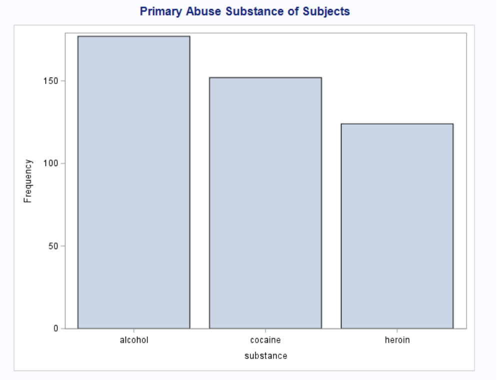
The data set used to illustrate graphing with SAS is the HELP study (data name is new), which was a clinical trial for adult inpatients recruited from a detoxification unit. The variables that we use throughout this tutorial include depression (`cesd`), homelessness status (`homeless`), primary abuse substance (`substance`), patient's age (`age`), and patient's sex (`sex`).

Univariate Graphing

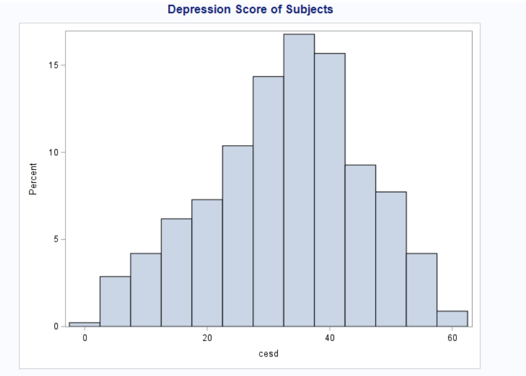
- Suppose we would like a plot of a single categorical variable.

```
proc sgplot data=new;
title 'Primary Abuse Substance of Subjects';
vbar substance;
run;
```

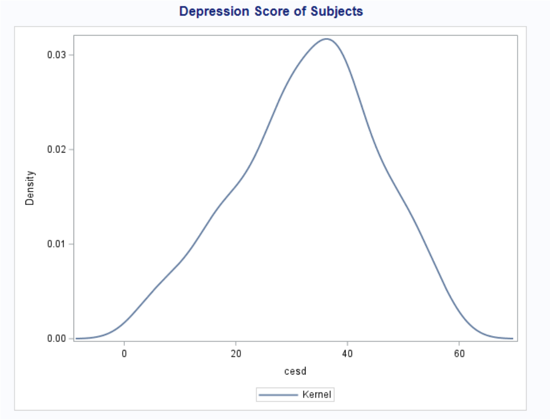


- Now for a plot of a single quantitative variable

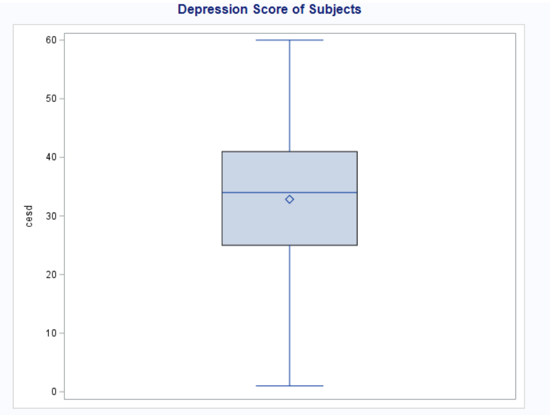
```
/*HISTOGRAM*/

proc sgplot data=new;
title 'Depression Score of Subjects';
histogram cesd;
run;
```



```
/*DENSITY PLOT*/

proc sgplot data=new;
title 'Depression Score of Subjects';
density cesd/ type=kernel;
run;
```



```
/*BOXPLOT*/
proc sgplot data=new;
title 'Depression Score of Subjects';
vbox cesd;
run;
```

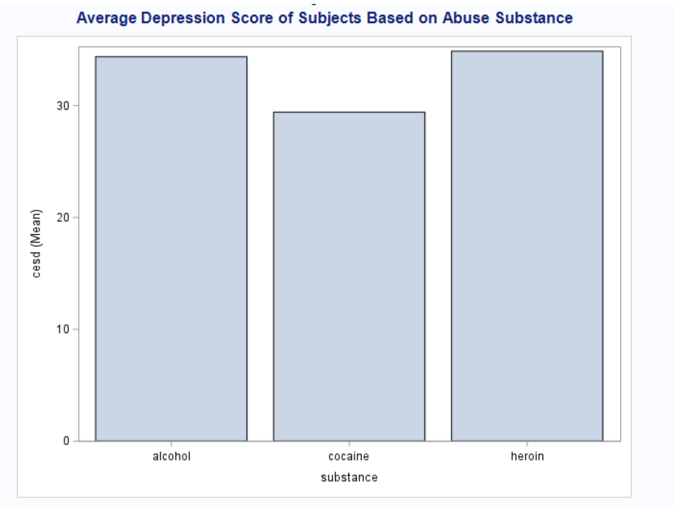


Bivariate Graphing

C→ Q

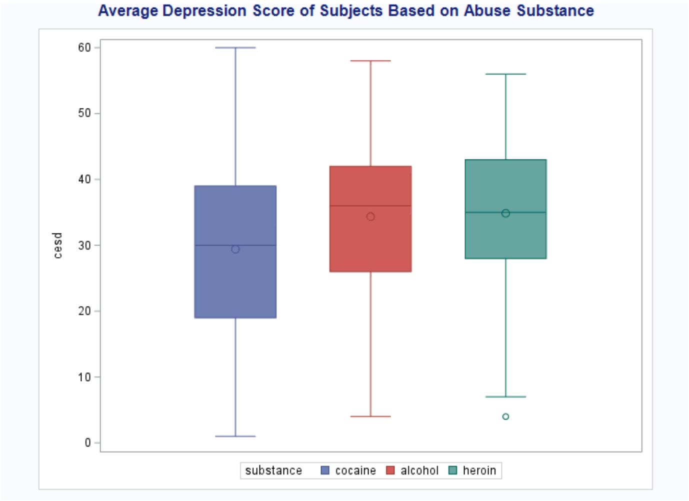
- OPTION 1: Construct a bar plot with mean of response variable on y-axis.

```
proc sgplot data=new;
title 'Average Depression Score of Subjects Based on Abuse Substance ';
vbar substance / response=cesd stat=mean;
run;
```



