rachellab06

Rachel Brunner

11:59PM April 10, 2022

#Visualization with the package ggplot2

I highly recommend using the ggplot cheat sheet as a reference resource. You will see questions that say "Create the best-looking plot". Among other things you may choose to do, remember to label the axes using real English, provide a title and subtitle. You may want to pick a theme and color scheme that you like and keep that constant throughout this lab. The default is fine if you are running short of time.

Load up the `GSSvocab` dataset in package `carData` as `X` and drop all observations with missing measurements. This will be a very hard visualization exercise since there is not a good model for vocab.

```
pacman::p_load(ggplot2)
pacman::p_load(carData)
X = carData::GSSvocab
X = na.omit(X)
```

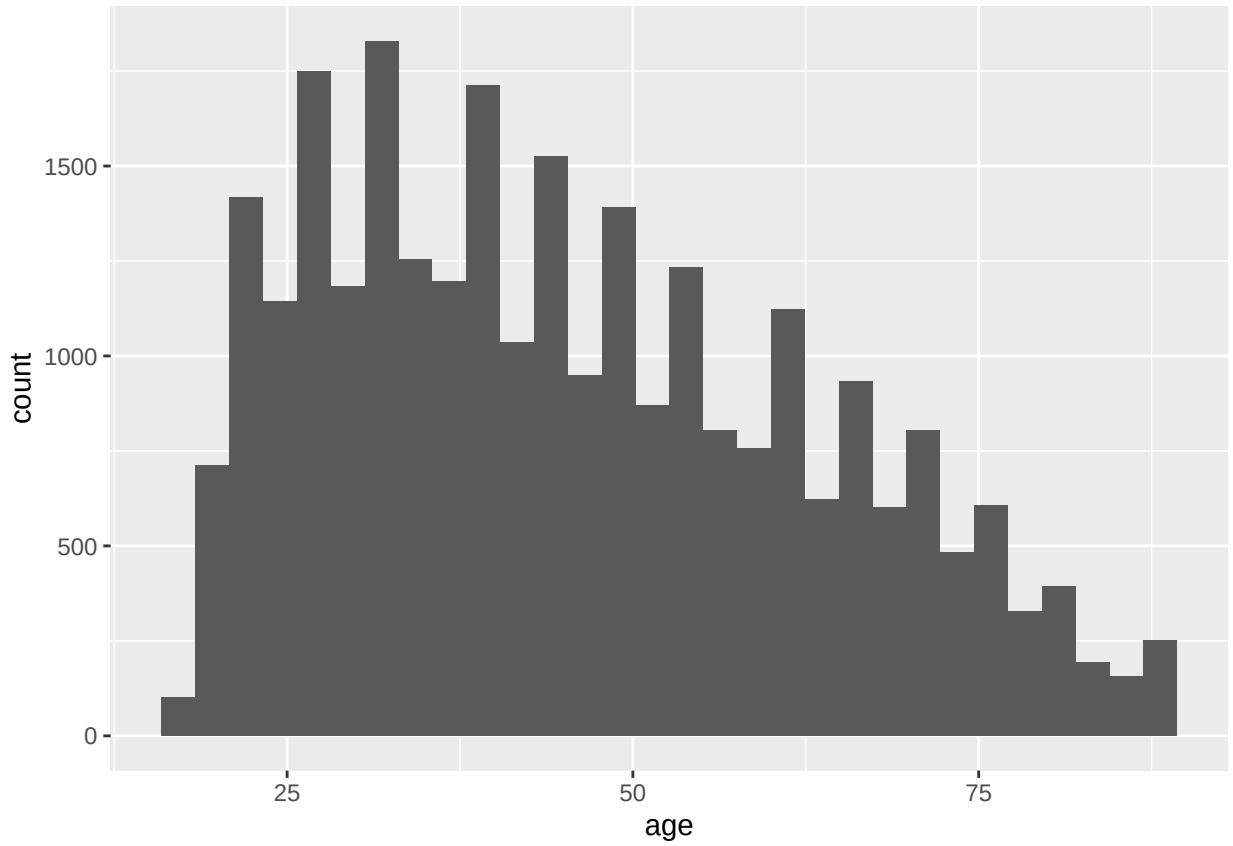
Briefly summarize the documentation on this dataset. What is the data type of each variable? What do you think is the response variable the collectors of this data had in mind? ?carData::GSSvocab do this in console and read descriptions of each variable #TO-DO age and education are both continuous and derived categorical features

Create two different plots and identify the best-looking plot you can to examine the `age` variable. Save the best looking plot as an appropriately-named PDF.

