

more.on.graphs

Your name:

September 6,2023

```
#click on this-  
#Note:you may need to import some packages but the program will tell you if  
so.  
#these libraries are needed for filter and more advanced plots  
library(ggplot2)  
library(dplyr)  
  
##  
## Attaching package: 'dplyr'  
  
## The following objects are masked from 'package:stats':  
##  
##   filter, lag  
  
## The following objects are masked from 'package:base':  
##  
##   intersect, setdiff, setequal, union
```

We will be using the datasets myfam2, titanic_tr, and unicefb2016. Download the files for these (titanic_tr, census1880midlancpa, unicefbas2016) from resources on our webpage. In order to knit this rmarkdown with these datasets you will need to have load statements similar to those in the following snippet. To get what your load statements should be, load each of the files for these by first finding your files and then loading them to R by going to files in the lower right hand window, locating your file, and clicking on it. A load statement will appear in the console. Copy that and paste it into the code snippet below. Do this for each of the files. (You can also go to history, select the load statements, and click to source.) Remove my load statements and click the green arrow.

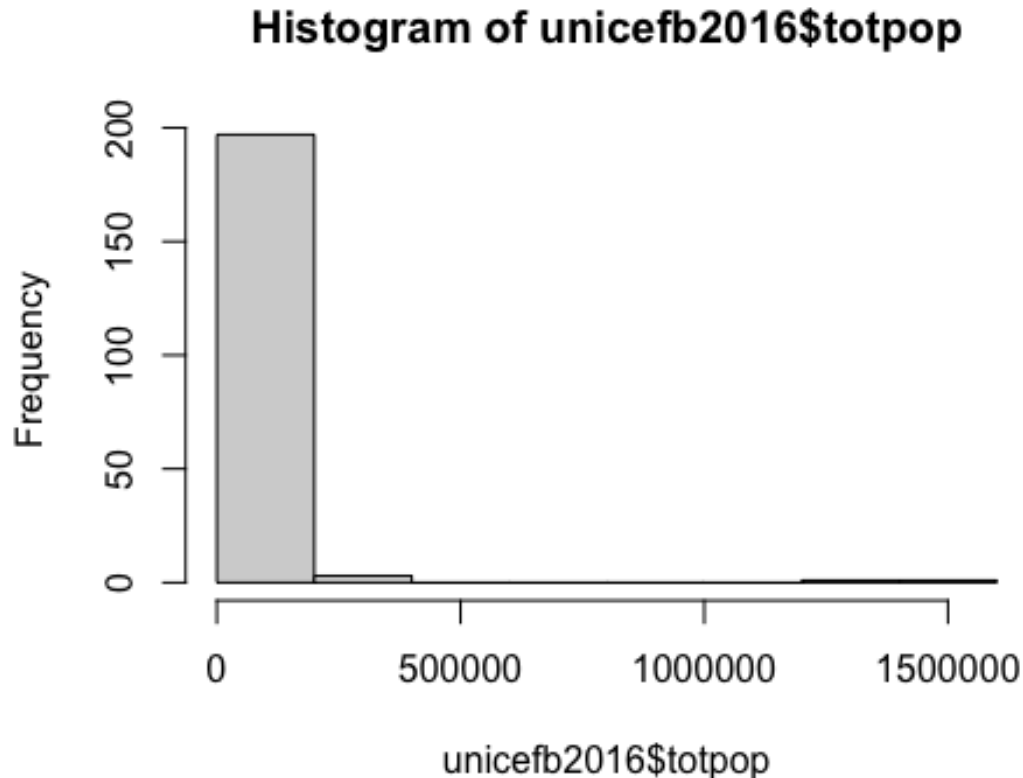
```
load("~/Desktop/Rnotes/census1880midlancpa.rda")  
load("~/Desktop/Rnotes/titanic_tr.rda")  
load("~/Desktop/Rnotes/unicefbas2016.rda")
```

Make a histogram of the variable birthyear in the dataset thefam2. Explain what that histogram tells you.

```
#hist( )
```

Make a histogram of the variable Age in titanic_train. Make a stem plot of the variable Fare in titanic_train. Explain to your neighbor what each of these tell you, including information on their shape.

```
hist(unicefb2016$totpop)#very skewed right-why?
```



The unicef dataset variables codebook can be found in the course resource webpage. Take a minute to look at it. Choose two numerical variables. Then create histograms for those. Write a statement explaining what each histogram tells you.