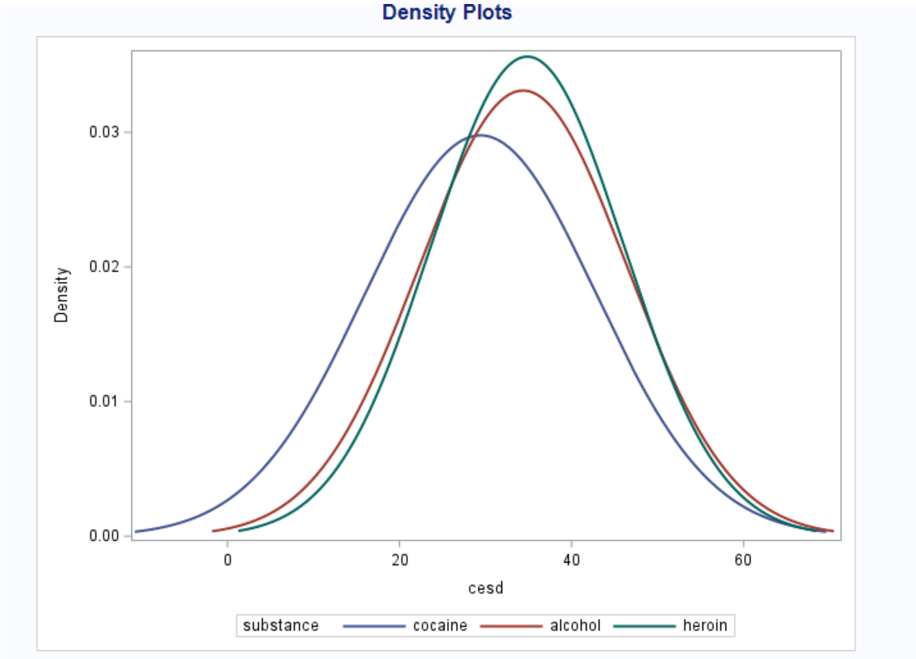
- OPTION 2: Boxplots

```
proc sgplot data=new;
title 'Average Depression Score of Subjects Based on Abuse Substance ';
vbox cesd / group=substance;
run;
```



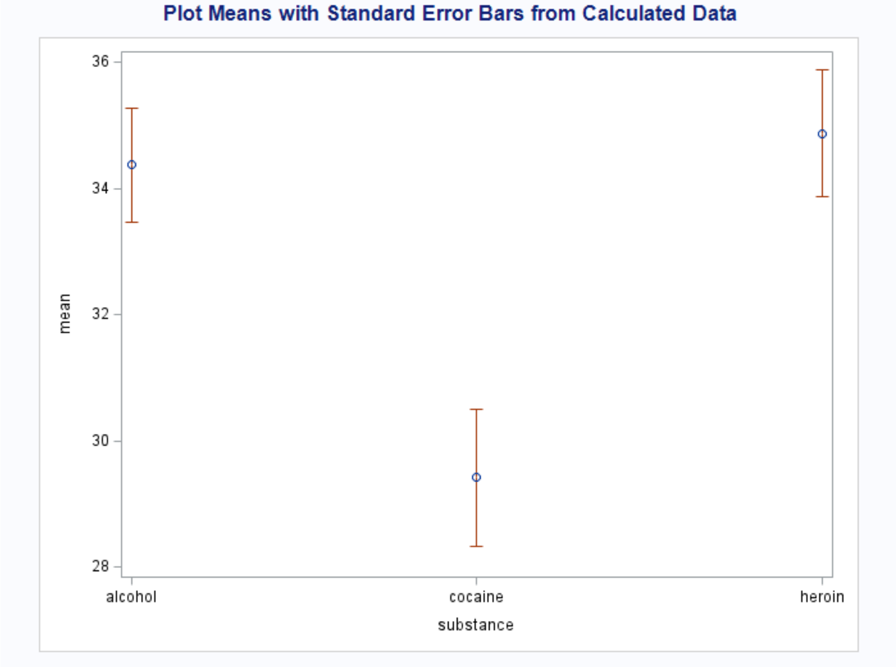
- OPTION 3: Density Plots

```
proc sgplot data=new;
  title 'Density Plots';
  density cesd  / type=kernel group=substance;
  run;
```



- OPTION 4: Mean of Response with Error Bars

```
/*REQUIRES DATA PREP WORK AND CONSTRUCTION OF NEW WORKING DATA SET*/

/* Sort data and then find group means and standard deviations. Save results to meansout */
proc sort data=new;
   by substance;
run;

proc means data=new noprint;
   by substance;
   var cesd;
   output out=meansout mean=mean stderr=stderr;
run;

/* Calculate the upper and lower error bar values. */
data reshape(drop=stderr);
   set meansout;
   lower=mean - stderr;
   upper=mean + stderr;
run;


proc sgplot data=reshape;
   scatter x=substance y=mean / yerrorlower=lower
                           yerrorupper=upper;
   title1 'Plot Means with Standard Error Bars from Calculated Data';
run;
```

