Print

Demo: Exploring Associations Using PROC SGPLOT

Filename: **st102d01.sas**

In this demo, you use PROC SGPLOT to create box plots, and visually examine the association between the categorical predictor Central_Air and the continuous response SalePrice in the ameshousing3 data. As shown in the syntax, PROC SGPLOT has many graphical options to visually explore data and create a variety of plots.

PROC SGPLOT DATA=*SAS-data-set <options>***;**
HBAR *category-variables* **<** */ options>***;**
VBAR *category-variables* **<** */ options>***;**
HBOX *category-variables* **<** */ options>***;**
VBOX *category-variables* **<** */ options>***;**
RUN;

1. Open program st102d01.sas.

```
/*st102d01.sas*/  /*Part C*/
proc sgplot data=STAT1.ameshousing3;
    vbox SalePrice / category=Central_Air
                     connect=mean;
    title "Sale Price Differences across Central Air";
run;
```

   Part C of the program uses the VBOX statement to create a vertical box plot that shows the distribution of the data. It specifies the variable for the Y axis, SalePrice, followed by a forward slash. The CATEGORY= option specifies the category variable and creates different box plots for each distinct value.This means it will create a plot for homes with no central air, and one for those with central air. To help us visually assess the relationship between our variables, we'll use connect=mean to include the regression line and connect the means of Y at each value of X.

2. Submit this step.

3. Review the output.

   In the SGPlot Procedure box plot, SalePrice is on the Y axis, and Central_Air is on the X axis. The No group is on the left, the Yes group is on the right, and there are some outliers. The regression line is definitely not horizontal. Clearly, there appears to be an association between Central_Air and SalePrice. It seems that homes with central air tend to sell for higher prices than homes without.

   Exploring associations with box plots helps prepare you for what you might encounter as you analyze your data. However, don't use these plots exclusively to determine which variables to include in your model. They represent only simple relationships between one predictor variable and the response variable. When you start putting multiple variables in the model, the picture of associations can become very different.

Close