For totpop we cannot see what is happening very well so we want to restrict to those with $\text{totpop} \leq 400000$. (This is in thousands.)

First let's see how many countries are bigger than 400000 thousand.

#The line below gives us a Boolean vector of trues and falses.

```
unicefb2016$totpop > 400000
```

```
## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [37] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```

FALSE
## [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE
FALSE
## [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

```

#These are also encoded as zeros and ones for false and true. Thus the following sum gives us the number of ones or # the number of

```

sum(unicefb2016$totpop>400000)

```

```

## [1] 2

```

#Note that this is a very useful and common construction.

If we want to see what countries those are we would enter this vector in unicefb2016\$countries.and.areas. The output is those entries with index true.

```

unicefb2016$countries.and.areas[unicefb2016$totpop>400000]

```

```

## [1] "China" "India"

```

Change the above to see which countries have totpop greater than 300000 thousand.

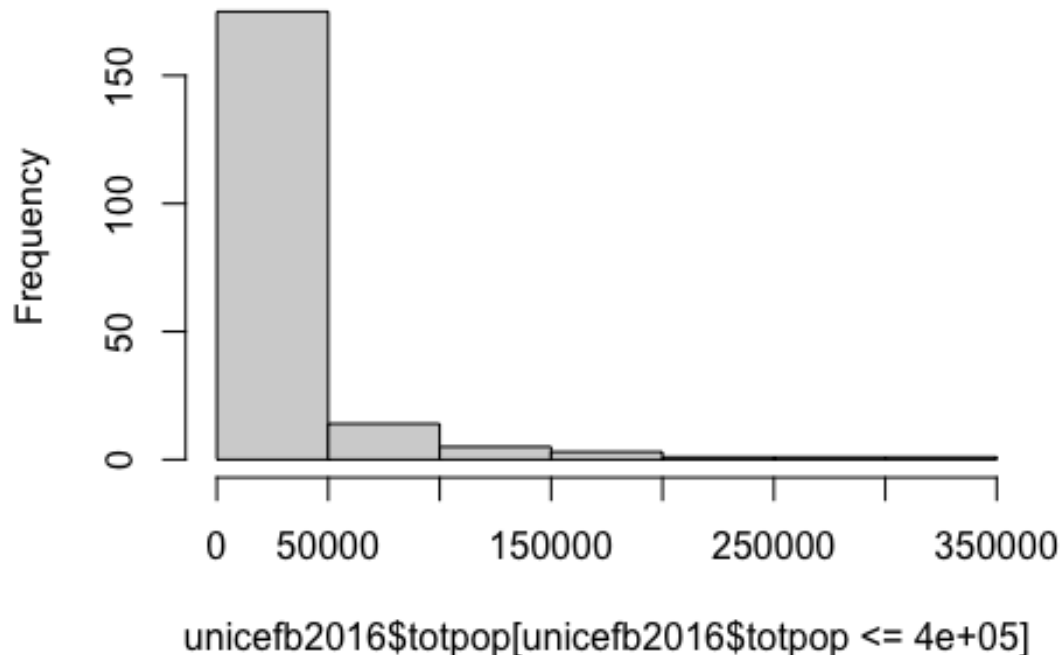
Now let's create a histogram of the total populations less than or equal to 400000 thousand.

```

hist(unicefb2016$totpop[unicefb2016$totpop<=400000])

```

Histogram of unicefb2016\$totpop[unicefb2016\$totpop <=



#Let's try for 100000; then 10000

Now make a histogram of the literacy rates in countries with less than 5 million total population.

Make a histogram of all literacy rates.

Make a histogram of literacy rates for countries in SubSaharan Africa (unicefb2016\$classf=='ssa'). Note the double equal signs. When one is checking a condition of equality, the double equal sign is used.