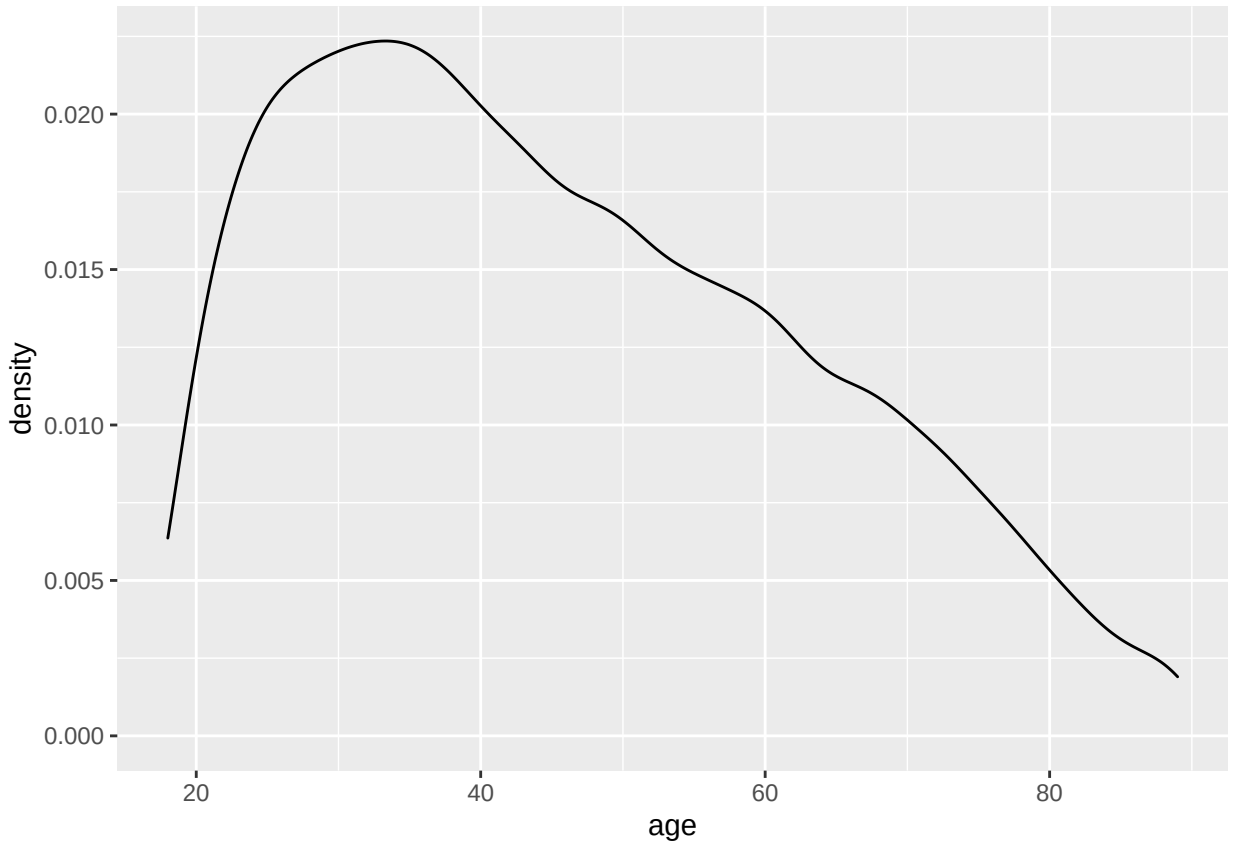
```
ggplot(X,aes(age))
```


```
base = ggplot(X, aes(age))
base + geom_histogram()
```

