Graphing in R with `gpplot2`

- First we begin by installing the graphing package `ggplot2`. As a reminder, you only need to run this line once, after which the package is permanently installed on your computer.

```
install.packages("ggplot2")
```

- Then you will need to load the library. This line must be run every time you open R and use the graphical commands within `ggplot2`.

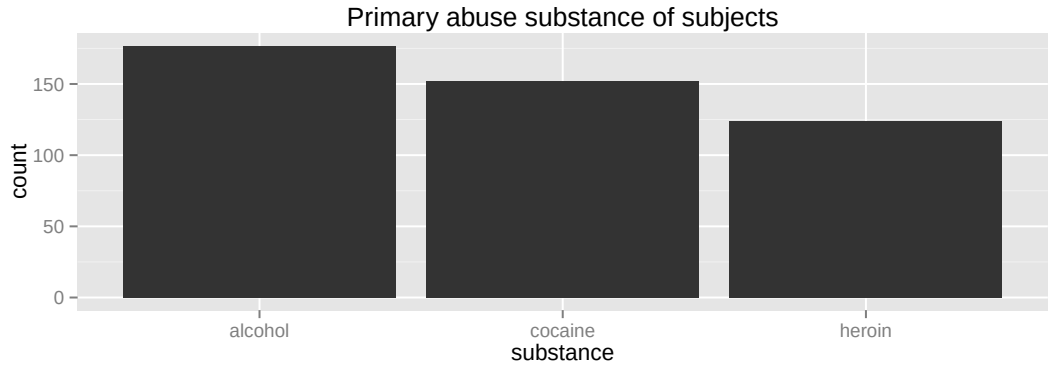
```
library("ggplot2")
```

- The data set used to illustrate the `ggplot2` commands is the HELP study (data name is `HELPrct`), 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 gender (`sex`).

Univariate Graphing

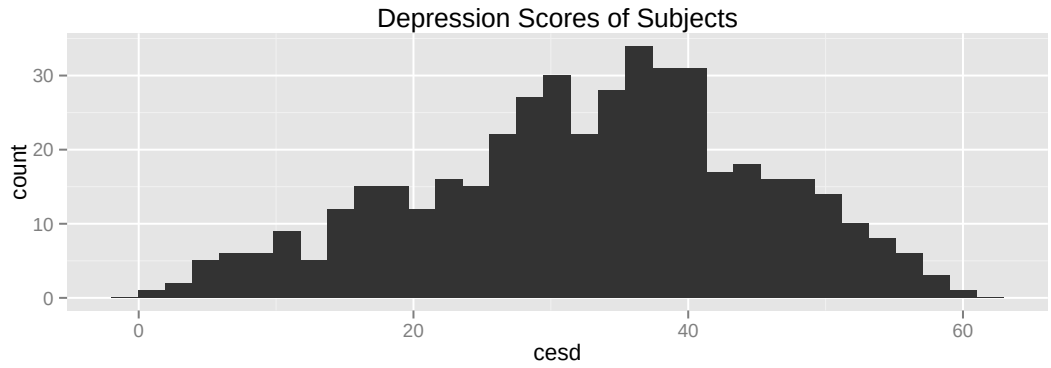
- Suppose we would like a plot of a single categorical variable.

```
ggplot(data=HELPrct)+
  geom_bar(aes(x=substance))+
  ggtitle("Primary abuse substance of subjects")
```



- Now for a plot of a single quantitative variable

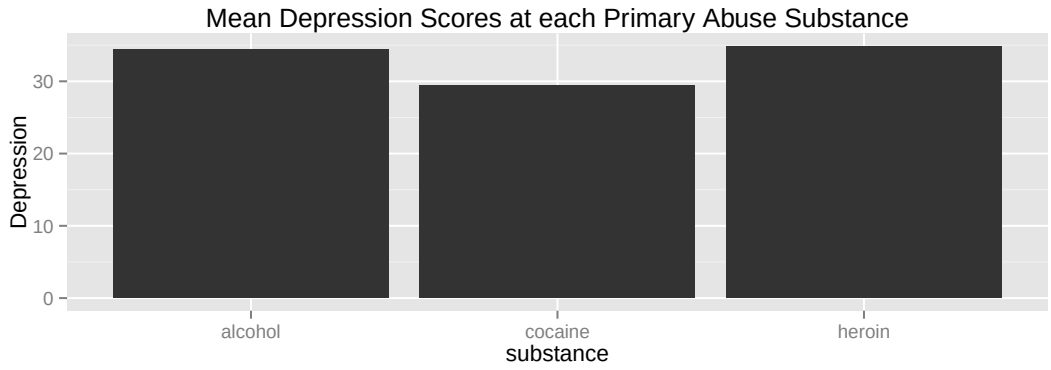
```
ggplot(data=HELPrct)+
  geom_histogram(aes(x=cesd))+
  ggtitle("Depression Scores of Subjects")
```



