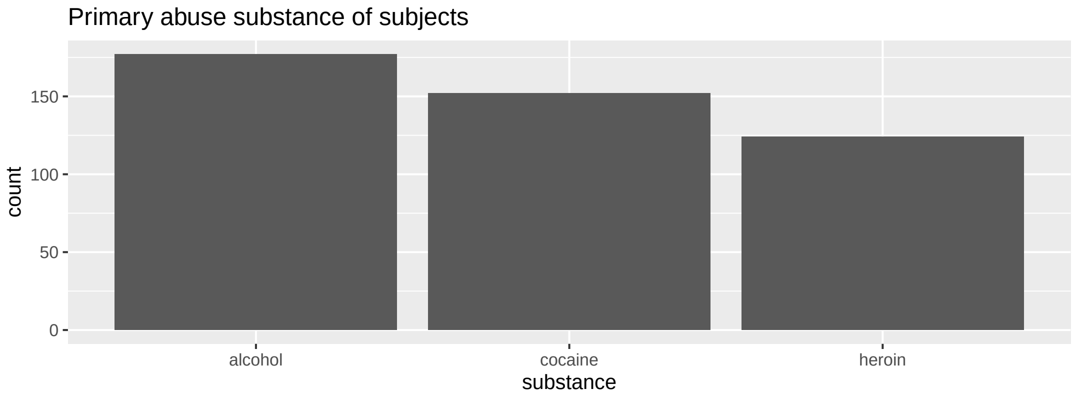
Graphing in R with gpplot2

- 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

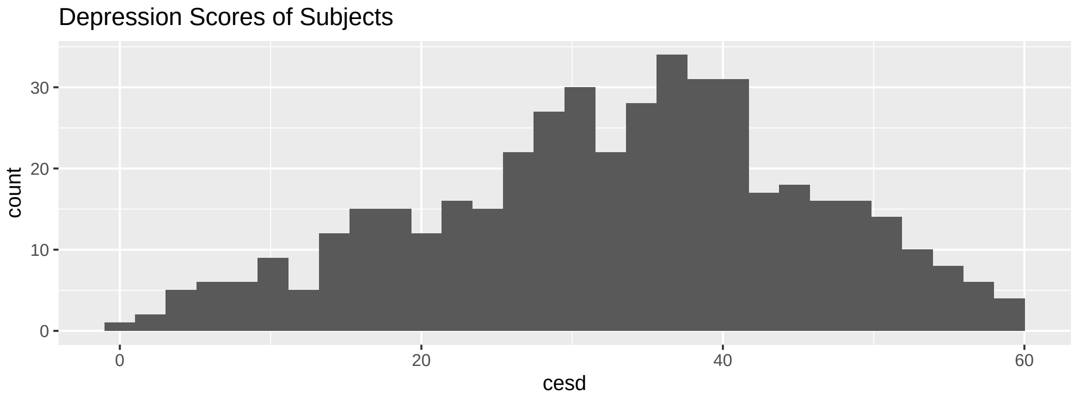
- Suppose we would like a plot of a single categorical variable.