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

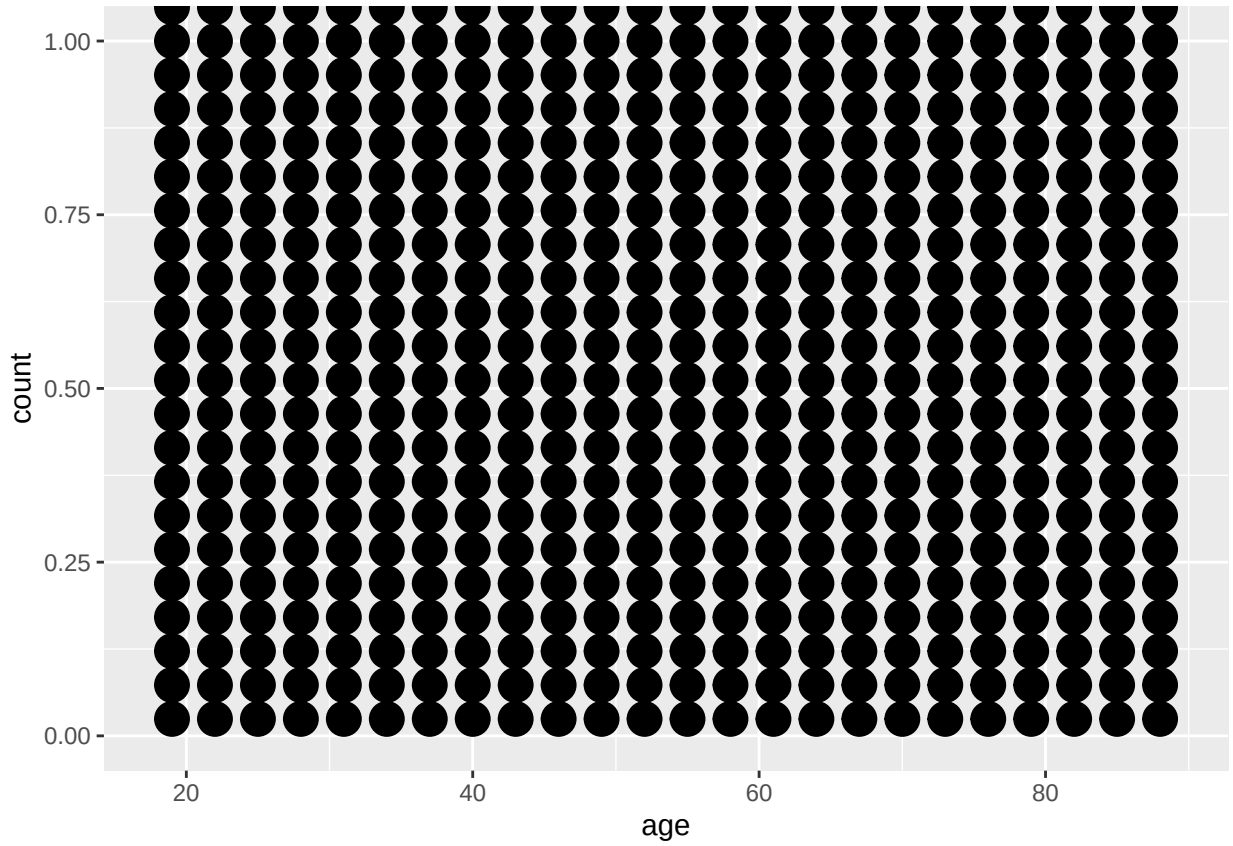



```
base + geom_density()
```



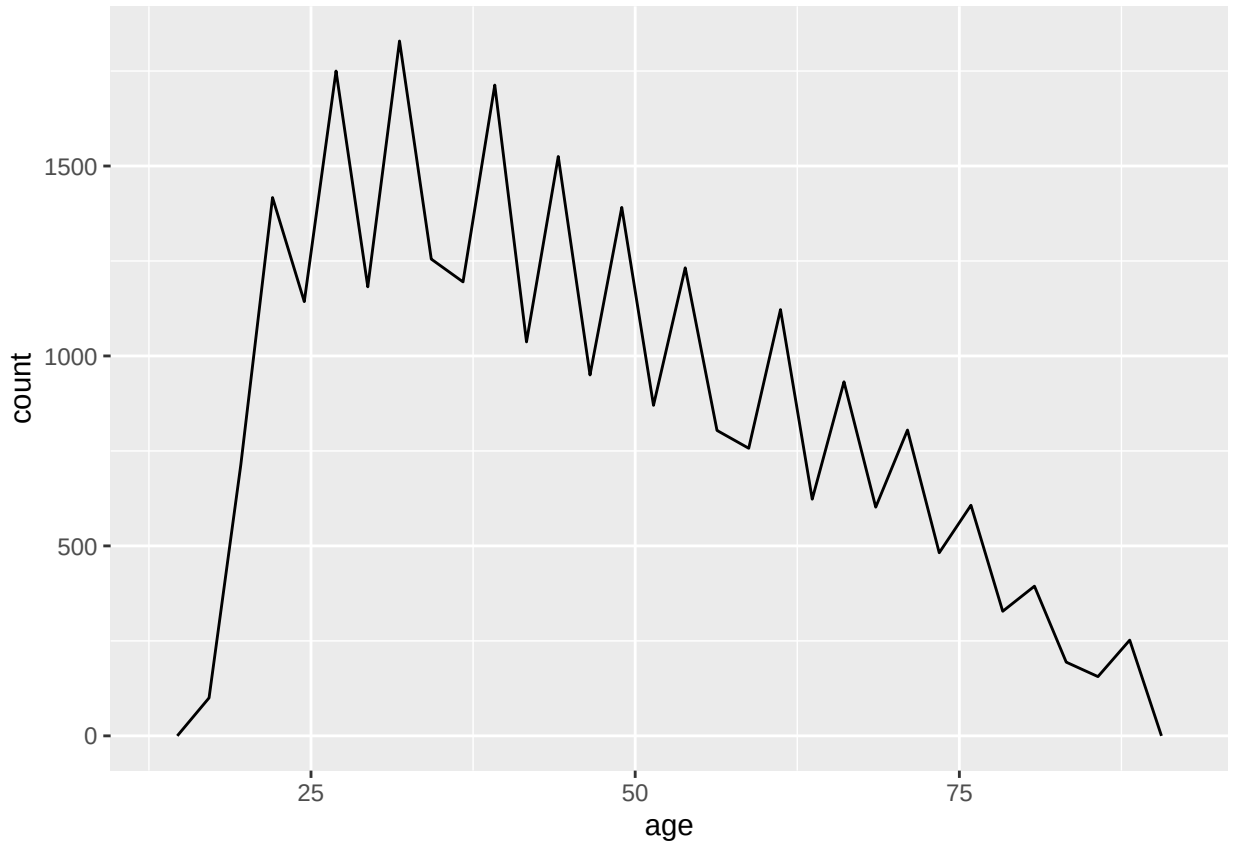

```
base + geom_dotplot()
```

```
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
```



