Stat 230 - Basic R you may remember

P.B. Matheson adapted from Shu-Min Liao

Feb 7, 2022

Basics

1. We will always use the mosaic package - begin with require(mosaic)
2. Sometimes we will use mosaic data sets - use require(mosaicData)
3. We will use the ggformula commands to create graphics
4. You should spell check your work (see the Edit pulldown menu and click Check Spelling)
5. Sometimes we will load some of the data sets from our book (Stat2Data) - A link will be provided.

Help

1. Watch the "Getting Started" videos from Nick Horton (in numerical order) at //nhorton.people.amherst.edu/rstudio/
2. If you aren't sure how to use a particular command type ? and the command in the console to learn more, for example - ?tally
3. See the resources in Moodle under "All Things R"

Examples of R commands below are given based on variables from the dataset called HELPrct.
The HELP study was a randomized clinical trial (rct) for adult inpatients recruited from a detoxification unit. Patients with no primary care physician were randomized to receive a multidisciplinary assessment and a brief motivational intervention or usual care, with the goal of linking them to primary medical care. The cesd variable represents a Center for Epidemiologic Studies Depression measure at baseline (high scores indicate more depressive symptoms). The mcs variable represents the SF-36 Mental Component Score (measured at baseline, lower scores indicate worse status).

We are using the four step modeling process of CHOOSE, FIT, ASSESS and USE. The first step is CHOOSE and all the commands given below will help you choose the right statistics and visualizations. The type of analysis that you can do depends on the type of data you have and how many variables you are considering. Is it qualitative (categorical like dogs vs. cats) or quantitative (IQ or height)? Do you have 1 variable, 2 variables, or 3 variables?

UNIVARIATE - examining one variable

The basic structure is:

command (~ varname, data = ds) where varname is the name of the variable you want to look at and ds is the data file you are working on

Qualitative/Categorical variables - see examples in the R chunk below

a) can use numerical summaries such as frequencies, relative frequencies (proportions) via the tally command

```
tally (~substance, data=HELPrct)
```

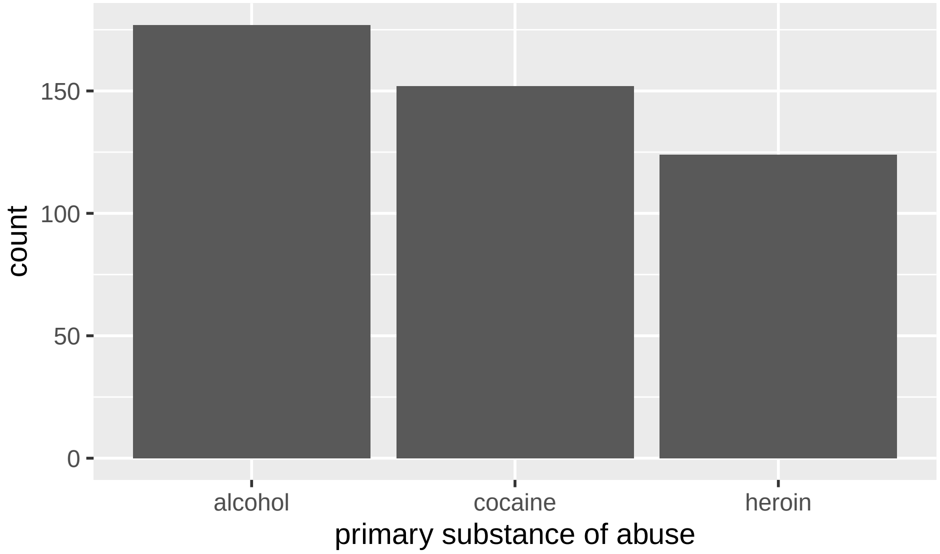
```
## substance
## alcohol cocaine  heroin
##     177     152     124
```

```
tally (~substance, format="proportion", data=HELPrct)
```

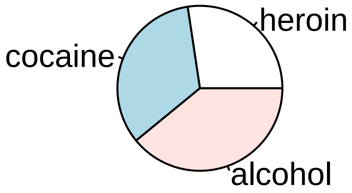
```
## substance
##  alcohol  cocaine   heroin
## 0.390728 0.335541 0.273731
```

b) can use graphical displays such as bar chart, pie chart

```
gf_bar (~substance, data=HELPrct)
```



```
pie ( sort( tally(~substance, data=HELPrct) ) )
```