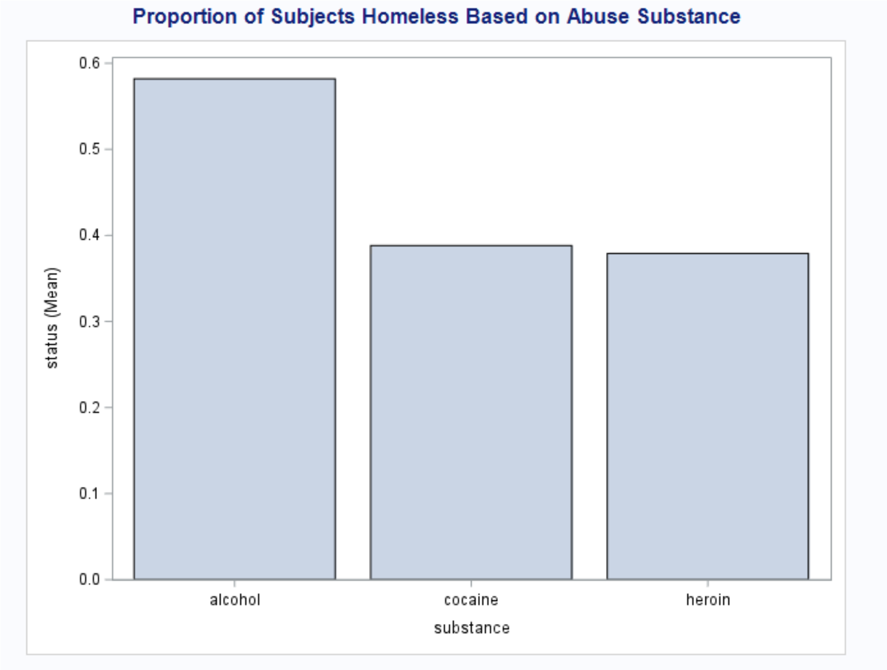


C→ C

- If you have a binary response variable (that is, a response variable that takes on two possible values) - you can display the proportion of participants at an indicated response level for each level of a categorical variable.

```
data new; set myFolder.HELP;
if homeless='homeless' then status=1;
if homeless='housed' then status=0;
```

```
proc sgplot data=new;
title 'Proportion of Subjects Homeless Based on Abuse Substance ';
vbar substance / response=status stat=mean;
run;
```

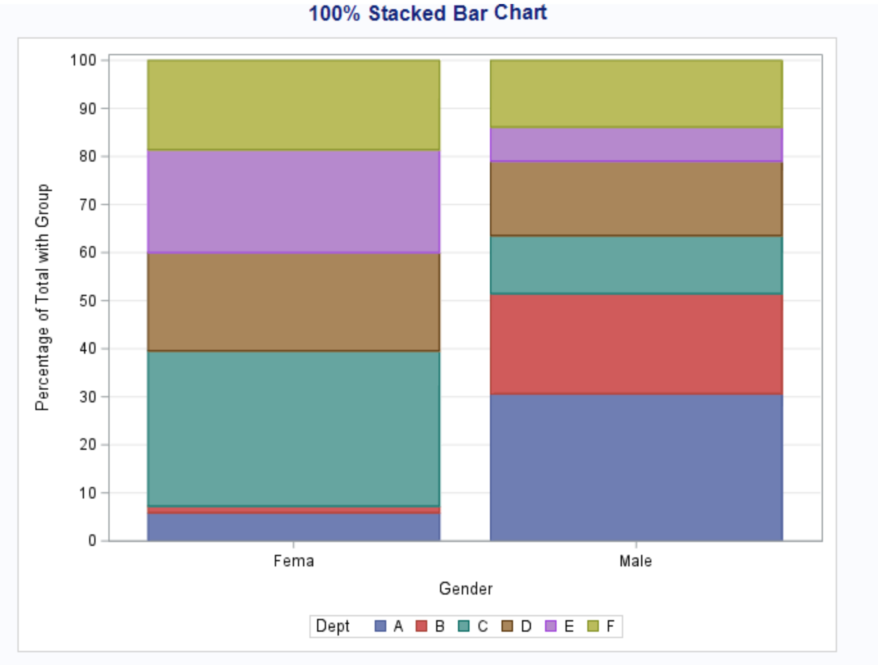


- If you have a categorical response variable that takes on more than 2 categories, then perhaps the graph below will work well for you (the graph also works for a response variable with two levels, but the above plot is typically a better choice for that scenario). It will display the proportion of participants at each level of the response variable based on a categorical explanatory variable. In the example below, suppose we would like to understand how Gender relates to Department choice at UC Berkeley in 1973. (Notice that there seems to be some relationship between Gender and Department selection - Males and Females are prone to apply to different departments. For example, Males were prone to apply to Departments A and B, and it was among one of the least likely departments for females to apply to.)

```
proc sort data=UCBADMISSIONS;
by gender;
run;
```

```
proc freq data=UCBADMISSIONS;
by gender;
tables dept / out=FreqOut;
run;
```

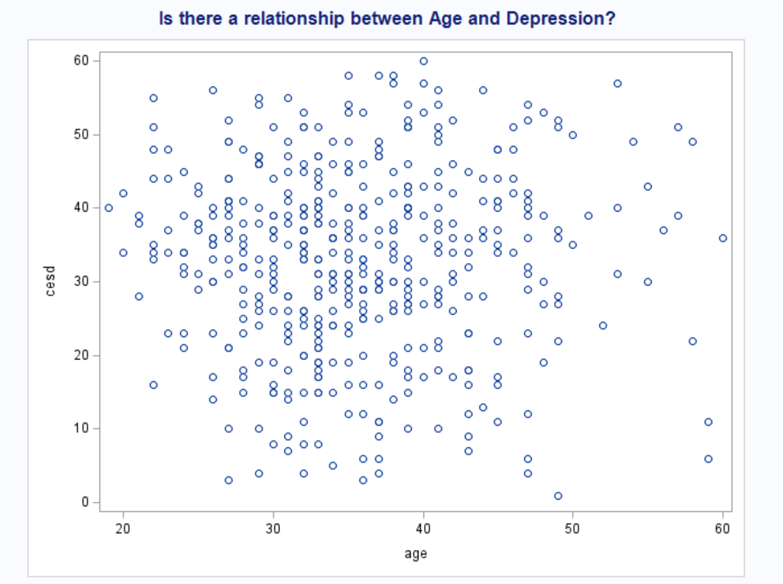
```
proc sgplot data=FreqOut;
title "100% Stacked Bar Chart";
vbar gender / response=Percent group=dept groupdisplay=stack;
xaxis discreteorder=data;
yaxis grid values=(0 to 100 by 10) label="Percentage of Total with Group";
run;
```