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

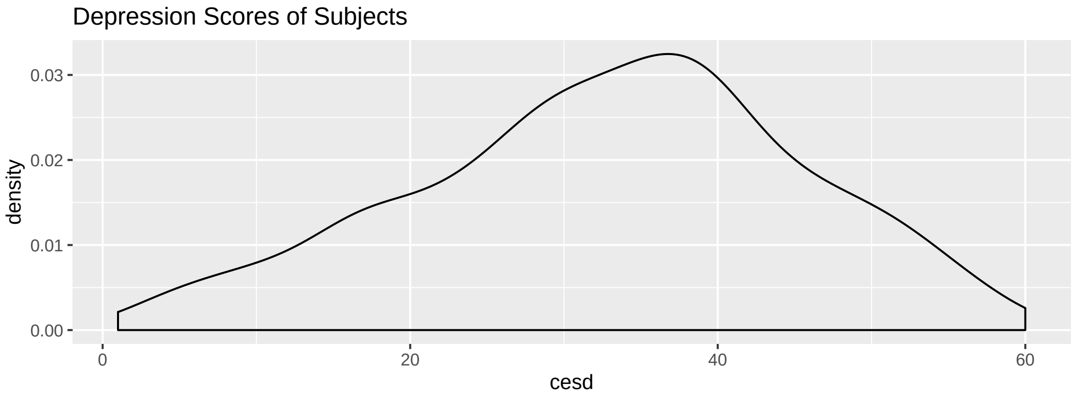


- Now for a plot of a single quantitative variable

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



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

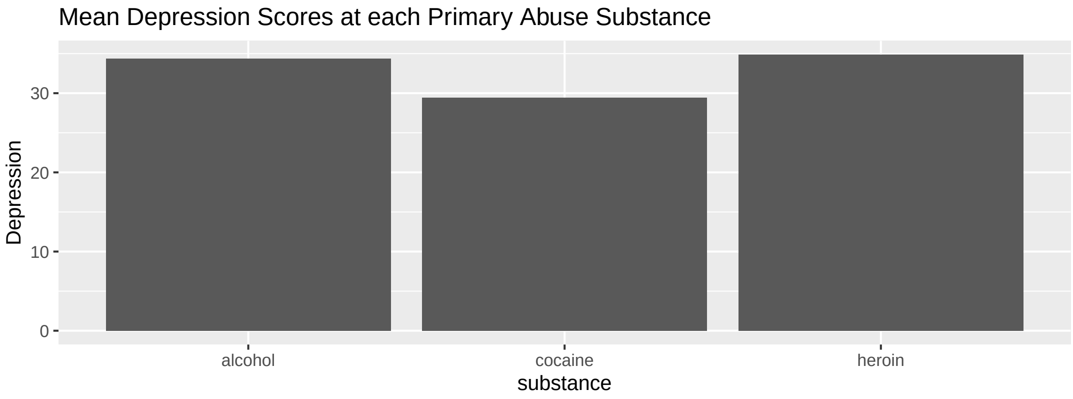


Bivariate Graphing

C→ Q

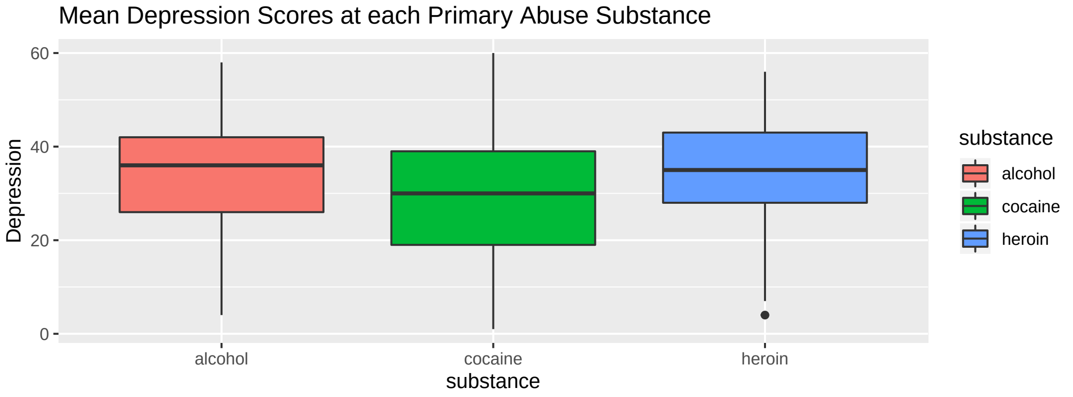
- OPTION 1: Construct a bar plot with mean of response variable on y-axis.

```
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")
```



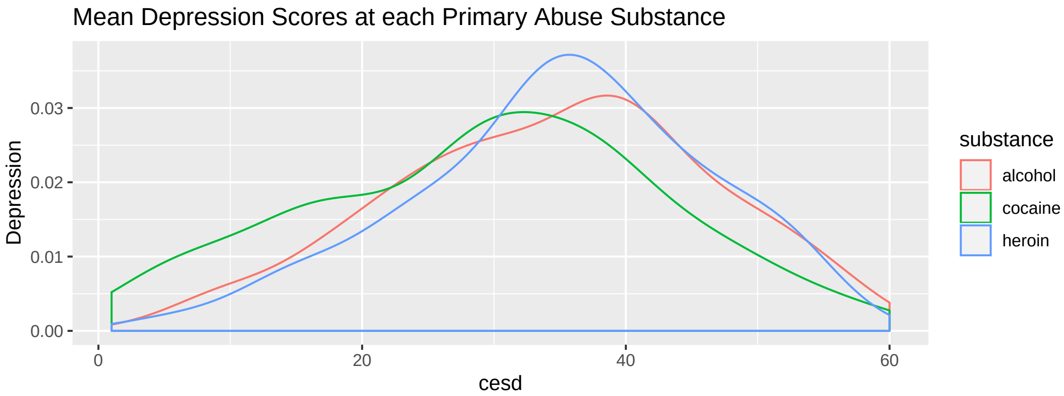
- OPTION 2: Boxplots

```
ggplot(data=HELPrct)+
  geom_boxplot(aes(x=substance, y=cesd, fill=substance))+
  ylab("Depression")+
  ggtitle("Mean Depression Scores at each Primary Abuse Substance")
```