Quantitative variables - see examples in the R chunk below

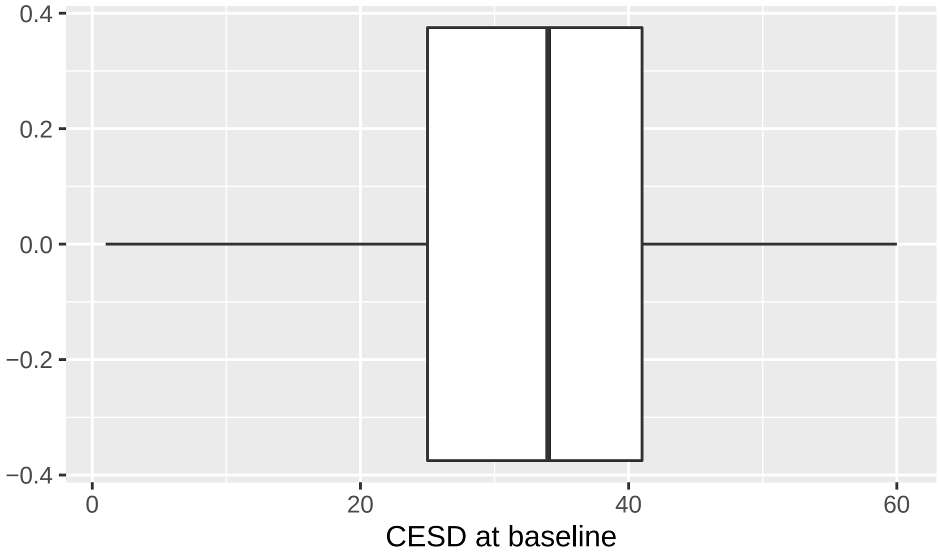
a) can use summary statistics: 5 number, median and IQR, Mean and SD

```
favstats(~cesd, data=HELPrct)
```

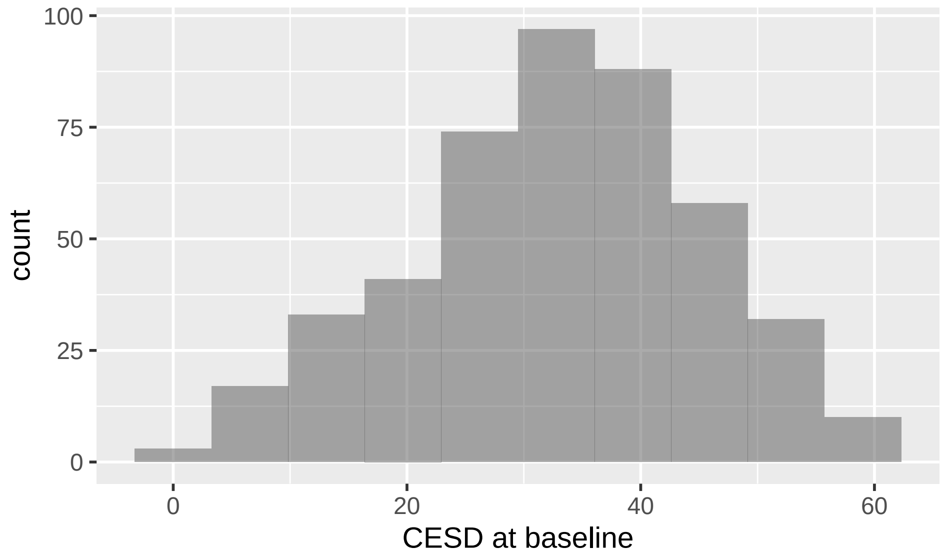
```
##  min Q1 median Q3 max    mean      sd   n missing
##    1 25     34 41  60 32.8477 12.5145 453       0
```

b) can use graphical displays: histogram, boxplot, density plot

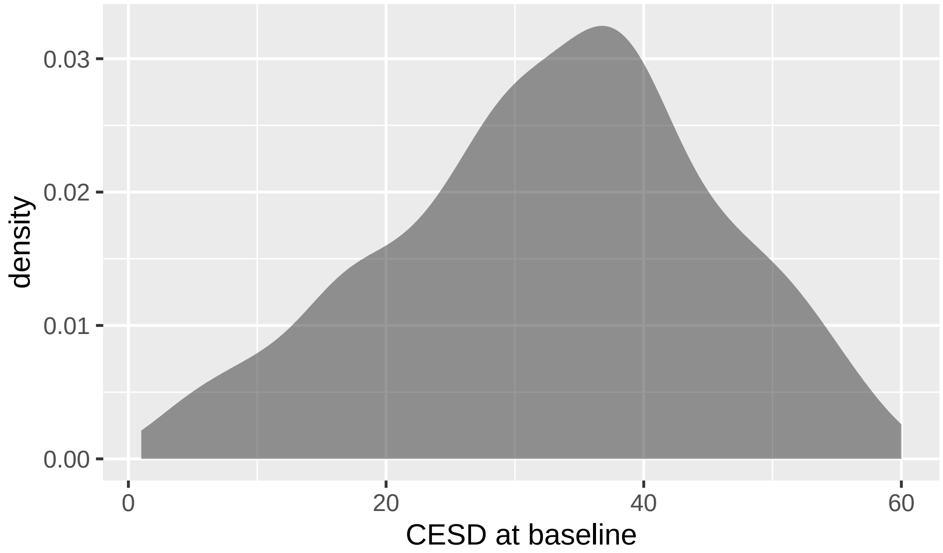
```
gf_boxplot (~cesd, data=HELPrct)
```



```
gf_histogram (~cesd, bins=10, data=HELPrct)
```



```
gf_density (~cesd, data=HELPrct)
```



BIVARIATE - examining two variables

The general form is:

command (Y ~ X, data=ds) where Y is the outcome and X is the explanatory variable, ds is the name of the dataset you are using.