Q→ Q

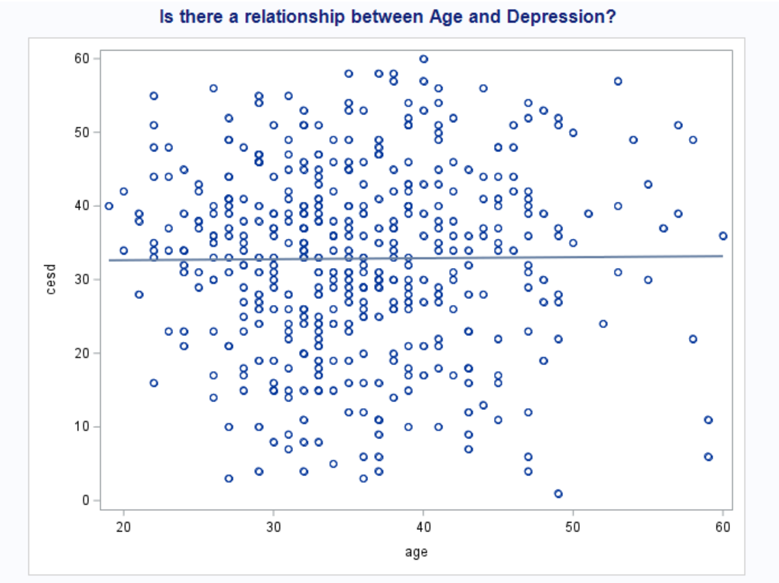
- Now, let's look at an explanatory and response variable which are both quantitative.

```
proc sgplot data=new;
      title 'Is there a relationship between Age and Depression?';
      scatter x=age y=cesd;
   run;
```

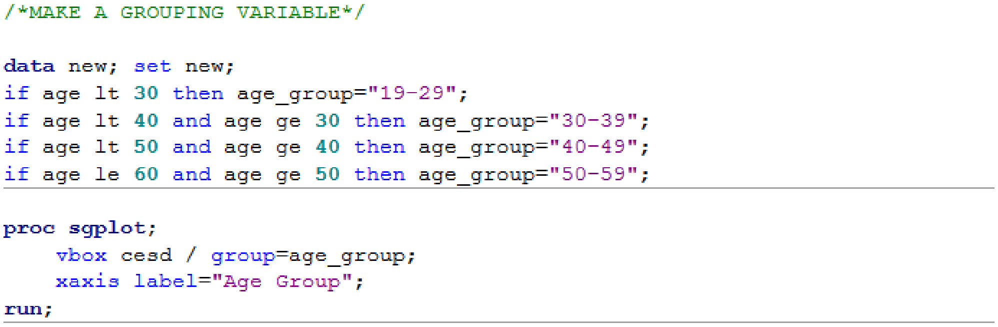


- This may be a bit much to look at and it is difficult to see overall trends. You may want to make a line of best fit to help determine whether a linear trend exists.

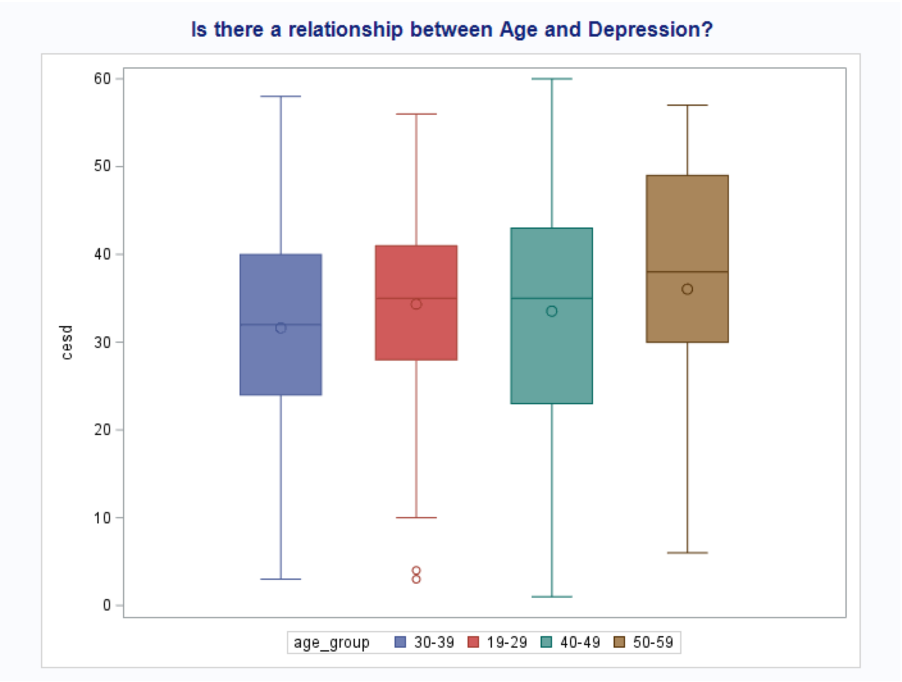
```
proc sgplot data=new noautolegend;
      title 'Is there a relationship between Age and Depression?';
      scatter x=age y=cesd;
      reg y=cesd x=age;
   run;
```



- Another alternative is to create a categorical version of age and find the relationship between depression score for each age group. This may be helpful when the trend in age is not linear.