Bivariate Graphing

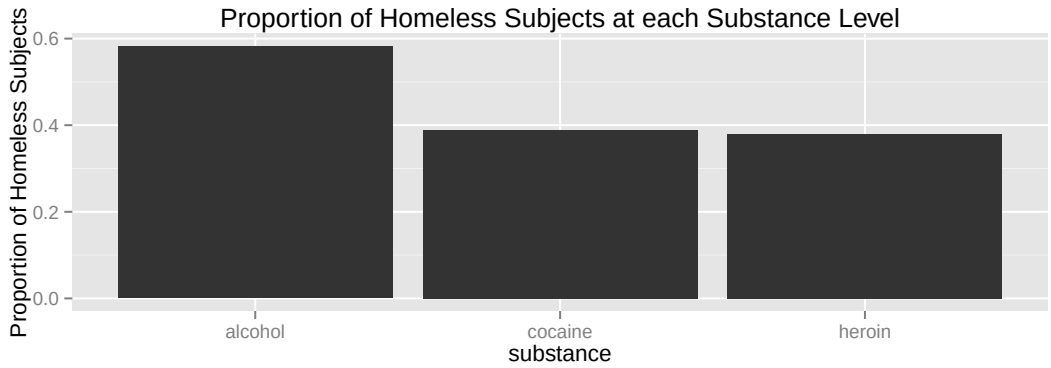
- Display the mean of one quantitative response variable based on a categorical explanatory variable.

```
ggplot(data=HELPrct)+
  stat_summary(aes(x=substance, y=cesd), fun.y=mean, geom="bar")+
  ylab("Depression")+
  ggtitle("Mean Depression Scores at each Primary Abuse Substance")
```



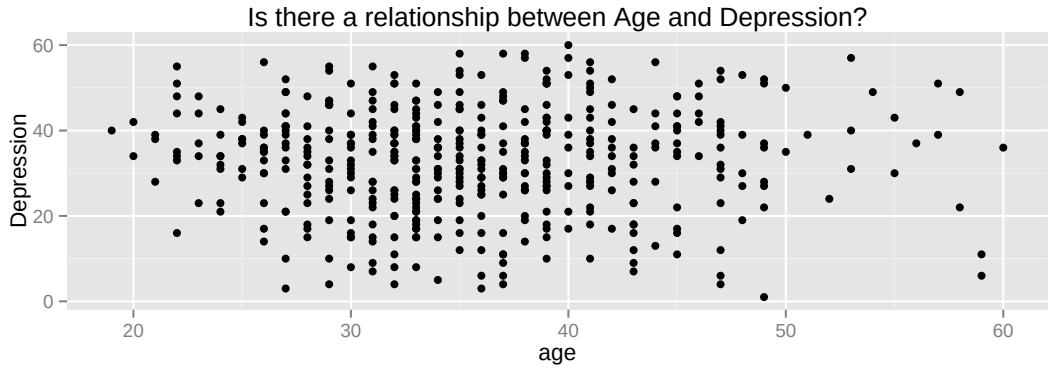
- Display the proportion of participants at a particular level of the response variable based on a categorical variable. (First we will need to create a dummy variable that corresponds to the level of the response variable of interest).

```
HELPrct$IsHomeless<-ifelse(HELPrct$homeless=="homeless",1,0)
ggplot(data=HELPrct)+
  stat_summary(aes(x=substance, y=IsHomeless), fun.y=mean, geom="bar")+
  ylab("Proportion of Homeless Subjects")+
  ggtitle("Proportion of Homeless Subjects at each Substance Level")
```



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

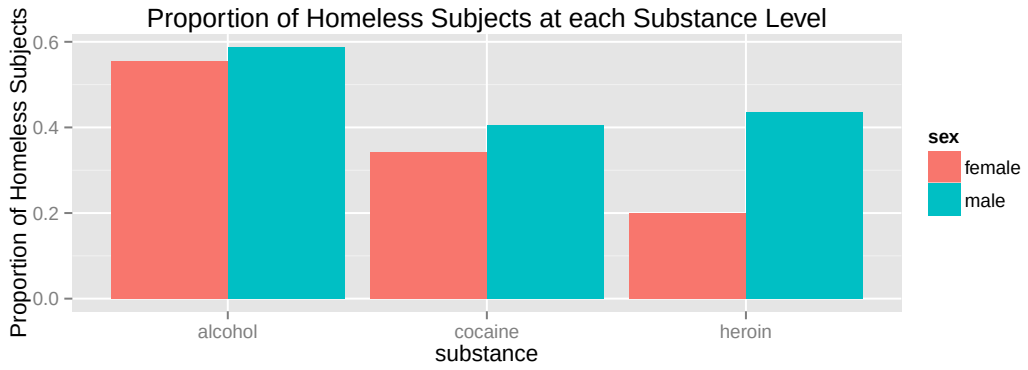
```
ggplot(data=HELPrct)+
  geom_point(aes(x=age, y=cesd))+
  ylab("Depression")+
  ggtitle("Is there a relationship between Age and Depression?")
```