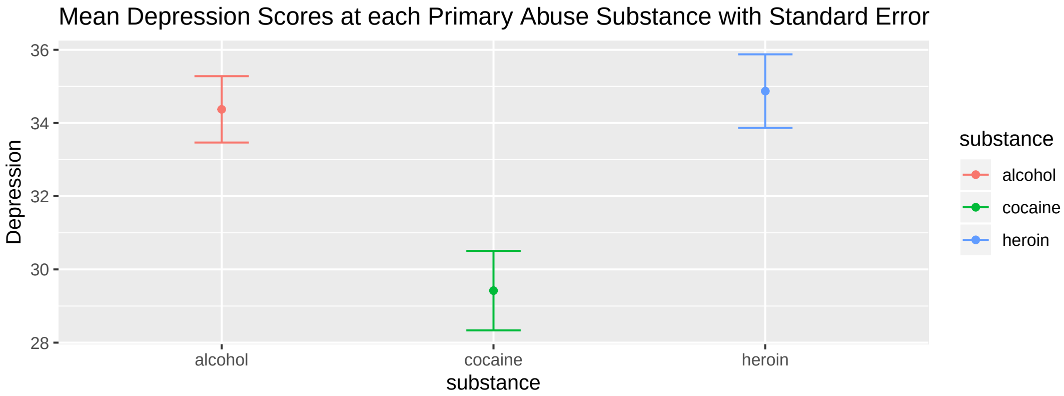
- OPTION 3: Density Plots

```
ggplot(data=HELPrct)+
  geom_density(aes(x=cesd, color=substance))+
  ylab("Depression")+
  ggtitle("Mean Depression Scores at each Primary Abuse Substance")
```



- OPTION 4: Mean of Response with Error Bars

```
ggplot(data=HELPrct)+
  stat_summary(aes(x=substance, y=cesd, color=substance),
               fun.data="mean_se", geom="errorbar", width=0.2)+
  stat_summary(aes(x=substance, y=cesd, color=substance),
               fun.y="mean", geom="point")+
  ylab("Depression")+
  ggtitle("Mean Depression Scores at each Primary Abuse Substance with Standard Error")
```

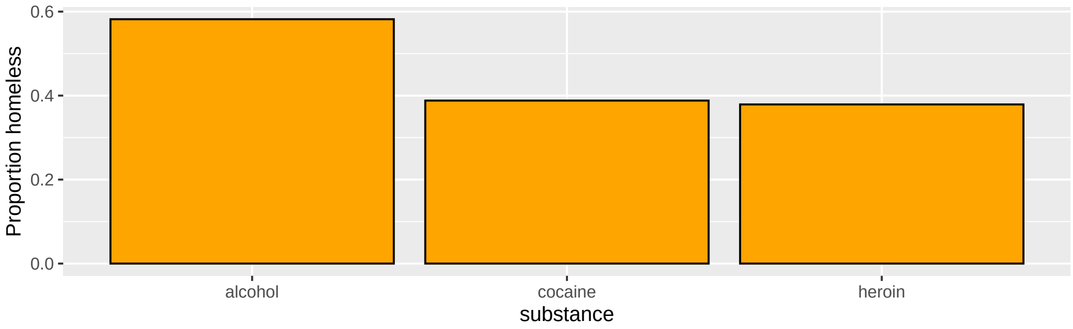


C→ C

- 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.

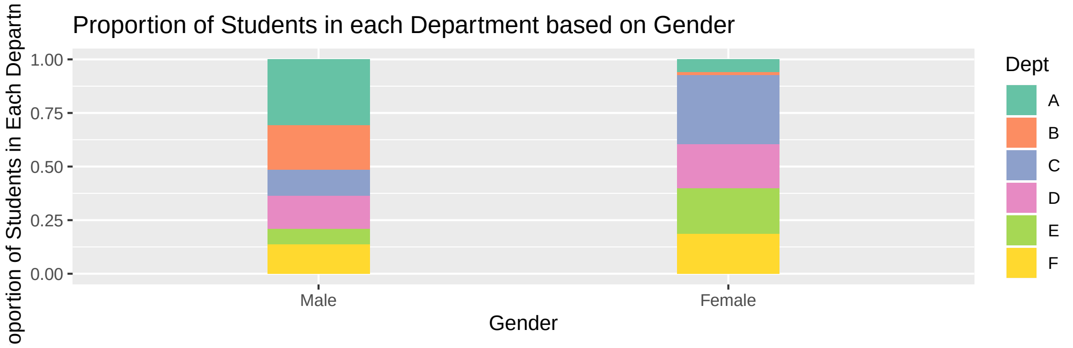
```
HELPrct$homeless_status[HELPrct$homeless=="homeless"]<-1
HELPrct$homeless_status[HELPrct$homeless=="housed"]<-0

ggplot(data=HELPrct)+
  stat_summary(aes(x=substance, y=homeless_status),
               fun.y="mean", geom="bar", fill="orange",
               color="black")+
  ylab("Proportion homeless")
```



- 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.)