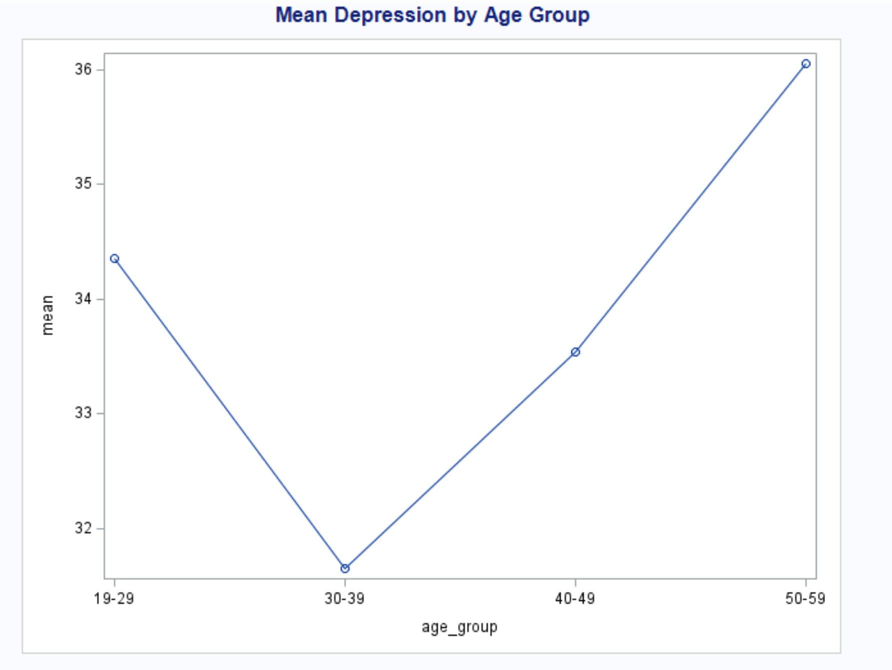


```
/*MAKE A GROUPING VARIABLE*/

data new; set new;
if age lt 30 then age_group="19-29";
if age lt 40 and age ge 30 then age_group="30-39";
if age lt 50 and age ge 40 then age_group="40-49";
if age le 60 and age ge 50 then age_group="50-59";

proc sgplot;
    vbox cesd / group=age_group;
    xaxis label="Age Group";
run;
```



- You could also use the grouping variable above to look exclusively at the means of each group.

```
/*Option to show means by group*/
proc sort;
by age_group;

proc means data=new;
   by age_group;
   var cesd;
   output out=meansout mean=mean;
run;

proc sgplot data=meansout noautolegend;
   scatter x=age_group y=mean;
   series x=age_group y=mean;
   title1 'Mean Depression by Age Group';
run;
```



Multivariate Graphing

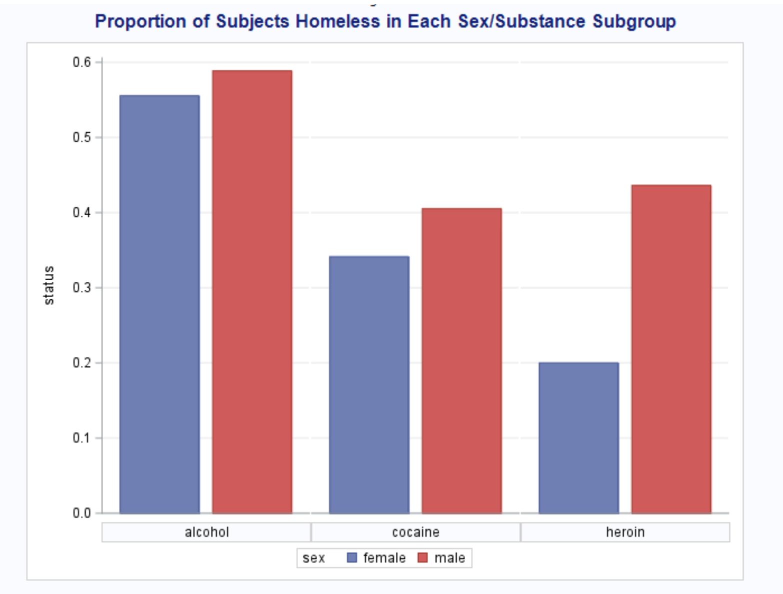
C→ C with Categorical Third Variable

- **When response variable is binary**: Suppose we wish to visualize the relationship between two categorical variables, controlling for an additional categorical variable. For this example, suppose that substance abused is the explanatory variable and homelessness status is the response variable. Assume further that sex is an additional explanatory variable of interest.

```
proc sort data=new; by substance sex;
```

```
proc freq data=new;
by substance sex;
tables status / out=FreqOut;
run;
```

```
proc sgpanel data=new;
title 'Proportion of Subjects Homeless in Each Sex/Substance Subgroup';
 panelby substance / layout=columnlattice onepanel
         colheaderpos=bottom rows=1 novarname noborder;
 vbar sex / group=sex response=status stat=mean group=sex nostatlabel;
 colaxis display=none;
 rowaxis grid;
 run;
```

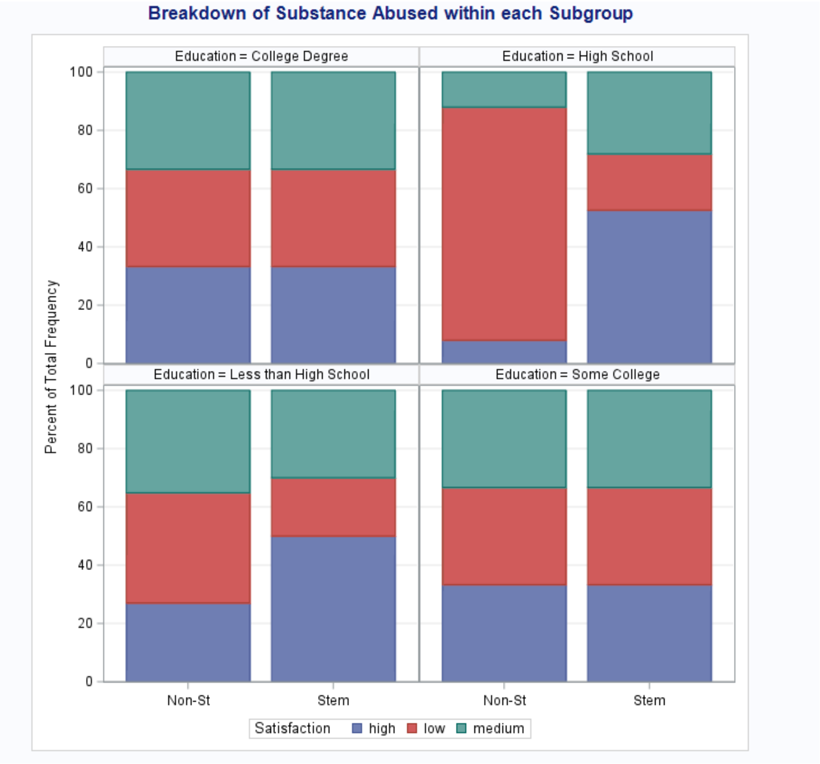


- **When response variable has more than 2 categorical levels**: Suppose we wish to predict level of job satisfication (low, medium, high) based on profession (Stem, Non-Stem). Additionally, we want to see whether level of education (Less than High School, High School, Some College, College Degree) plays a role in this relationship.

```
proc sort data=job_data;
by Profession Education;
```