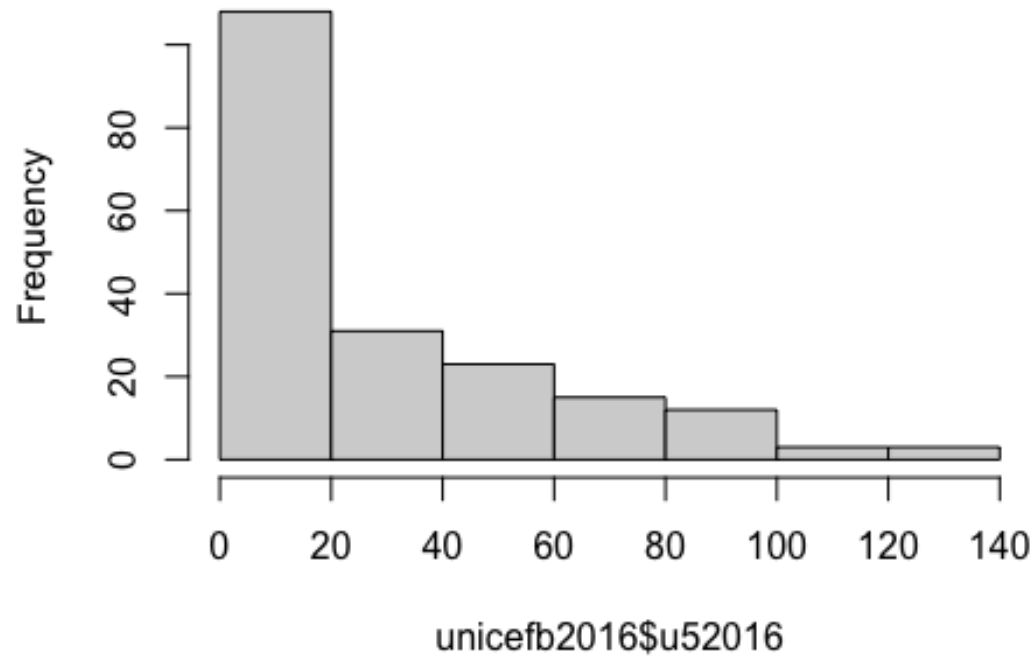
Make a barplot of the regions in unicef (classf)

Make a stemplot of the ages of the first 30 people in the titanic dataset.

Some homework type problems 1.36 Make a histogram of the mortality rates for children under 5 in countries. The book dataset is for 2015 and gives rates per 1000 children. We will use our unicef dataset, which is for 2016. a. Why do you think the mortality rates are measured in deaths per 1000 children under age 5, rather actual number of deaths? b. Make a histogram. Describe the shape, center, variability of the distribution. Do any countries appear to be obvious outliers?

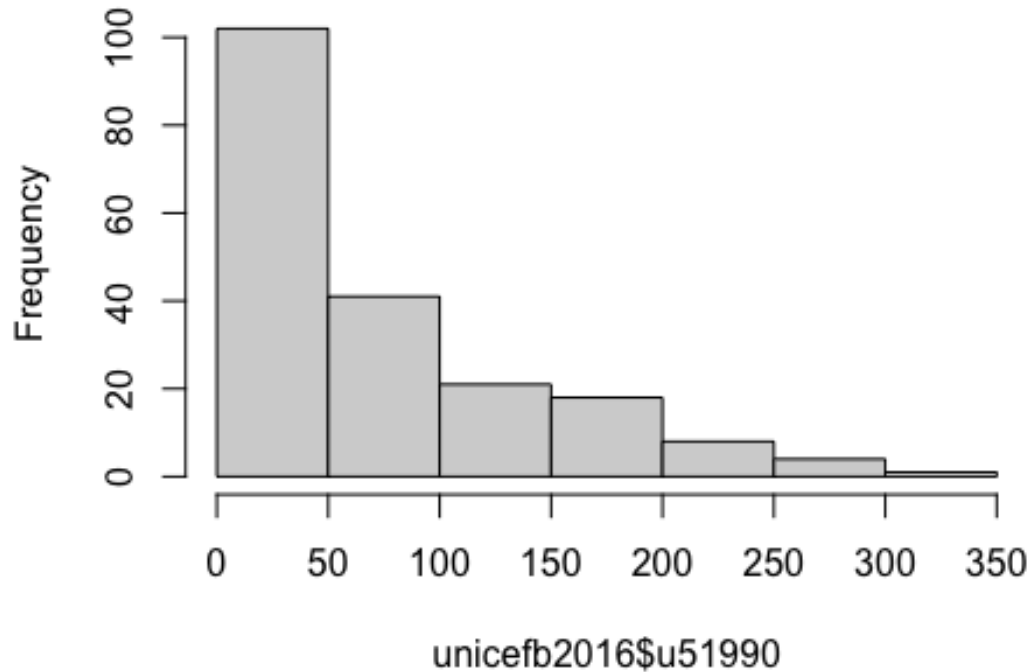
```
hist(unicefb2016$u52016)
```

Histogram of unicefb2016\$u52016



```
hist(unicefb2016$u51990)#What is the difference?
```