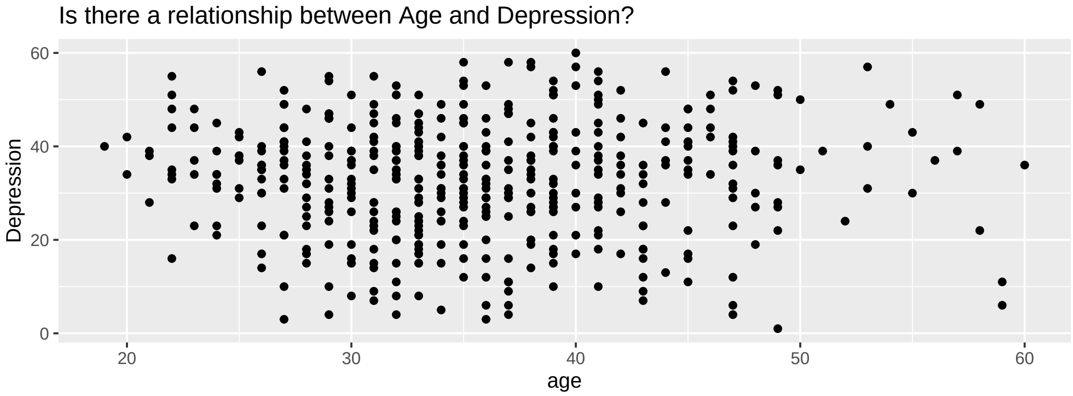
```
ggplot(data=UCBAdmissions)+
  geom_bar(aes(x=Gender, fill=Dept), position="fill", width=0.25)+
  ylab("Proportion of Students in Each Department")+
  ggtitle("Proportion of Students in each Department based on Gender")+
  scale_fill_brewer(palette="Set2")
```



Q→ Q

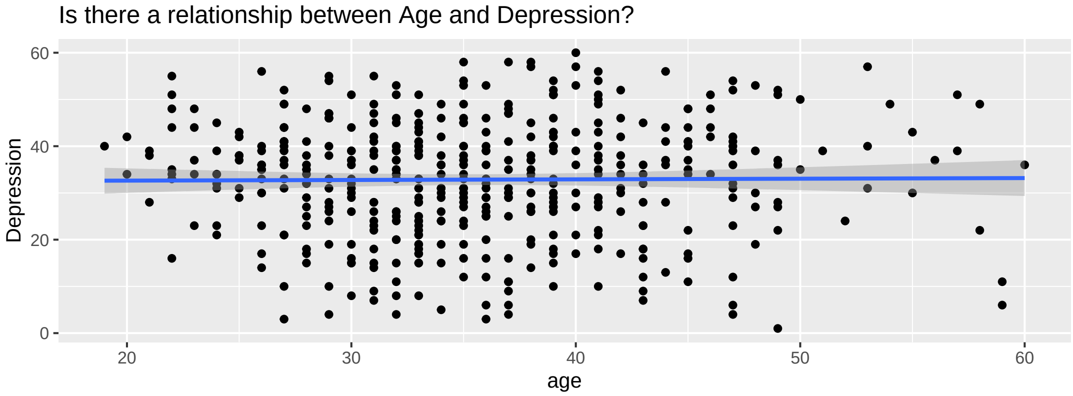
- Now, let's look at an explanatory and response variable which are both quantitative.

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



- 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.

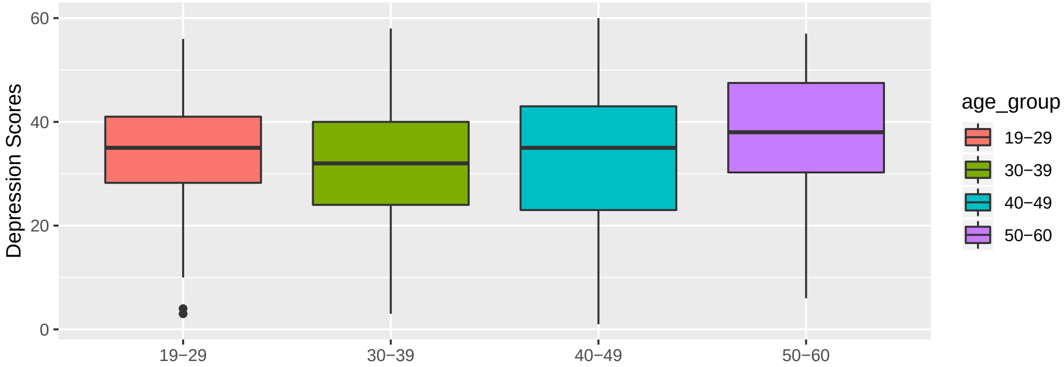
```
ggplot(data=HELPrct)+
  geom_point(aes(x=age, y=cesd))+
  geom_smooth(aes(x=age, y=cesd), method="lm")+
  ylab("Depression")+
  ggtitle("Is there a relationship between Age and Depression?")
```