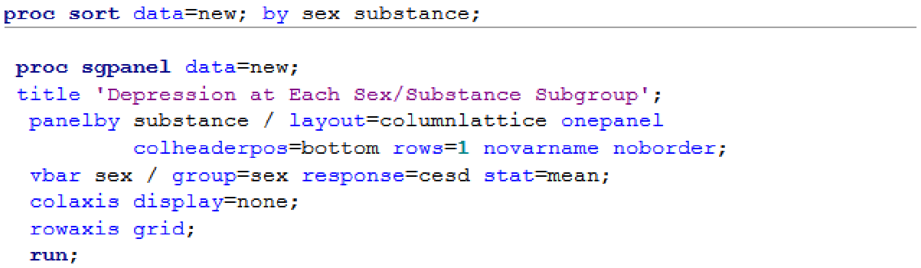
```
proc freq data=job_data;
by Profession Education;
tables Satisfaction / out=FreqOut;
run;
```

```
proc sgpanel data=FreqOut;
  title 'Breakdown of Substance Abused within each Subgroup';
  panelby Education;
  vbar Profession / response=Percent group=Satisfaction groupdisplay=stack;
  colaxis display=(nolabel);
  rowaxis grid;
  run;
```

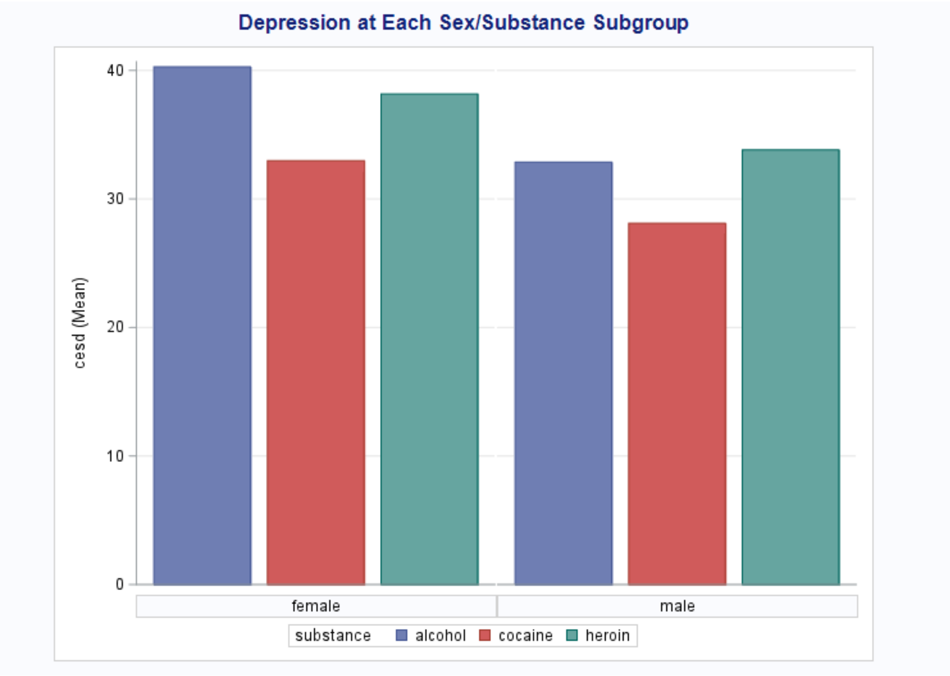


C→ Q with Categorical Third Variable

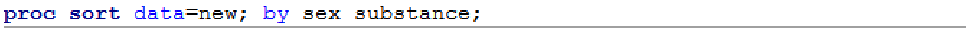
- Here, I wish to understand how abuse substance (explanatory variable) relates to depression of patients (response variable). In addition, I wish to understand how sex may play a role in this relationship.
- OPTION 1:



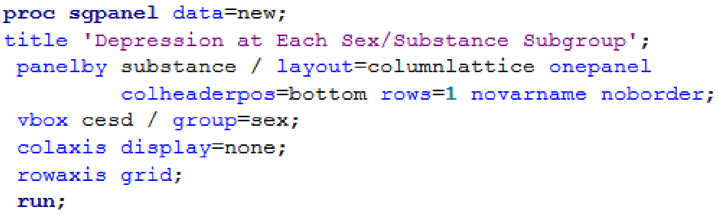
```
proc sort data=new; by sex substance;

proc sgpanel data=new;
title 'Depression at Each Sex/Substance Subgroup';
 panelby substance / layout=columnlattice onepanel
         colheaderpos=bottom rows=1 novarname noborder;
 vbar sex / group=sex response=cesd stat=mean;
 colaxis display=none;
 rowaxis grid;
 run;
```



