing datasets.
3. **Model Training:** Fit a linear regression model on training data using PROC GLMSELECT or PROC HPREG.
4. **Prediction:** Predict Price on the test set and calculate RMSE manually using DATA step or PROC SQL.
5. **Model Comparison:** Compare R², Adjusted R², and RMSE between Non-ML (PROC REG) and ML (PROC GLMSELECT) approaches.