2 Qualitative variables - see examples in the R chunk below

a) can use numerical summaries such as a two-way table

```
tally (sex ~ substance, data=HELPrct)
```

```
##         substance
## sex      alcohol cocaine heroin
##   female      36      41     30
##   male       141     111     94
```

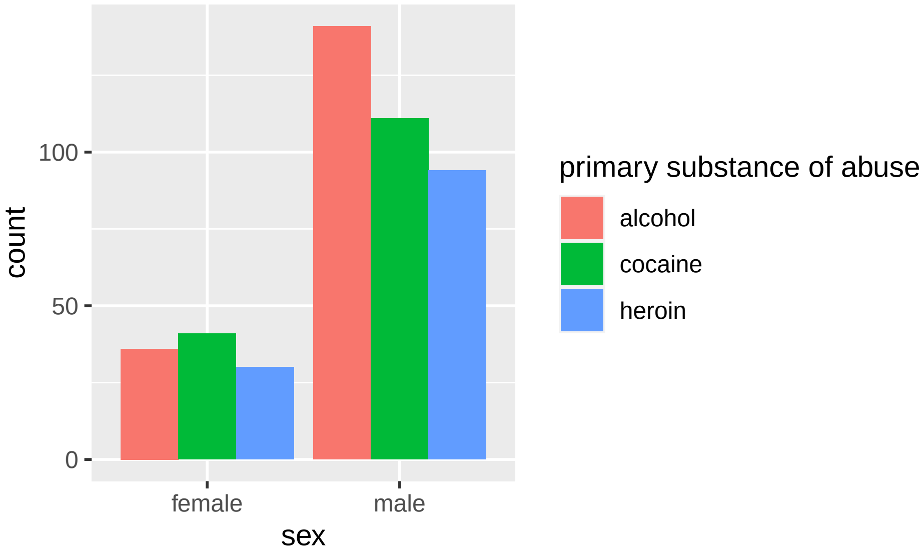
```
tally (sex ~ substance / sex, format="proportion", data=HELPrct)
```

```
## Warning in Ops.factor(substance, sex): '/' not meaningful for factors
```

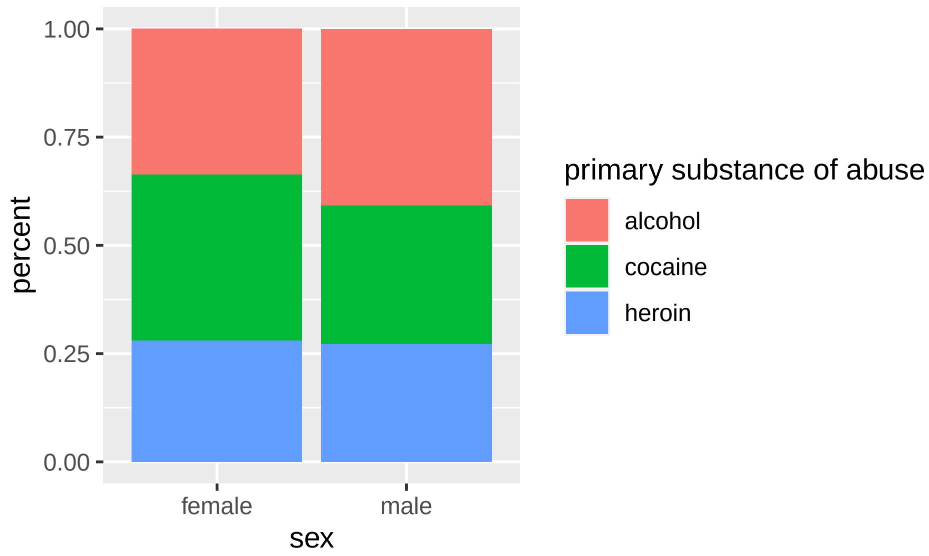
```
##         substance/sex
## sex      TRUE FALSE     <NA>
##   female                0.236203
##   male                  0.763797
```