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

```
HELPrct$age_group[HELPrct$age<30]<-"19-29"
HELPrct$age_group[HELPrct$age<40&HELPrct$age>=30]<-"30-39"
HELPrct$age_group[HELPrct$age<50&HELPrct$age>=40]<-"40-49"
HELPrct$age_group[HELPrct$age<=60&HELPrct$age>=50]<-"50-60"

ggplot(data=HELPrct)+
  geom_boxplot(aes(x=age_group, y=cesd, fill=age_group))+
  xlab("")+
  ylab("Depression Scores")
```



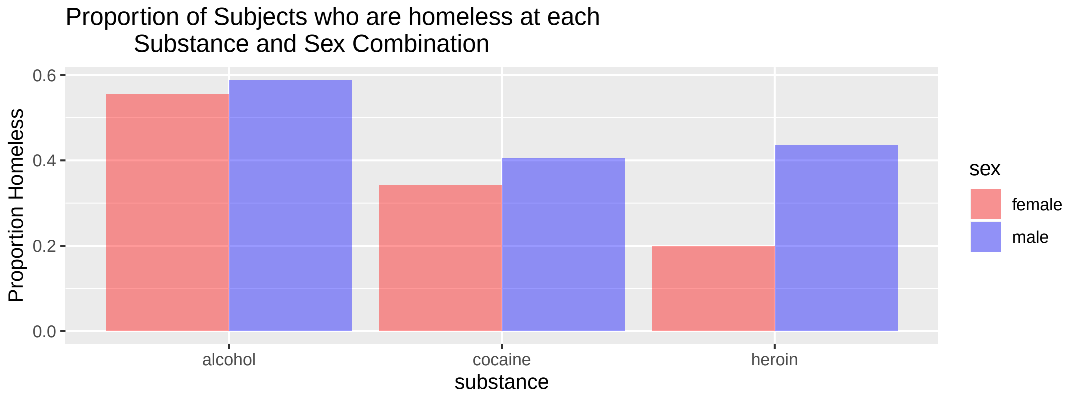
Multivariate Graphing

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.

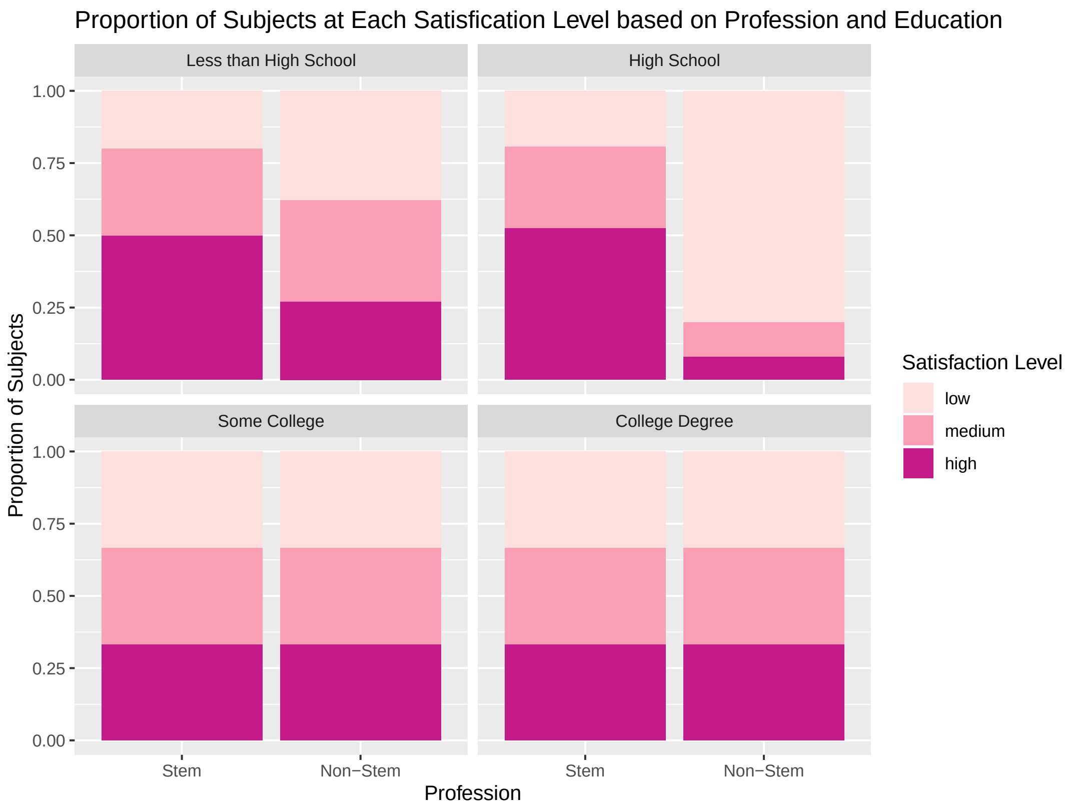
```
HELPrct$homeless_status[HELPrct$homeless=="homeless"]<-1
HELPrct$homeless_status[HELPrct$homeless=="housed"]<-0


ggplot(data=HELPrct)+
  stat_summary(aes(x=substance, fill=sex, y=homeless_status),
               fun.y="mean", geom="bar", position="dodge", alpha=0.4)+
  ylab("Proportion Homeless")+
  ggtitle("Proportion of Subjects who are homeless at each
          Substance and Sex Combination")+
  scale_fill_manual(values=c("red","blue"))
```