Histogram of unicefb2016\$u51990

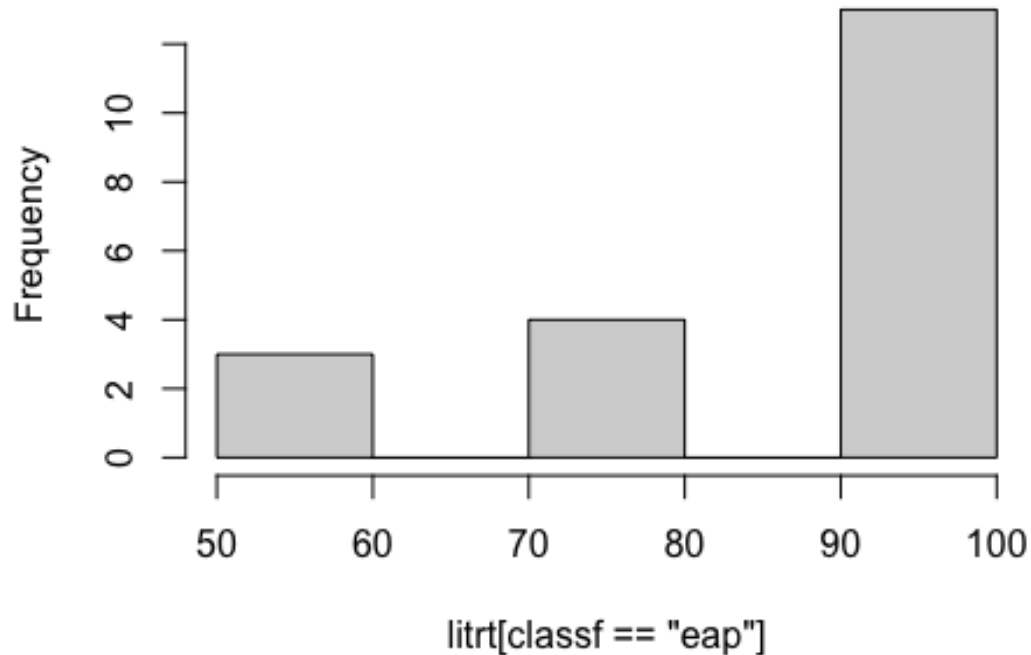


1.19 Look at in book. Interpreting a stemplot.

Some shortcuts: 1) Instead of typing the dataset multiple times in a statement one can use `with(dataset,code)`. Example below. Note there are some instances where this construct will not work.

```
with(unicefb2016, hist(litrt[classf=='eap']))
```

