# Demo: Exploring Ames Housing Data

Filename: **st101d01.sas**

In this demonstration, we use the FREQ and UNIVARIATE procedures to explore the Ameshousing3 table, generating graphics and summary statistics to learn more about the data.

```
PROC FREQ DATA=SAS-data-set;
     TABLES table-request(s) < / options>;
     <additional statements>
RUN;
```

```
PROC UNIVARIATE DATA=SAS-data-set < / options>;
     VAR variables;
     HISTOGRAM variables < / options>;
     INSET keywords < / options>;
RUN;
```

1. Select Libraries, My Libraries, and expand the STAT1 library. Double-click AMESHOUSING3, which contains a random sample of 300 homes from the original data. We use this table in most of our analyses, so we'll explore some of the variables.

2. Select Column labels in the View field to display more descriptive labels. The categorical variables include Style of dwelling, such as 1Story, 2Story, etc., Original construction year, Number of fireplaces, Foundation Type, such as Concrete/Slab or Cinder Block, and masonry veneer or not. The continuous variables include the Lot size in square feet, Above ground living area in square feet, Sale price in dollars, Basement area in square feet, Number of full bathrooms and half bathrooms, and Age of house when sold, in years.

3. Open program st101d01.sas.

```
/*st101d01.sas*/  /*Part A*/
/*Exploration of all variables that are available for analysis.*/
/*%let statements define macro variables containing lists of */
/*dataset variables*/
%let categorical=House_Style Overall_Qual Overall_Cond Year_Built
         Fireplaces Mo_Sold Yr_Sold Garage_Type_2 Foundation_2
         Heating_QC Masonry_Veneer Lot_Shape_2 Central_Air;
%let interval=SalePrice Log_Price Gr_Liv_Area Basement_Area
         Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr
         Full_Bathroom Half_Bathroom Total_Bathroom ;


/*st101d01.sas*/  /*Part B*/
/*PROC FREQ is used with categorical variables*/
ods graphics;
```

```
proc freq data=STAT1.ameshousing3;
    tables &categorical / plots=freqplot ;
    format House_Style $House_Style.
           Overall_Qual Overall.
           Overall_Cond Overall.
           Heating_QC $Heating_QC.
           Central_Air $NoYes.
           Masonry_Veneer $NoYes.
           ;
    title "Categorical Variable Frequency Analysis";
run;

/*st101d01.sas*/  /*Part C*/
/*PROC UNIVARIATE provides summary statistics and plots for */
/*interval variables.  The ODS statement specifies that only */
/*the histogram be displayed.  The INSET statement requests */
/*summary statistics without having to print out tables.*/
ods select histogram;
proc univariate data=STAT1.ameshousing3 noprint;
    var &interval;
    histogram &interval / normal kernel;
    inset n mean std / position=ne;
    title "Interval Variable Distribution Analysis";
run;
title;
```

Part A defines macro variables to help organize the data set variables and make modifying the SAS code easier. The %LET statements are used to name the macro variables and set their values. The first %LET statement creates a macro variable named categorical, and assigns it a space-delimited list of the categorical variables in the table. The next %LET statement creates the macro variable named interval, and assigns it the names of all the interval variables. Now, instead of typing a long list of variable names over and over in your programs, you can simply reference macro variable values by placing an ampersand in front of the macro variable name.

In Part B, the PROC FREQ step uses the ameshousing3 data to generate frequency tables and plots summarizing the categorical variables. In PROC FREQ, you list the analysis variables in the TABLES statement. The macro variable reference &categorical is replaced with the macro variable's value, the list of categorical variables, when you submit the step. The PLOTS= option requests a frequency plot. And by including the FORMAT statement, the data will be formatted and grouped before being analyzed.

4. Submit the code in Parts A and B and check the log to verify that there are no error or warning messages.

5. Review the output.

   The Table House_Style, One-Way Frequencies table indicates that from the 300 homes in our sample, almost 200 are one-story homes. There are few homes with other styles, and there are only six observations with the house style 2nd level unfinished. There are too few members to analyze, so they'll be merged with the One story and Two story levels in the variable House_Style2.

   In the Tables Overall_Qual and Overall_Cond, the One-Way Frequencies and Distribution Plots indicate that the variables representing the overall quality and overall condition of the homes are predominantly average. Both variables have many levels with small frequencies. For example, there's only one home each with an overall quality of 1, the poorest level, and 9, the best level. We'll trichotomize these two variables into Below Average, Average, and AboveAverage, in the variables Overall_Qual2 (overall quality 2) and Overall_Cond2 (overall condition 2).

In the Table Year_Built, the One-Way Frequencies and Distribution Plots indicate that Year_Built ranges from 1875 to 2009 and has more values than is practical to treat as a categorical variable in a statistical model with only 300 observations, so we'll treat it as interval.