- **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.

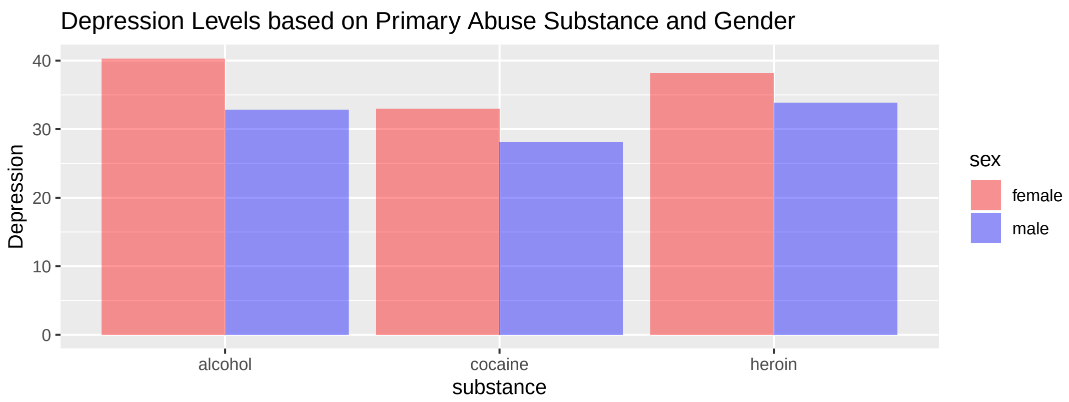
```
ggplot(data=job_data)+
  geom_bar(aes(x=Profession, fill=Satisfaction), position="fill")+
  facet_wrap(~Education)+
  ylab("Proportion of Subjects")+
  ggtitle("Proportion of Subjects at Each Satisfication Level based on Profession and Education")+
  scale_fill_brewer("Satisfaction Level",
                    palette="RdPu")
```



C→ Q with Categorical Third Variable

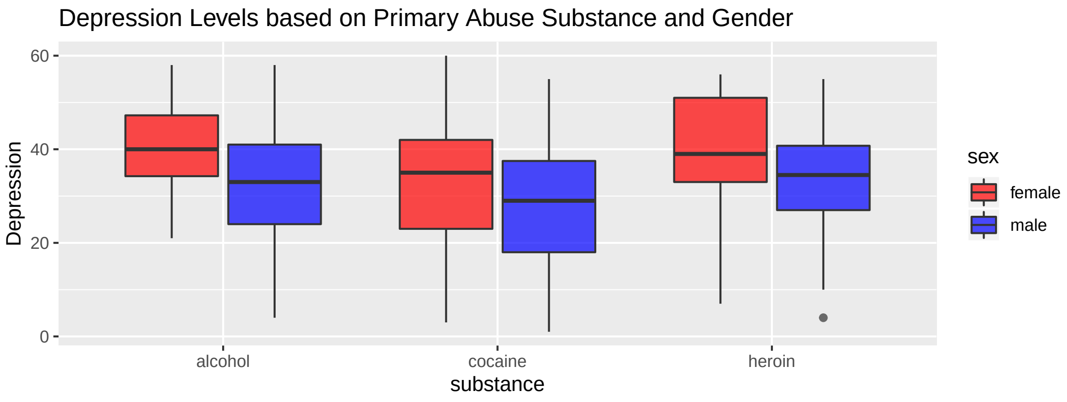
- Here, I wish to understand how abuse substance (explanatory variable) relates to depression of patients (response variable). In addition, I wish to understand how sex may play a role in this relationship.
- OPTION 1:

```r
ggplot(data=HELPrct)+
  stat_summary(aes(x=substance, y=cesd, fill=sex), fun.y=mean,
               geom="bar", position="dodge",
               alpha=0.4)+
  ylab("Depression")+
  ggtitle("Depression Levels based on Primary Abuse Substance and Gender")+
  scale_fill_manual(values=c("red","blue"))
```