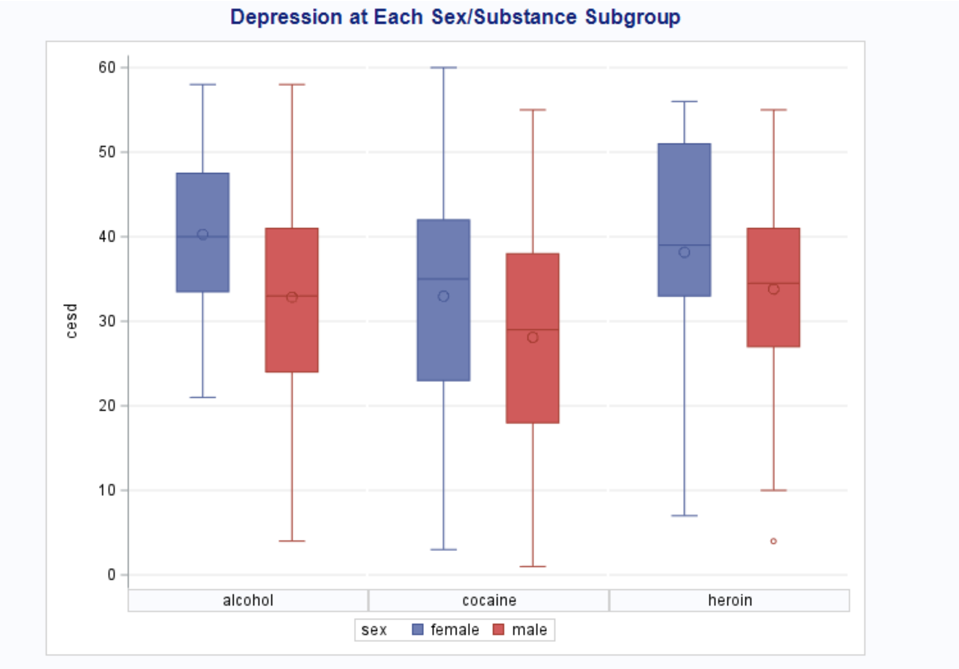
- OPTION 2:



```
proc sort data=new; by sex substance;
```



```
proc sgpanel data=new;
title 'Depression at Each Sex/Substance Subgroup';
 panelby substance / layout=columnlattice onepanel
          colheaderpos=bottom rows=1 novarname noborder;
 vbox cesd / group=sex;
 colaxis display=none;
 rowaxis grid;
 run;
```



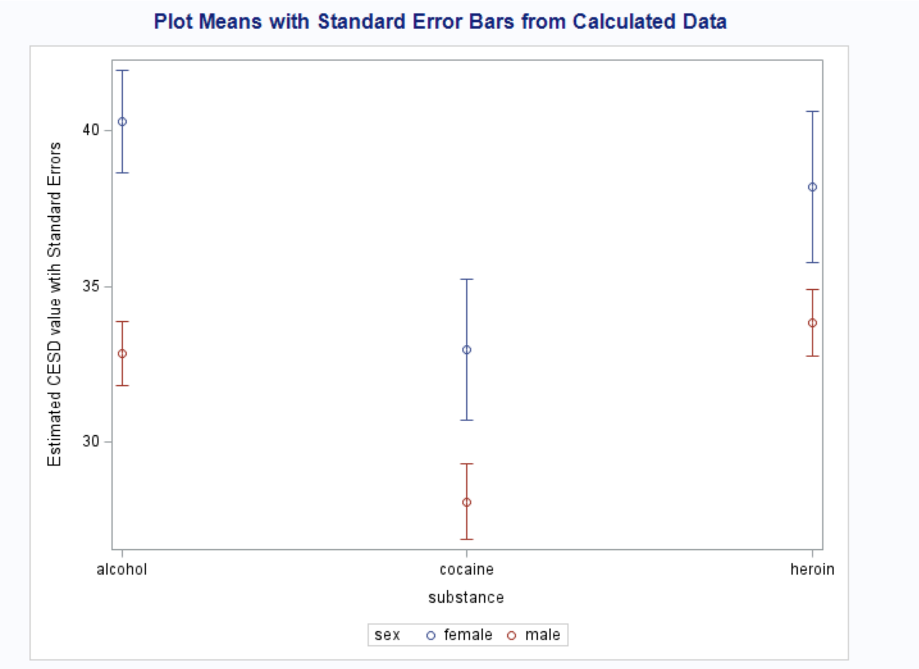
- OPTION 3:

```
/* Calculate the means and the standard errors. */
proc sort data=new;
   by substance sex;
run;
```

```
proc means data=new noprint;
   by substance sex;
   var cesd;
   output out=meansout mean=mean stderr=stderr;
run;
```

```
/* Calculate the upper and lower error bar values. */
data reshape(drop=stderr);
   set meansout;
   lower=mean - stderr;
   upper=mean + stderr;
run;
```

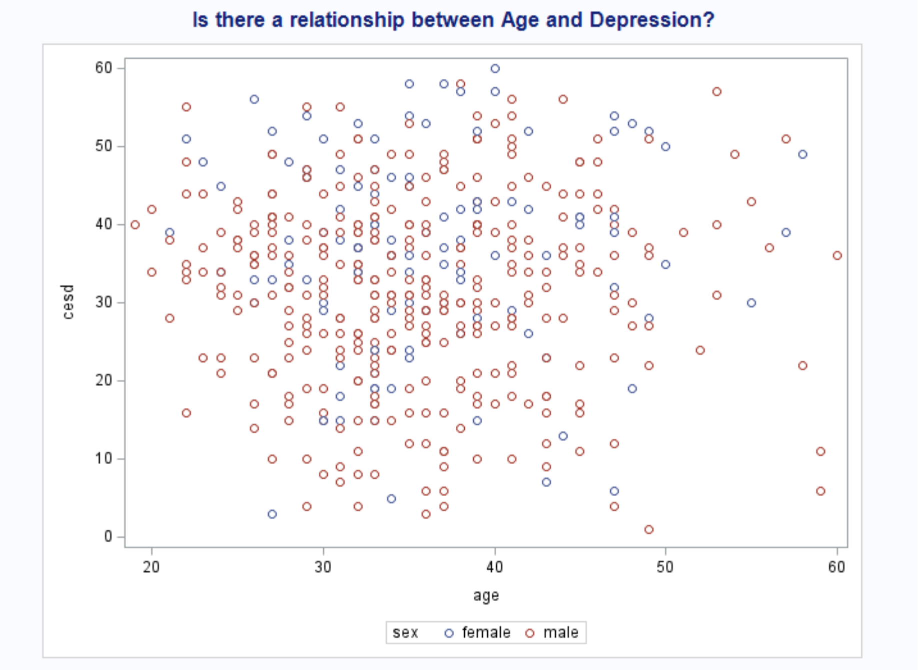
```
proc sgplot data=reshape;
   scatter x=substance y=mean / group=sex yerrorlower=lower
                           yerrorupper=upper;
   yaxis label="Estimated CESD value wtih Standard Errors";
   title1 'Plot Means with Standard Error Bars from Calculated Data';
run;
```



