Foundations of Data Analysis

The University of Texas at Austin

R Tutorials: Week 4

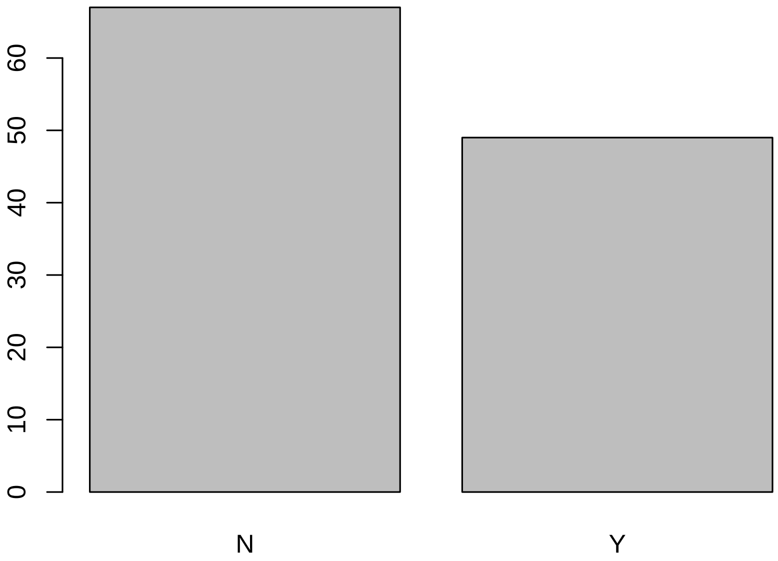
Grouped Bar Charts

In this R tutorial, we're going to show you how to create a grouped or stacked bar chart to visualize the distribution across two categorical variables. Similar to the last video, we're going to be using the Austin City Limits data set, which I've already imported into my workspace or environment. And we're going to be looking at the Grammy and gender variables just like we did before. So from the last tutorial, we created a frequency table of the Grammy variable by itself and then also a contingency table of counts across the Grammy variable as well as the gender variable.

```
gtab <- table(acl$Grammy)
gtab2 <- table(acl$Grammy, acl$Gender)
```

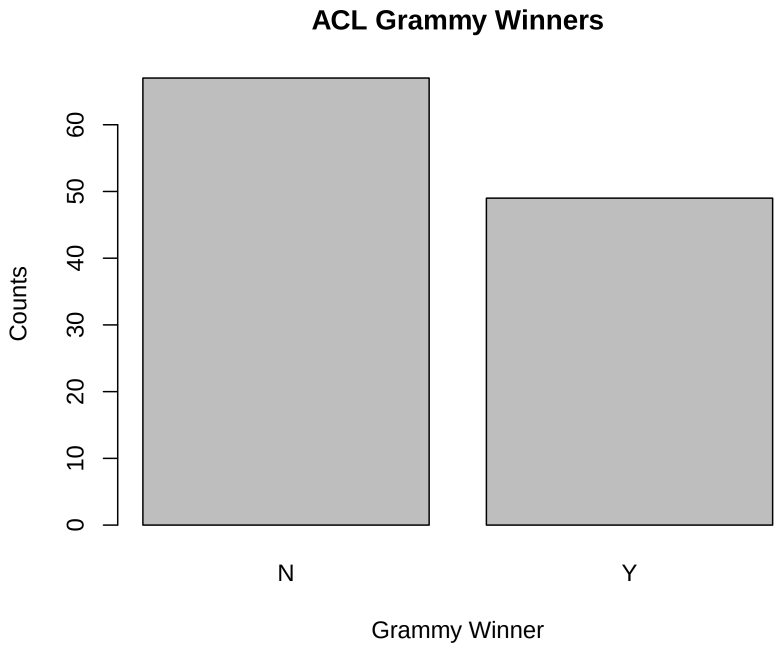
So just to review how we make a bar chart across one variable, we can ask R to give us a bar chart with the "**barplot()**" function. And then we just need to give it a single categorical variable. So in this case, my table of the Grammy variable was called gtab. So if I just ask for a **barchart()** of gtab, I'm going to get a very dry looking frequency graph of how many yeses and no's there were in our data set for this Grammy variable.

```
barplot(gtab)
```



So similar to what we've done before, we're going to want to make our graph a little bit nicer by adding a main title, ACL Grammy winners, an x label for x-axis which will appear under the yes and no categories, and then also a y label to show so these are the counts or frequency of occurrences in our data set.