Histogram of `litrt[classf == "eap"]`



this is shorthand for `hist(unicefb2016$litrt[unicefb2016$classf=='eap'])`

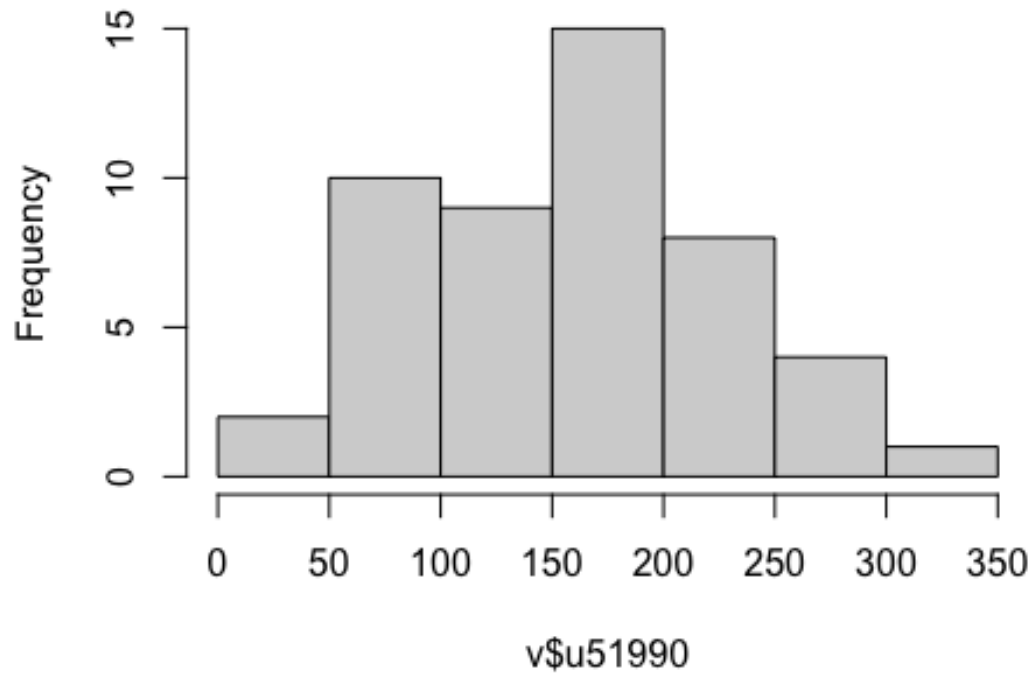
2. This one needs `dplyr` activated before use. (We did this earlier with the statement `library(dplyr)` in a previous code snippet.) Here one creates a smaller dataset by filtering and then uses the variable of that smaller dataset.

```
v=unicefb2016%>%filter(classf=='ssa')
```

#v consists of those lines/observations in the unicef dataset with classf = ssa.

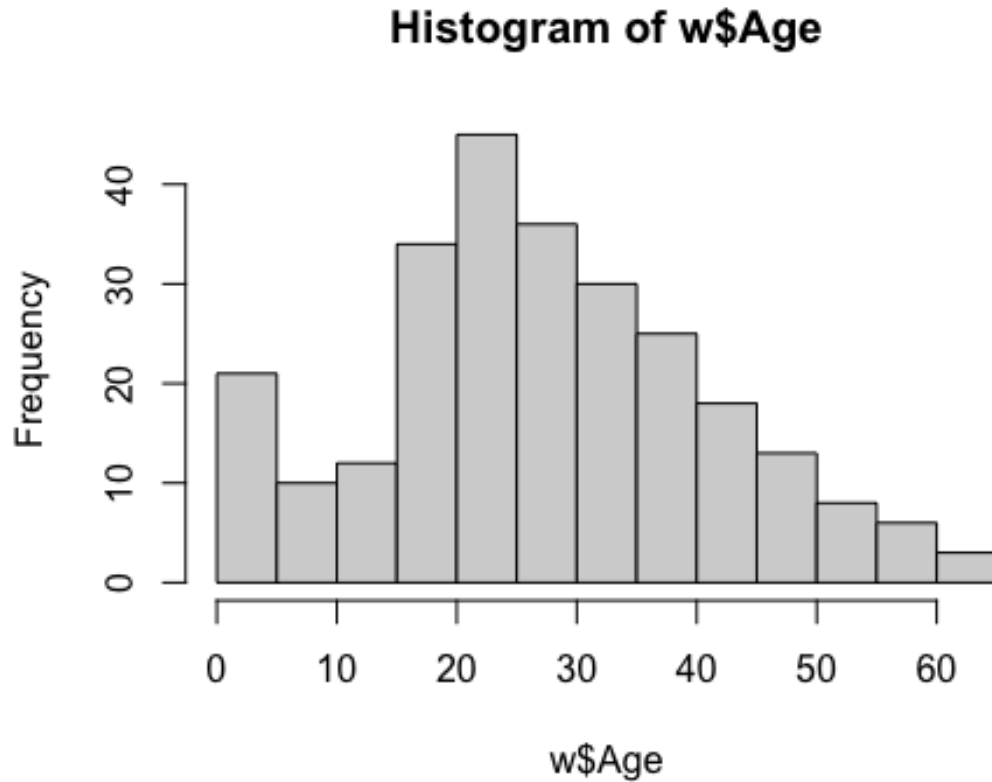
```
hist(v$u51990)
```

Histogram of v\$u51990



Let's look at the histogram of ages of women in titanic

```
w=titanic_train%>%filter(Sex=='female')  
hist(w$Age)
```