Q→ Q with Categorical Third Variable

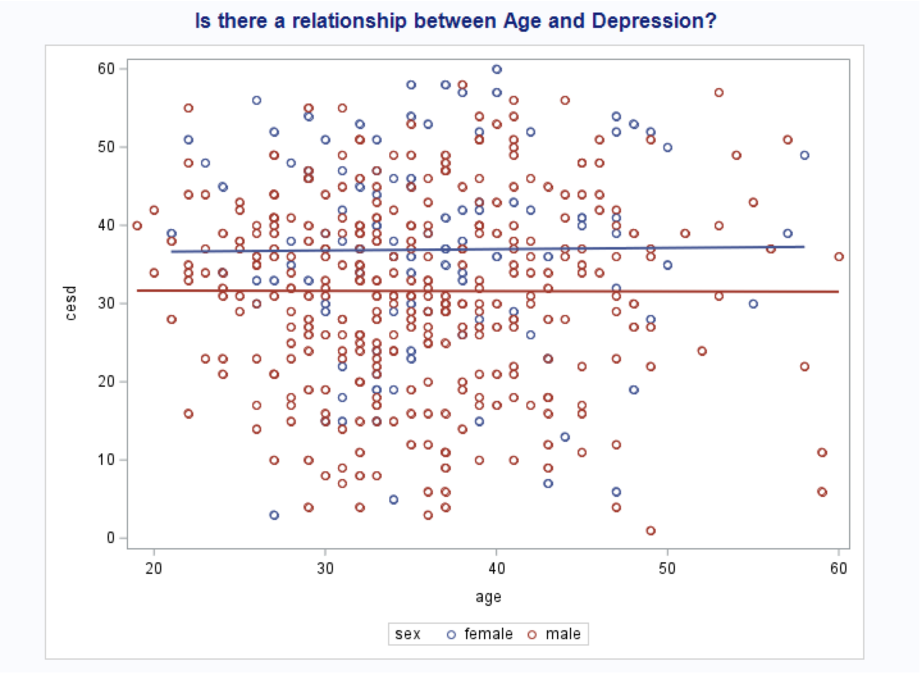
- Suppose we wish to visualize the relationship between age (explanatory variable) and cesd (a measure of depression and the response variable in this study) based on sex.
- OPTION 1:

```
proc sgplot data=new;
      title 'Is there a relationship between Age and Depression?';
      scatter x=age y=cesd/group=sex;
   run;
```



- OPTION 2: Suppose we wish to add separate regression lines for each sex.

```
proc sgplot data=new;
      title 'Is there a relationship between Age and Depression?';
      scatter x=age y=cesd/group=sex;
      reg y=cesd x=age/group=sex;
   run;
```



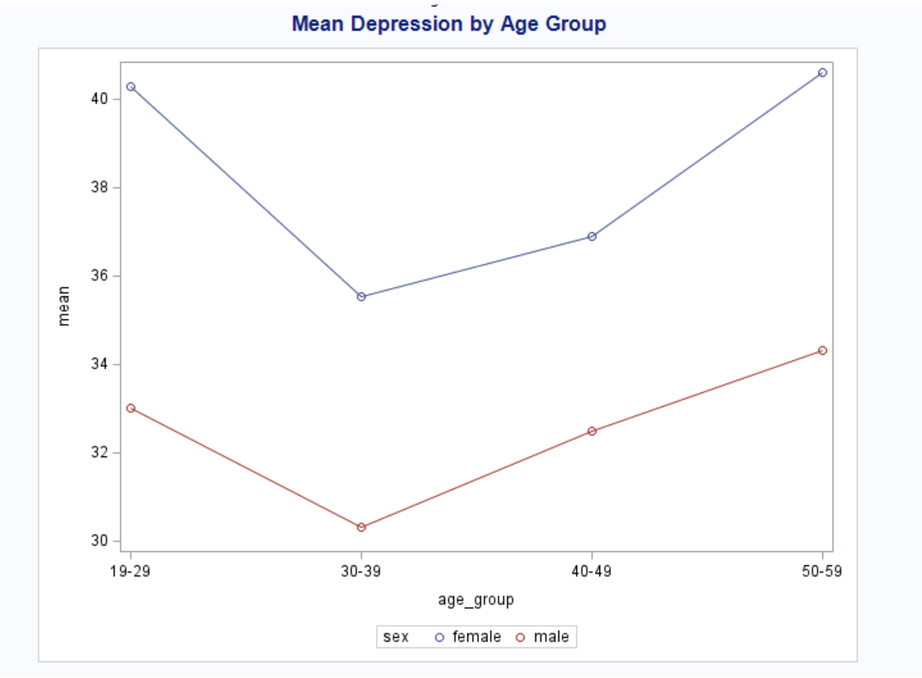
- OPTION 3: Or, again, we could use the categorical version to see how the trends in depression based on age group for males and females are similar.

```
/*MAKE A GROUPING VARIABLE*/

data new; set new;
if age lt 30 then age_group="19-29";
if age lt 40 and age ge 30 then age_group="30-39";
if age lt 50 and age ge 40 then age_group="40-49";
if age le 60 and age ge 50 then age_group="50-59";
```

```
proc sort data=new;
by age_group sex;
```

```
/*MAKE MEANS WITHIN EACH COMBINATION OF AGE_GROUP/SEX AND STORE VALUES*/

proc means data=new;
   by age_group sex;
   var cesd;
   output out=meansout mean=mean;
run;
```

```
proc sgplot data=meansout;
   scatter x=age_group y=mean/group=sex;
   series x=age_group y=mean/group=sex;
   title 'Mean Depression by Age Group';
run;
```


Mean Depression by Age Group
mean
age_group
sex ○ female ○ male