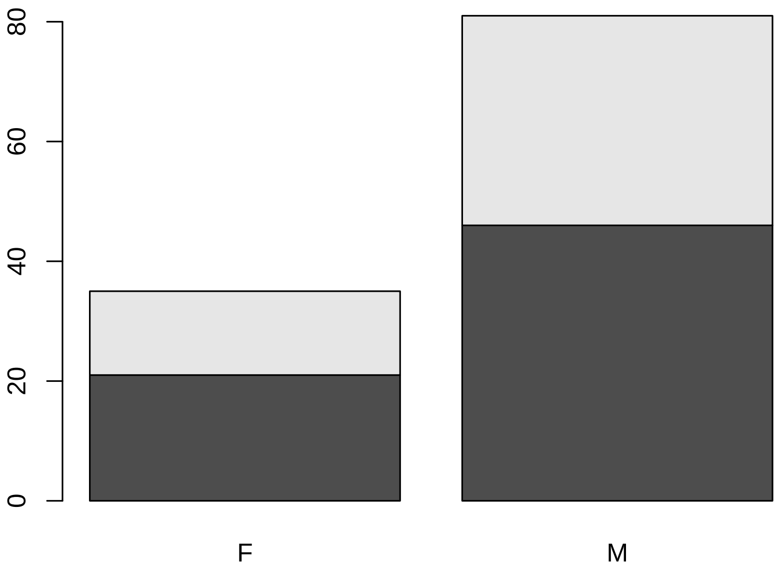
```
barplot(gtab, main = "ACL Grammy Winners", xlab = "Grammy Winner",
    ylab = "Counts")
```



So I've run that now we get a little bit nicer output with some titles and axis labels.

Now if we want to visualize two categorical variables at the same time, we can make either a stacked or side by side bar plot. So similar to what we did with one variable, I'm going to use the **barplot()** function. Instead of giving it the frequency table of one categorical variable, Grammy, I'm going to actually give it R gtab2, which was the contingency table of a Grammy and gender. So let's see what this gives us.

```r
barplot(gtab2)
```



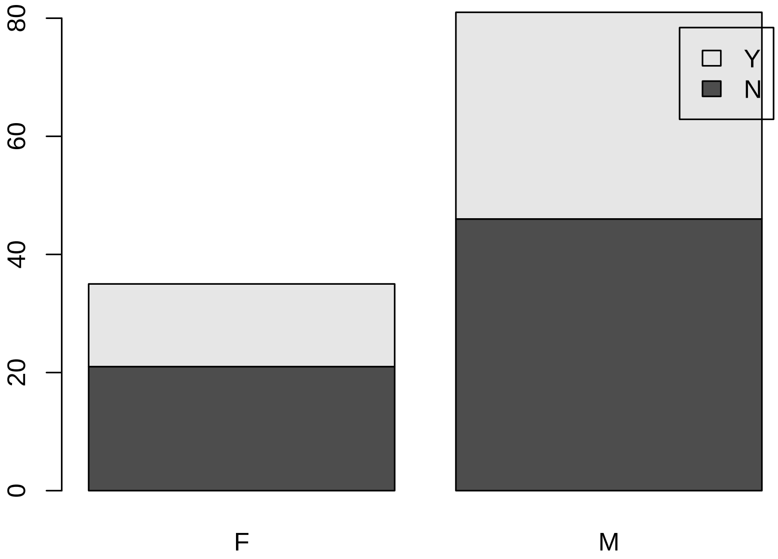
Now we can see R did a couple things. It put the gender variable along the x-axis now. So we see males and females. And then it colored the bars and kind of stacked them based off of the response to Grammy. So if we look at R gtab2, we're going to see that there were 21 no's and 14 yeses for females, which adds up here based off of our count on our y-axis. And then there were 46 no's and 35 yeses for the males.

```
gtab2
```

```
##
##      F  M
##   N 21 46
##   Y 14 35
```

Unfortunately, our bar chart here isn't showing us which color corresponds to which value of Grammy that we're on. So we're going to need a legend for our bar chart. And we can add that as an option in our bar plot function by saying legend equals true, or just T. It should turn light blue for you in RStudio.

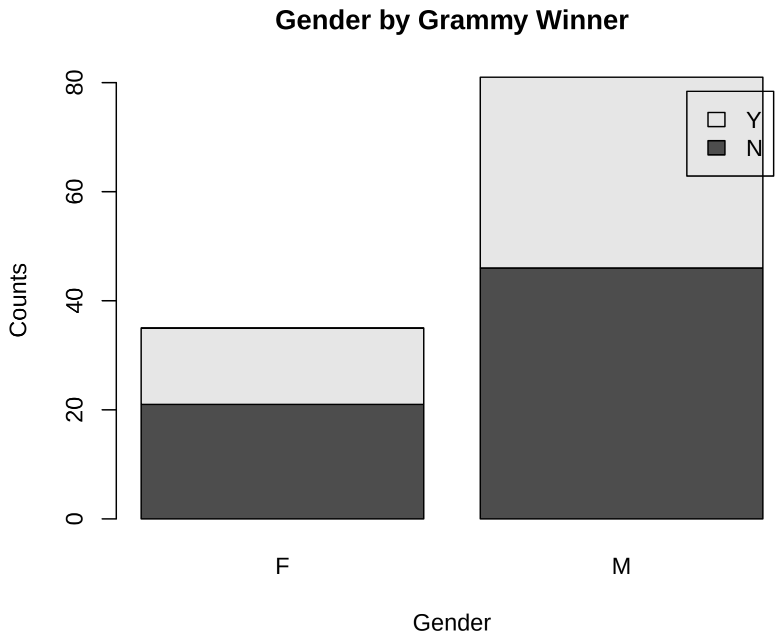
```
barplot(gtab2, legend = TRUE)
```



If we add this, we're going to get a legend that tells us which color corresponds to which value for the variable that we don't see along the x-axis.