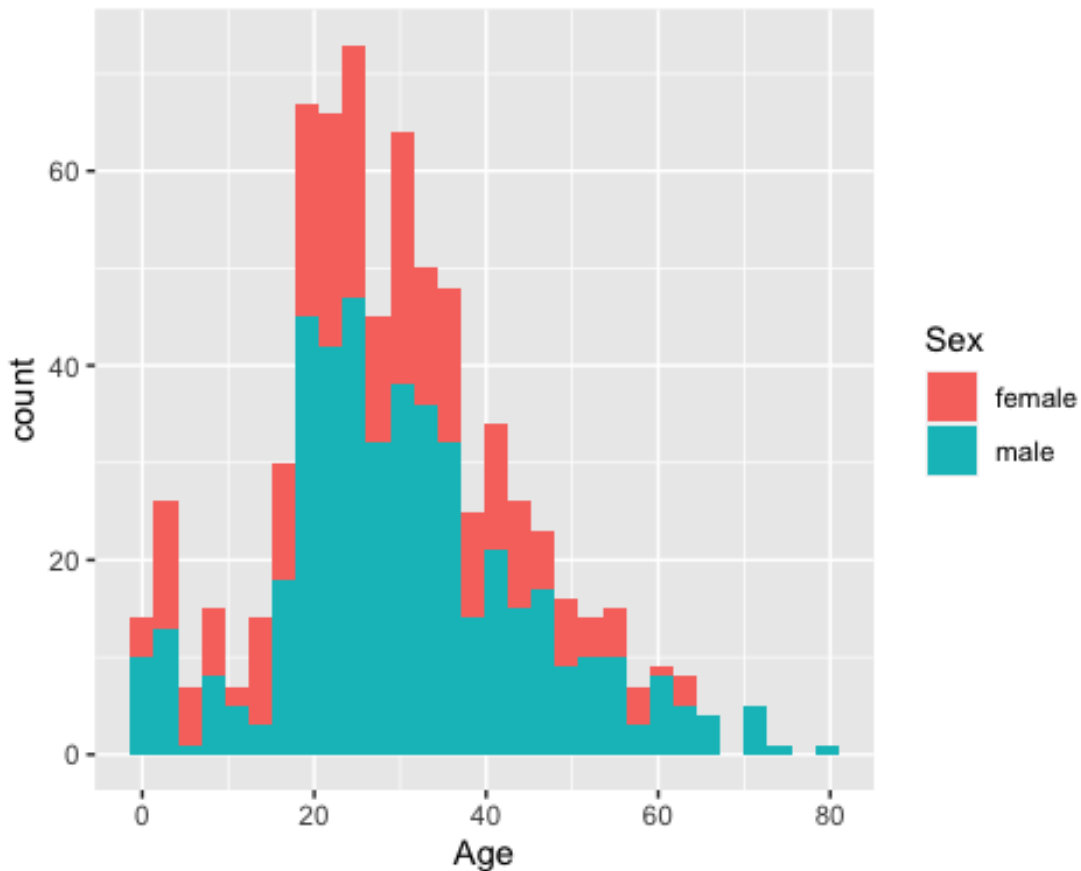
#redo this for males

A fancy stacked histogram

```
ggplot(titanic_train,aes(x=Age,fill=Sex))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 177 rows containing non-finite values (`stat_bin()`).
```



*#This uses the more advanced graphing package ggplot which you will not be required to learn. I may ask you occasionally to modify code snippets that have ggplot in them but there will be minimal modification. One puts in the dataframe/dataset (here titanic_train); then inside aes: x=variable (here Age) ,fill(for fill the bars of the histogram) = the variable you are grouping by(here Sex)
#Change Sex to Pclass in this code.*

In the snippet below: 1. Create a dataset called uneapwe which consists of the countries from eap (eastern Asia Pacific) and we (Western Europe) using filter(classf=='eap'| classf=='we'). The | stands for or. Modify the code in the above snippet to create a stacked histogram of literacy rates in eap and we. What do you notice from the graph?

Summary: loading a file in an rmarkdown; restricting the dataset based on values of a variable; histograms filter