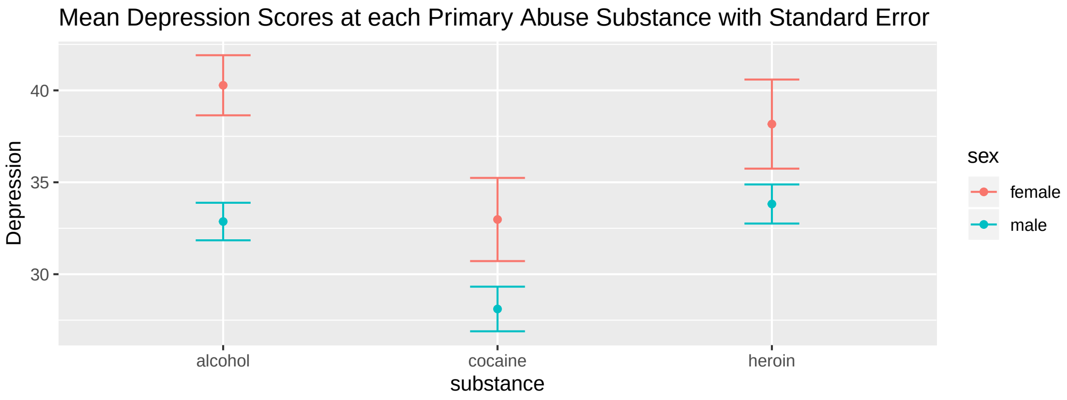
- OPTION 2:

```
ggplot(data=HELPrct)+
  geom_boxplot(aes(x=substance, y=cesd, fill=sex), alpha=0.7)+
  ylab("Depression")+
  ggtitle("Depression Levels based on Primary Abuse Substance and Gender")+
  scale_fill_manual(values=c("red","blue"))
```



- OPTION 3:

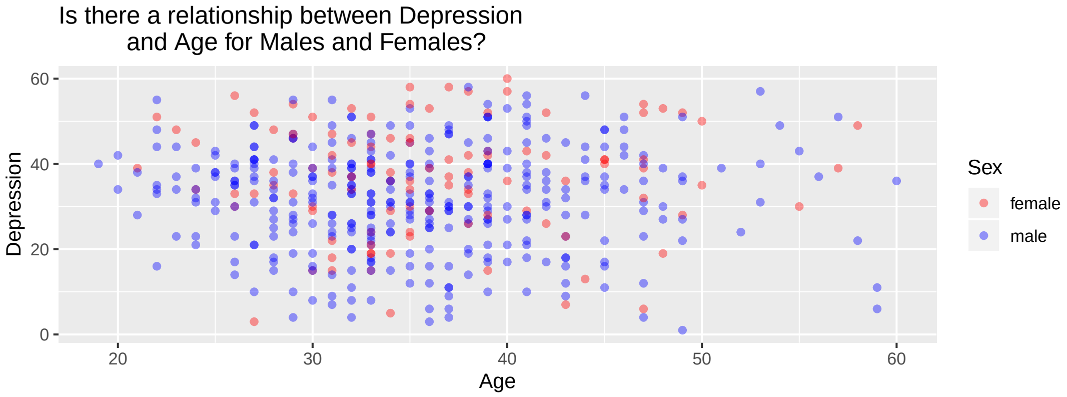
```
ggplot(data=HELPrct)+
  stat_summary(aes(x=substance, y=cesd, color=sex),
               fun.data="mean_se", geom="errorbar", width=0.2)+
  stat_summary(aes(x=substance, y=cesd, color=sex),
               fun.y="mean", geom="point")+
  ylab("Depression")+
  ggtitle("Mean Depression Scores at each Primary Abuse Substance with Standard Error")
```