b) can use graphical displays such as side-by-side or segmented bar chart

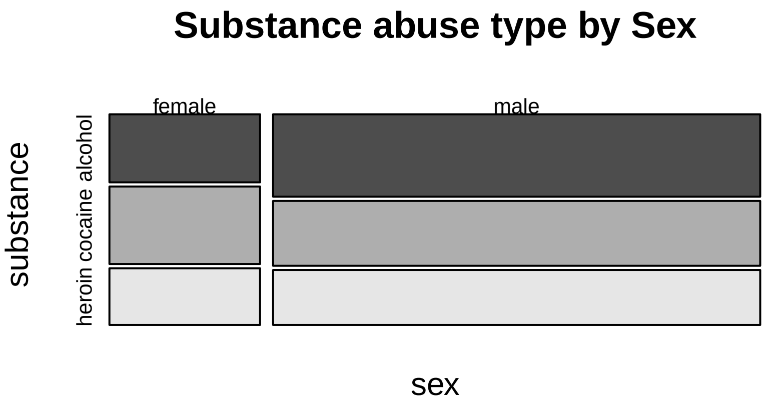
```
gf_counts(~sex, fill= ~substance, position=position_dodge(), data=HELPrct)
```



```
gf_percents(~sex, fill= ~substance, position="fill", data=HELPrct)
```



```
mosaicplot(sex ~ substance, color=TRUE, main="Substance abuse type by Sex", data=HELPrct)
```



1 Quantitative variable and 1 Qualititative var (grouping variable) - see examples in R chunk below

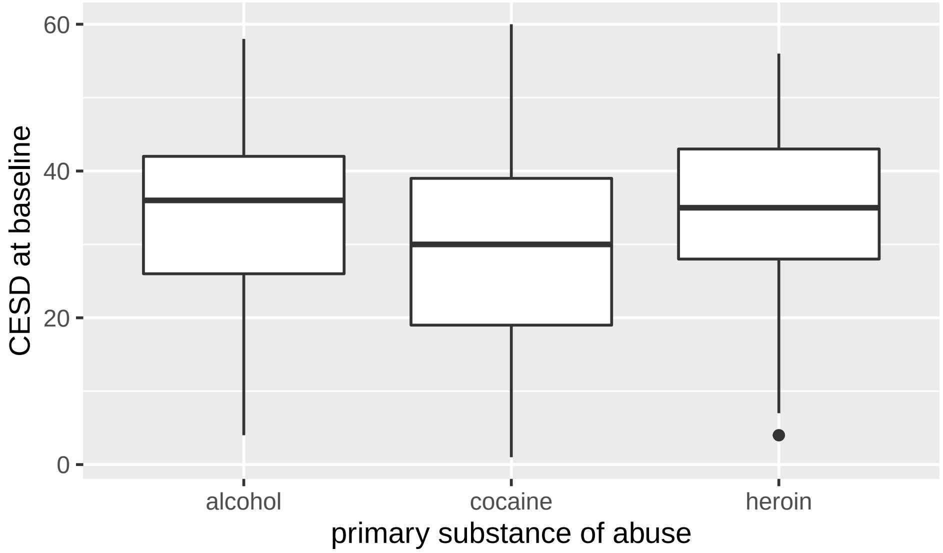
a) can use summary statistics such as favstats by group

```
favstats(cesd ~ substance, data=HELPrct)
```

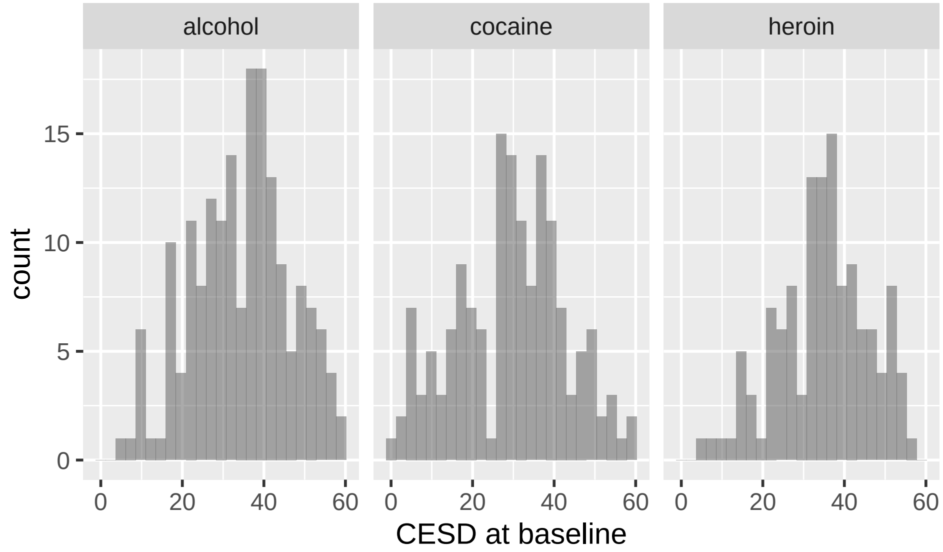
```
##   substance min Q1 median Q3 max    mean      sd   n missing
## 1   alcohol   4 26     36 42  58 34.3729 12.0504 177       0
## 2   cocaine   1 19     30 39   60 29.4211 13.3974 152       0
## 3    heroin   4 28     35 43   56 34.8710 11.1981 124       0
```

b) can use graphical displays such as side-by-side boxplots/histograms, overlaid densityplots