Multivariate Graphing

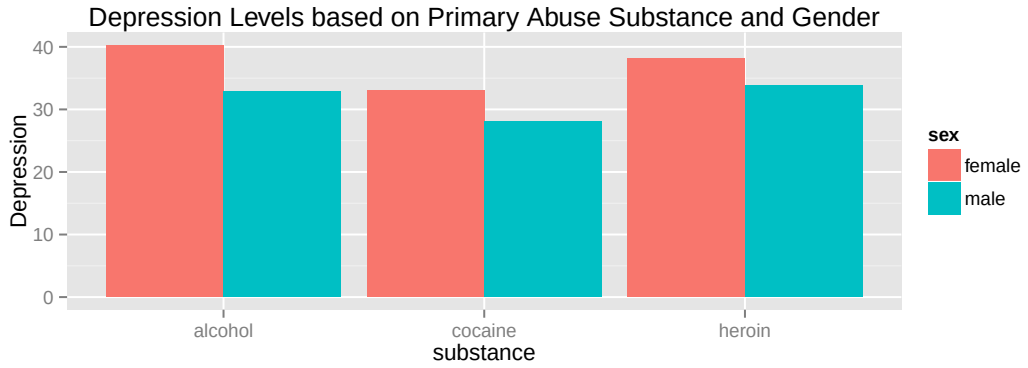
- Suppose we wish to visualize the relationship between two categorical variables, controlling for an additional categorical variable.

```
ggplot(data=HELPrct)+
  stat_summary(aes(x=substance, y=IsHomeless, fill=sex), fun.y=mean,
               geom="bar", position="dodge")+
  ylab("Proportion of Homeless Subjects")+
  ggtitle("Proportion of Homeless Subjects at each Substance Level")
```



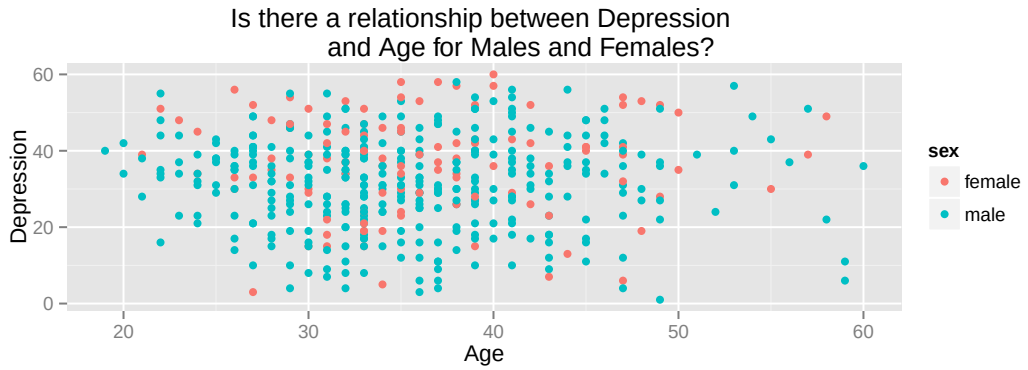
- Suppose we wish to visualize the relationship between a categorical explanatory variable and a quantitative response variable, controlling for an additional categorical variable.

```
ggplot(data=HELPrct)+
  stat_summary(aes(x=substance, y=cesd, fill=sex), fun.y=mean,
               geom="bar", position="dodge")+
  ylab("Depression")+
  ggtitle("Depression Levels based on Primary Abuse Substance and Gender")
```



- Suppose we wish to visualize the relationship between two quantitative variables, controlling for a categorical variable.

```
ggplot(data=HELPrct)+
  geom_point(aes(x=age, y=cesd, color=sex))+
  ylab("Depression")+
  xlab("Age")+
  ggtitle("Is there a relationship between Depression
          and Age for Males and Females?")
```



- Suppose we wish to add separate regression lines for each gender. We just need to add the function `geom_smooth` to our previous plot with the command `method="lm"` to denote that we want an overlaid regression line (opposed to a smoothed curve).

```
ggplot(data=HELPrct)+
  geom_point(aes(x=age, y=cesd, color=sex))+
  geom_smooth(aes(x=age, y=cesd, color=sex), method="lm")+
  ylab("Depression")+
  xlab("Age")+
  ggtitle("Is there a relationship between Depression
          and Age for Males and Females?")
```