In the Table Fireplaces, the One-Way Frequencies and Distribution Plots indicate that 195 homes have no fireplace, 93 have a single fireplace, and 12 homes have two fireplaces. Because the number of fireplaces has a natural ordering, we can treat Fireplaces as an ordinal variable.

In the Table Mo_Sold, the One-Way Frequencies and Distribution Plots indicate that the variable representing month sold shows a clear trend toward sales in the summer months, July and June. Some months have small numbers, so instead of analyzing by month, we created Season_Sold to use in subsequent analyses. Season 1 is from month 12 to month 2; season 2 is from month 3 to month 5; season 3 is from month 6 to month 8; and season 4 is from month 9 to month 11.

In the Table Yr_Sold, the One-Way Frequencies and Distribution Plots indicate that Yr_Sold is fairly uniform, meaning there were a similar number of homes sold each year between 2006 and 2010.

In the Table Garage_Type_2, the One-Way Frequencies and Distribution Plots indicate that the Garage_Type_2 variable shows that 159 homes have an attached garage, 109 have a detached garage, and 29 homes do not have a garage (represented by NA). The table also states that there are three homes with missing information.

The Table Foundation_2, the One-Way Frequencies and Distribution Plots show that most homes are on cinder block, followed by concrete and then brick tile or stone.

The Table Heating_QC, the One-Way Frequencies and Distribution Plots show that there are four levels of heating quality, excellent, fair, good, and average. Fortunately, most homes have excellent or average heating quality.

The Table Masonry_Veneer, the One-Way Frequencies and Distribution Plots show that most homes do not have masonry veneer. Only 89 do.

The Table Lot_Shape_2, the One-Way Frequencies and Distribution Plots show that most homes have a regular lot shape, and a majority have central air.

6. Use PROC UNIVARIATE to explore the continuous or interval variables, plotting histograms of the data to see the shape and spread, and also print the mean and standard deviation summary statistics. The PROC UNIVARIATE step in Part B performs a distribution analysis and plots the distribution of the continuous variables. The NOPRINT option suppresses the other output. We reference the interval macro variable in the VAR and HISTOGRAM statements, and request a normal curve, a kernel density estimate, and an inset box in the top right, or northeast corner, displaying the number of rows, the mean, and the standard deviation.

7. Submit this step.

8. Review the output.

   The first histogram, Distribution of SalePrice shows that for the continuous variable, SalePrice, the average sale price of homes in our sample is $137,524. The histogram is bell shaped, referring to a Gaussian, or normal distribution, a quality of the data that's important for our analyses in subsequent lessons. The blue line overlaying the plot is a normal density estimate and the red line is a kernel density estimate, which basically mimics the histogram. If these two overlayed lines are similar, the data are close to a normal distribution. We'll discuss this more in Lesson 1.

   Sometimes researchers use a log transformation on an outcome variable such as SalePrice to provide more bell-shaped or normal-looking data for future analyses. In reviewing the Log_Price histogram, we see that both the original variable and the log transformation provide bell-shaped data.

   The Gr_Liv_Area histogram indicate that on average, homes in Ames, Iowa, have 1,130 square feet of above ground living area. Most homes range from 900 to 1380 square feet.

The Basement_Area histogram is fairly bell shaped with a mean of 882 square feet. The Garage_Area histogram shows that the average garage area for homes with a garage is 369 square feet. The Deck_Porch_Area histogram is an example of a skewed distribution. Most of the observations, approximately 40%, have no deck, and then we see fewer and fewer larger decks. The Lot_Area histogram is fairly normal looking with a mean of 8,294 square feet. The Age_Sold histogram shows that the average age of homes sold in our sample is approximately 46 years, and the ages range from new to about 132 years old.

The Bedroom_AbvGr histogram shows that the number of bedrooms above ground, which could also be analyzed as a categorical variable, is 2.5 on average. Similarly, in the Full_Bathroom, Half_Bathroom, and Total_Bathroom histograms, the number of full, half and total bathrooms could also be analyzed as categorical variables, and the average numbers are 1.68, 0.25, and 1.70, respectively.

After looking at the variables in our course data, you might have some intuition as to which variables could accurately model sale price. For example, we could do an analysis of variance to see whether homes with central air are more likely to sell for higher prices, or whether homes with excellent heating condition are associated with higher priced homes. In addition, we could use regression to see whether the above ground living area, as a proxy for the size of homes, is correlated with SalePrice. Or perhaps we can model the probability of the home selling for more than $175,000 using both the number of fireplaces and basement area jointly. These are questions we'll be able to answer going forward.

Close