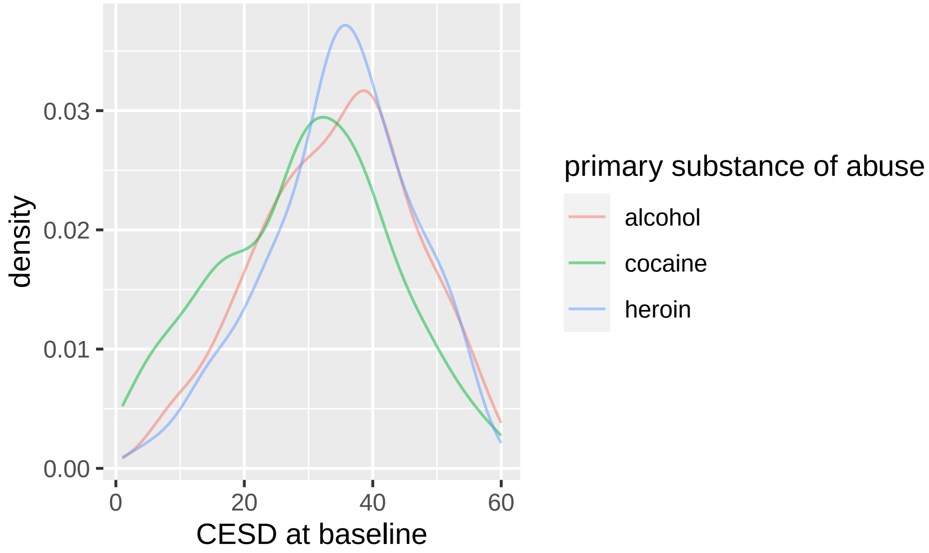
```
gf_boxplot(cesd ~ substance, data=HELPrct)
```



```
gf_histogram(~cesd | substance, data=HELPrct)
```



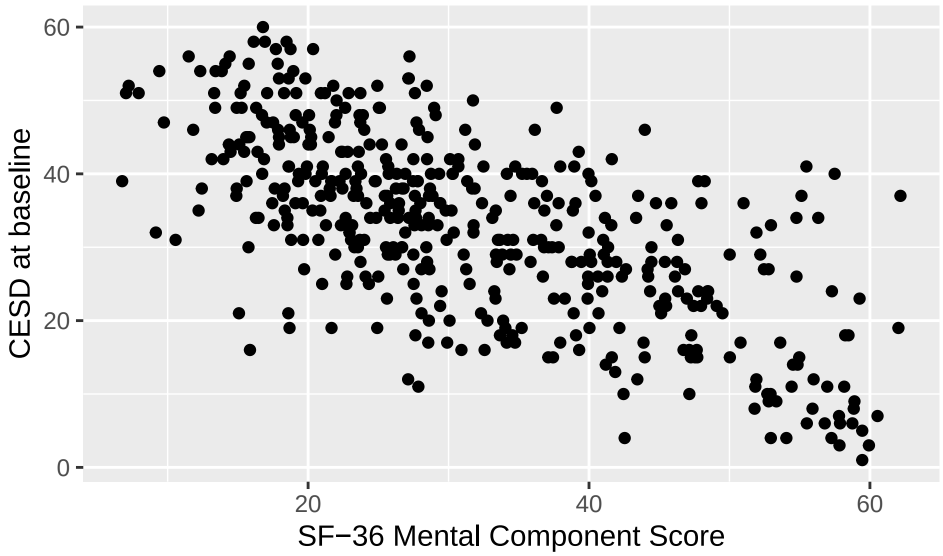
```
gf_dens(~cesd, color= ~substance, data=HELPrct)
```



2 Quantitative variables - see examples in the R chunk below

can use scatterplot, correlation, and simple linear regression (SLR)