Q→ Q with Categorical Third Variable

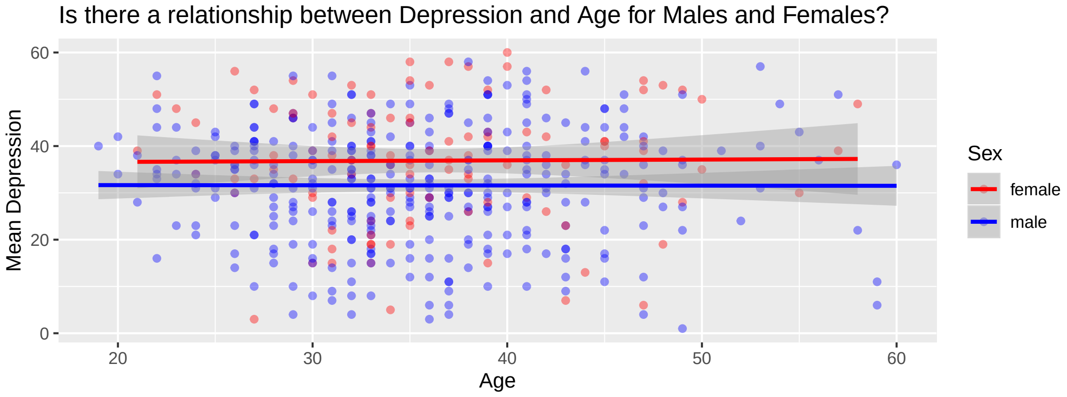
- Suppose we wish to visualize the relationship between age (explanatory variable) and cesd (a measure of depression and the response variable in this study) based on sex.
- OPTION 1:

```
ggplot(data=HELPrct)+
  geom_point(aes(x=age, y=cesd, color=sex), alpha=0.4)+
  ylab("Depression")+
  xlab("Age")+
  ggtitle("Is there a relationship between Depression
          and Age for Males and Females?")+
  scale_color_manual("Sex",values=c("red","blue"))
```



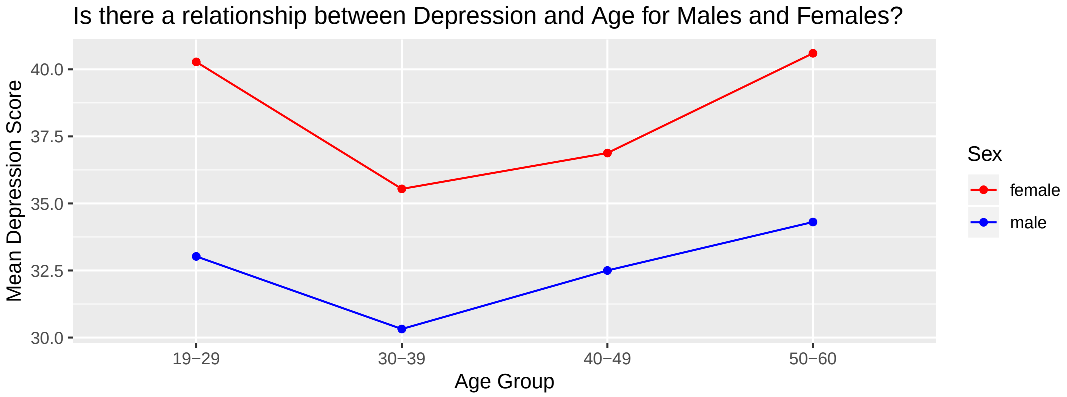
- OPTION 2: Suppose we wish to add separate regression lines for each sex. We just need to add the function `geom_smooth` to our previous plot with the command `method="lm"` to denote that we want an overlaid regression line (opposed to a smoothed curve).

```
ggplot(data=HELPrct)+
  geom_point(aes(x=age, y=cesd, color=sex), alpha=0.4)+
  geom_smooth(aes(x=age, y=cesd, color=sex), method="lm")+
  ylab("Mean Depression")+
  xlab("Age")+
  ggtitle("Is there a relationship between Depression and Age for Males and Females?")+
 scale_color_manual("Sex",values=c("red","blue"))
```



- 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.

```
ggplot(data=HELPrct)+
  stat_summary(aes(x=age_group, color=sex, y=cesd),
               fun.y="mean", geom="point")+
  stat_summary(aes(x=age_group, color=sex, y=cesd, group=sex),
               fun.y="mean", geom="line")+
  ylab("Mean Depression Score")+
  xlab("Age Group")+
  ggtitle("Is there a relationship between Depression and Age for Males and Females?")+
  scale_color_manual("Sex",values=c("red","blue"))
```


Is there a relationship between Depression and Age for Males and Females?
Mean Depression Score
Age Group
Sex
female
male