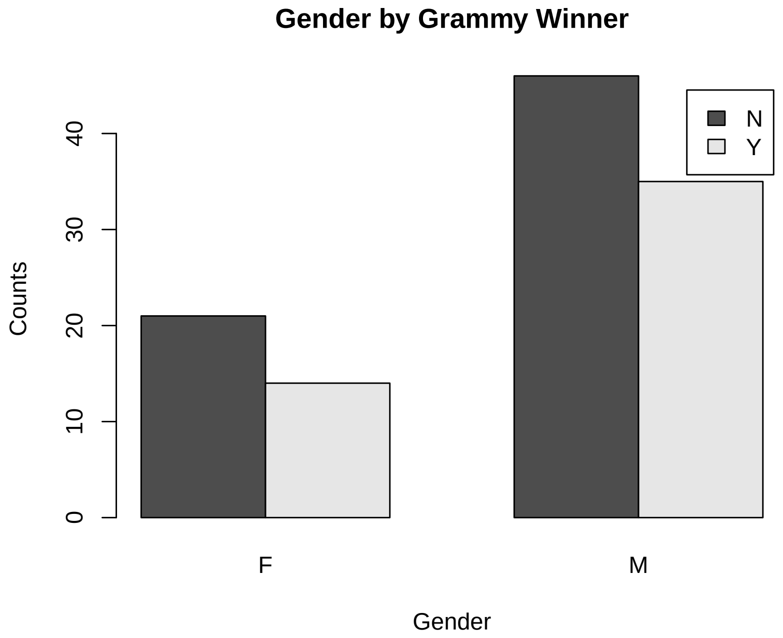
So similar to what we did before, I'm going to want to change the title and axis labels with some options and make that a little bit nicer. We could also include an x label for gender.

```
barplot(gtab2, legend = TRUE, main = "Gender by Grammy Winner",
    xlab = "Gender", ylab = "Counts")
```



Now there's one more option that we can do to switch between a stacked bar chart, which we have here, where literally the counts of the Grammy variable are stacked within each category of gender. I can include a "beside=TRUE" option and that will switch it to a side by side bar chart instead of a stacked bar chart.

```r
barplot(gtab2, legend = TRUE, main = "Gender by Grammy Winner",
    xlab = "Gender", ylab = "Counts", beside = TRUE)
```



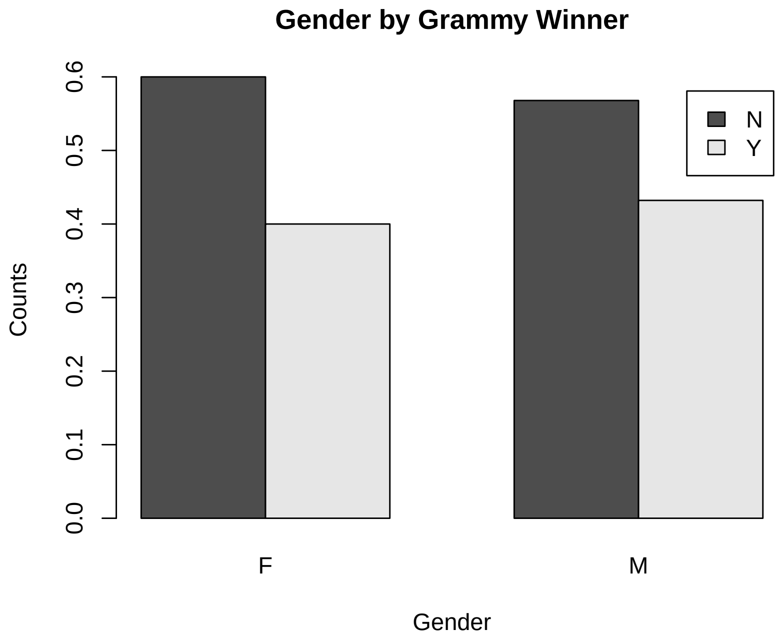
One more trick we can do with the **barplot()** function is to actually give us a relative frequency stacked bar chart, or that's sometimes called the mosaic plot. So remember when we did **prop.table()** of R gtab2 table and asked for the second option, which is the conditional probability of Grammy yes or no within each gender?

```
prop.table(gtab2, 2)
```

```
##
##             F         M
##   N 0.6000000 0.5679012
##   Y 0.4000000 0.4320988
```

If we give bar plot that same statement (‘prop.table(gtab2, 2)’), we’re going to actually get to a relative frequency along the y-axis.

```
barplot(prop.table(gtab2, 2), legend = TRUE, main = "Gender by Grammy Winner",
    xlab = "Gender", ylab = "Counts", beside = TRUE)
```



Instead of the count, we’re going to actually get the proportions. And then we can see within each gender what proportion were Grammy winners and non Grammy winners. And here with our mosaic plot, we can see that although we have more males in the data set, the proportion among males and females that are Grammy winners is pretty consistent at about 60%.