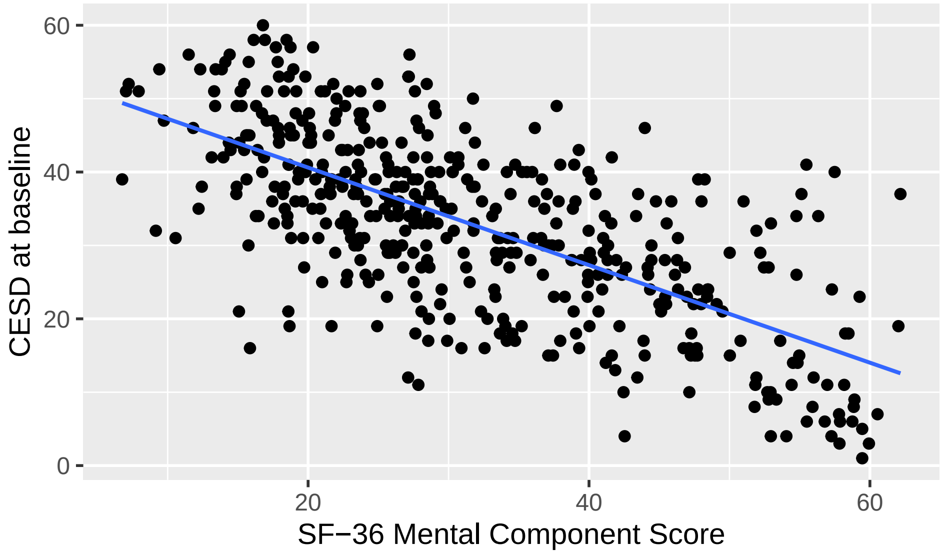
```
gf_point(cesd ~ mcs, data=HELPrct)
```



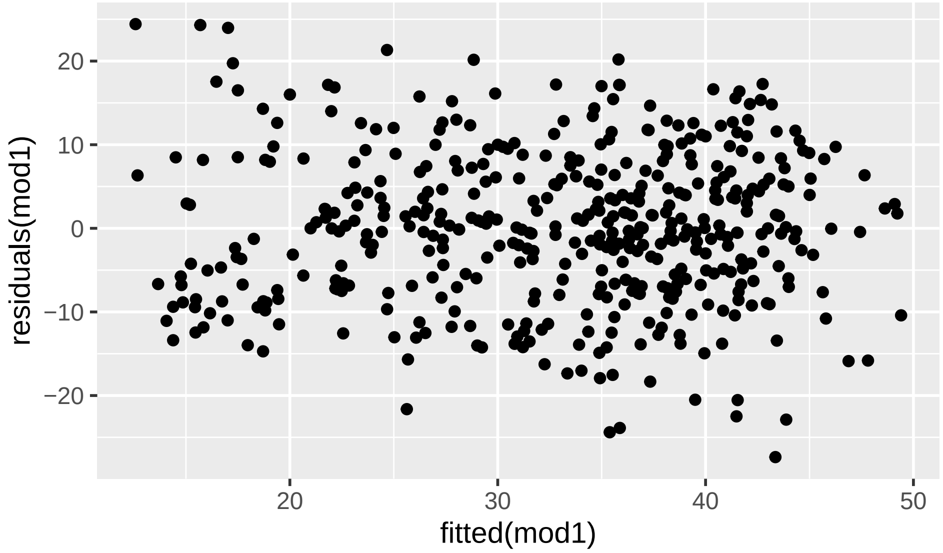
```
cor(cesd ~ mcs, data=HELPrct)
```

```
## [1] -0.681924
```

```
mod1 <- lm(cesd ~ mcs, data=HELPrct)
gf_point(cesd ~ mcs, data=HELPrct) %>%
   gf_lm()
```



```
gf_point(residuals(mod1) ~ fitted(mod1), data=HELPrct)
```



We will go more in depth about SLR in the next R activity.

MULTIVARIATE - examining 3 variables

The general form is:

command (Y ~ X | Z , data=ds) where Y is the outcome/response/dependent variable and X is the explanatory/predictor. Z denotes a third variable for which we want to get separate results for each level of Z.

3 Qualitative variables - see examples in the R chunk below

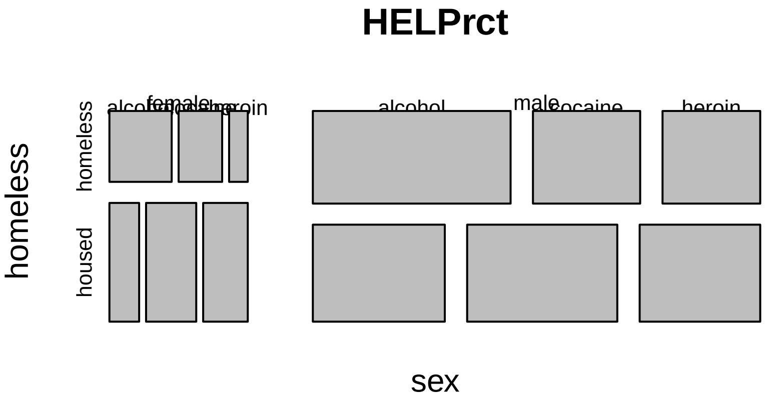
a) can use numerical summaries such as two 2-way tables

```
tally(homeless ~ substance|sex, data=HELPrct)
```

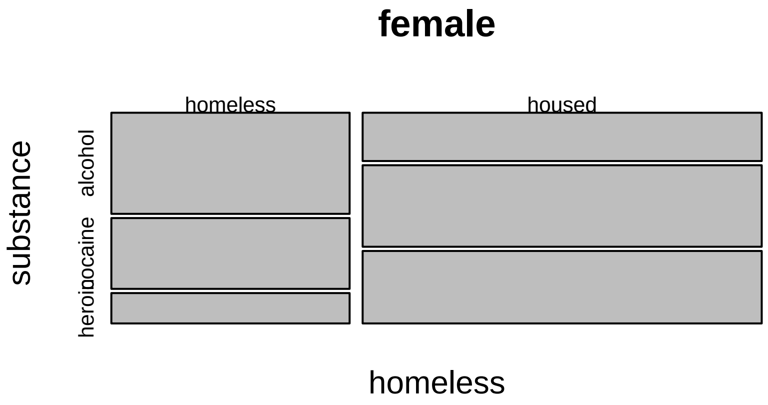
```
## , , sex = female
##
##             substance
## homeless   alcohol cocaine heroin
##   homeless      20      14      6
##   housed        16      27     24
##
## , , sex = male
##
##             substance
## homeless   alcohol cocaine heroin
##   homeless      83      45     41
##   housed        58      66     53
```

b) can use graphical displays such as a mosaic plot

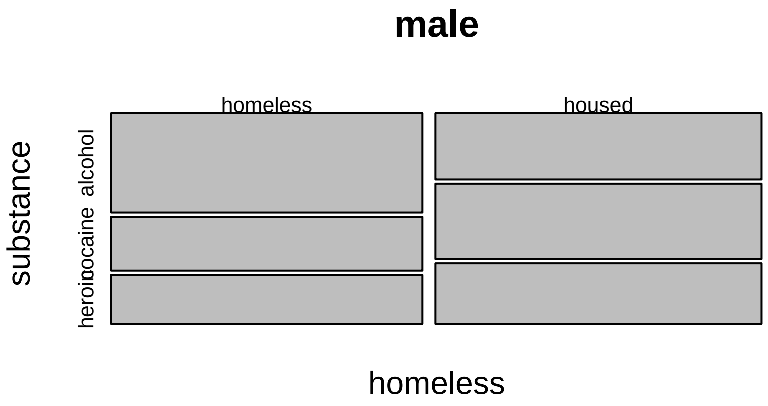
```
mosaicplot (~sex + homeless + substance, data=HELPrct)
```



```
female <- filter(HELPrct, sex == "female")
mosaicplot(~homeless + substance, data=female)
```



```
male <- filter(HELPrct, sex == "male")
mosaicplot(~homeless + substance, data=male)
```



1 Quantitative (outcome/response) and 2 Qualitative variables - see examples in the R chunk below

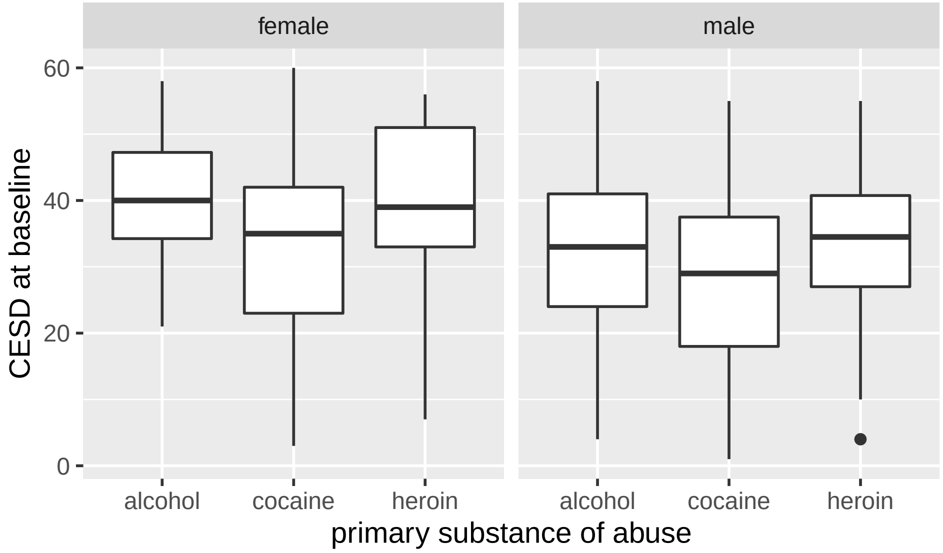
a) can use summary statistics such as favstats breaking down the quantitative variable by both the qualitative variables

```
favstats(cesd ~ substance + sex, data=HELPrct)
```

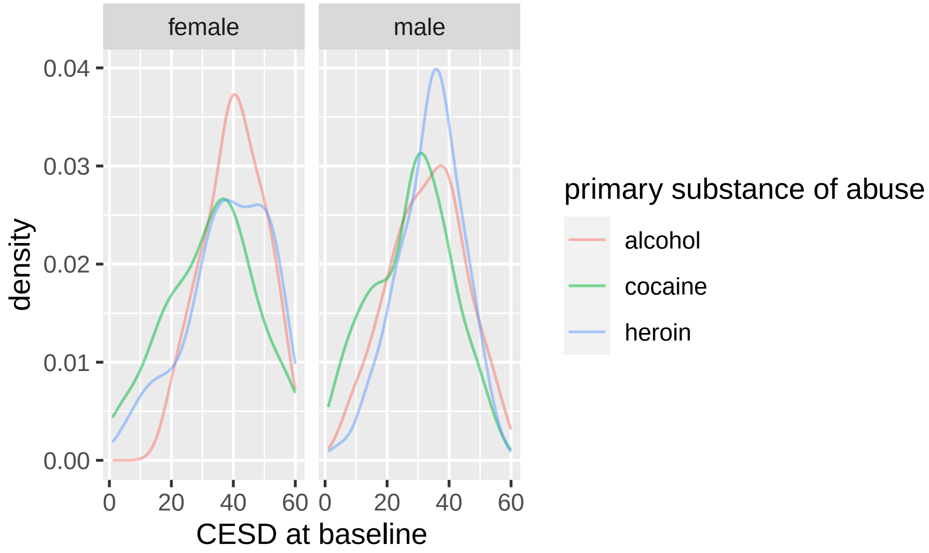
```
##     substance.sex min    Q1 median    Q3 max    mean       sd   n missing
## 1 alcohol.female  21 34.25   40.0 47.25  58 40.2778  9.80848  36       0
## 2 cocaine.female   3 23.00   35.0 42.00  60 32.9756 14.48704  41       0
## 3  heroin.female   7 33.00   39.0 51.00  56 38.1667 13.27451  30       0
## 4   alcohol.male   4 24.00   33.0 41.00  58 32.8652 12.13450 141       0
## 5   cocaine.male   1 18.00   29.0 37.50  55 28.1081 12.79158 111       0
## 6    heroin.male   4 27.00   34.5 40.75  55 33.8191 10.30916  94       0
```

b) can use graphical displays like boxplot or density plot of quantitative variable with different plots by for each level of the qualitative variable

```
gf_boxplot (cesd ~ substance | sex, data=HELPrct)
```



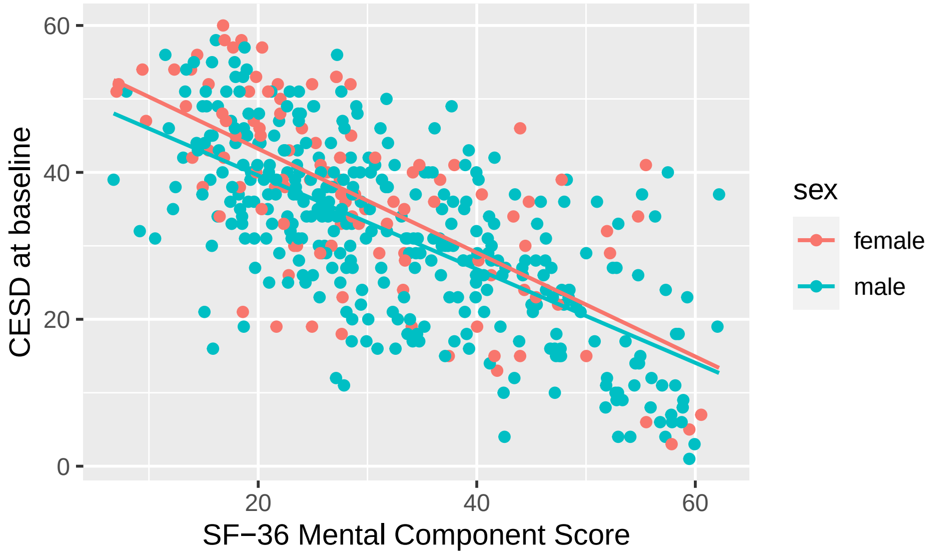
```
gf_dens(~cesd | sex, color= ~substance, data=HELPrct)
```



2 Quantitative variables and 1 Qualitative variables - see examples in the R chunk below

can use scatterplot of the two quantitative varibles differentiated by the qualitative variable

```
gf_point(cesd ~ mcs, color= ~ sex, data=HELPrct) %>%
   gf_lm()
```


CESD at baseline
SF−36 Mental Component Score
sex
female
male