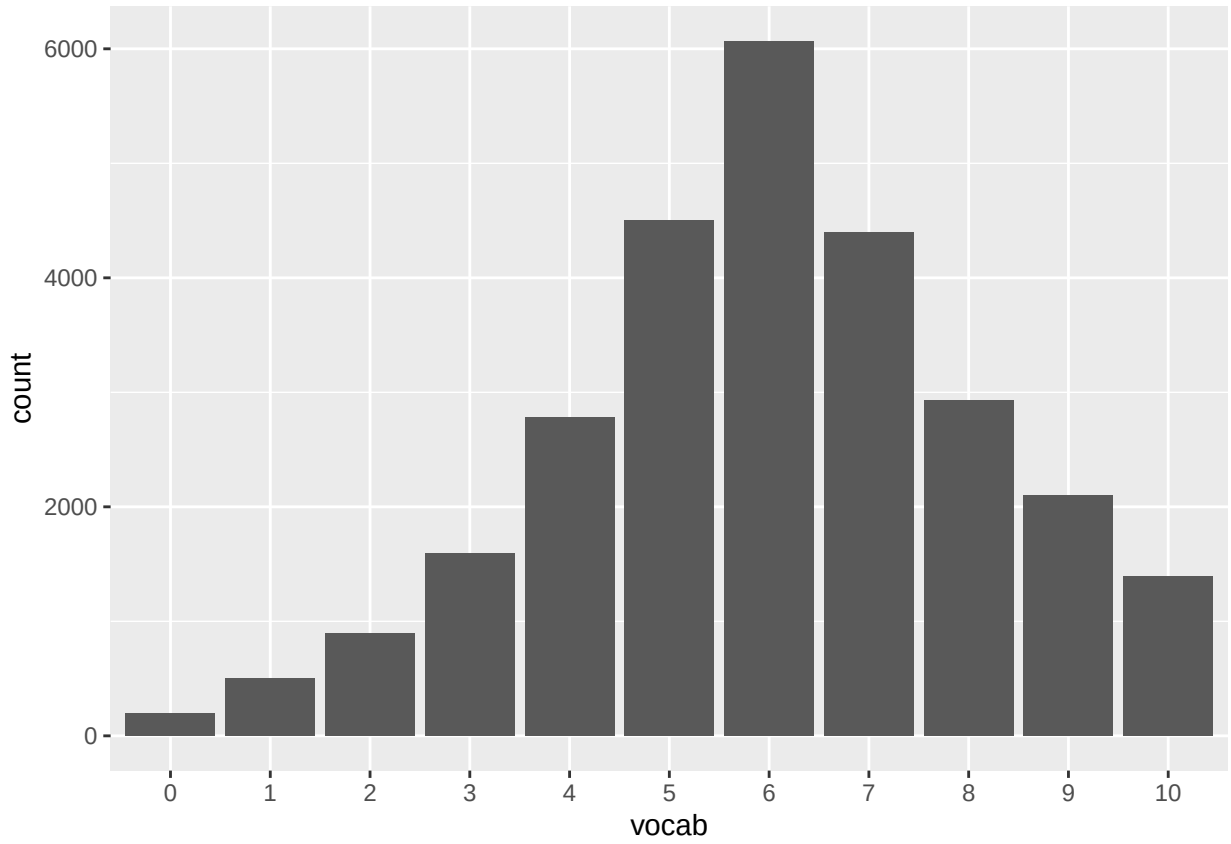

```
base + geom_freqpoly()
```

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



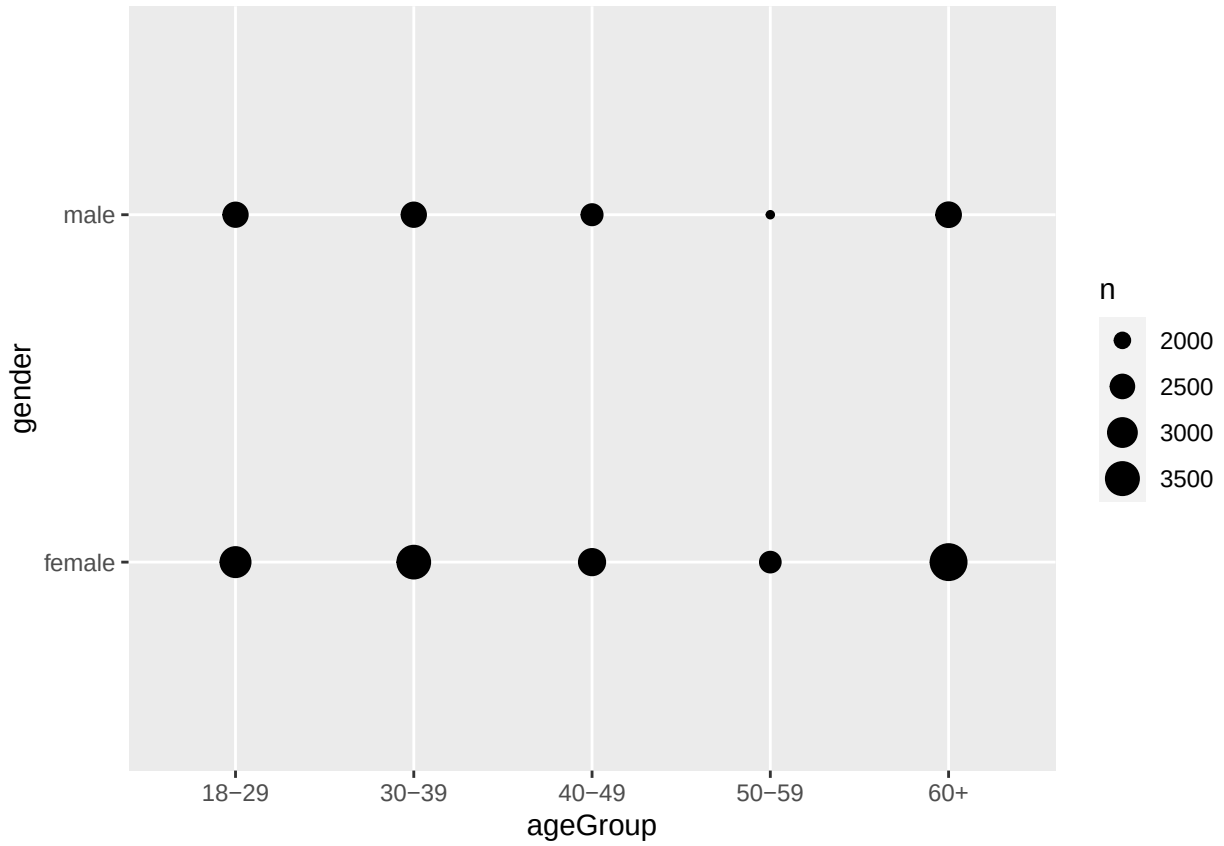
Create two different plots and identify the best looking plot you can to examine the `vocab` variable. Save the best looking plot as an appropriately-named PDF.

```
X$vocab = factor(X$vocab)
base = ggplot(X, aes(vocab))
base + geom_bar()
```

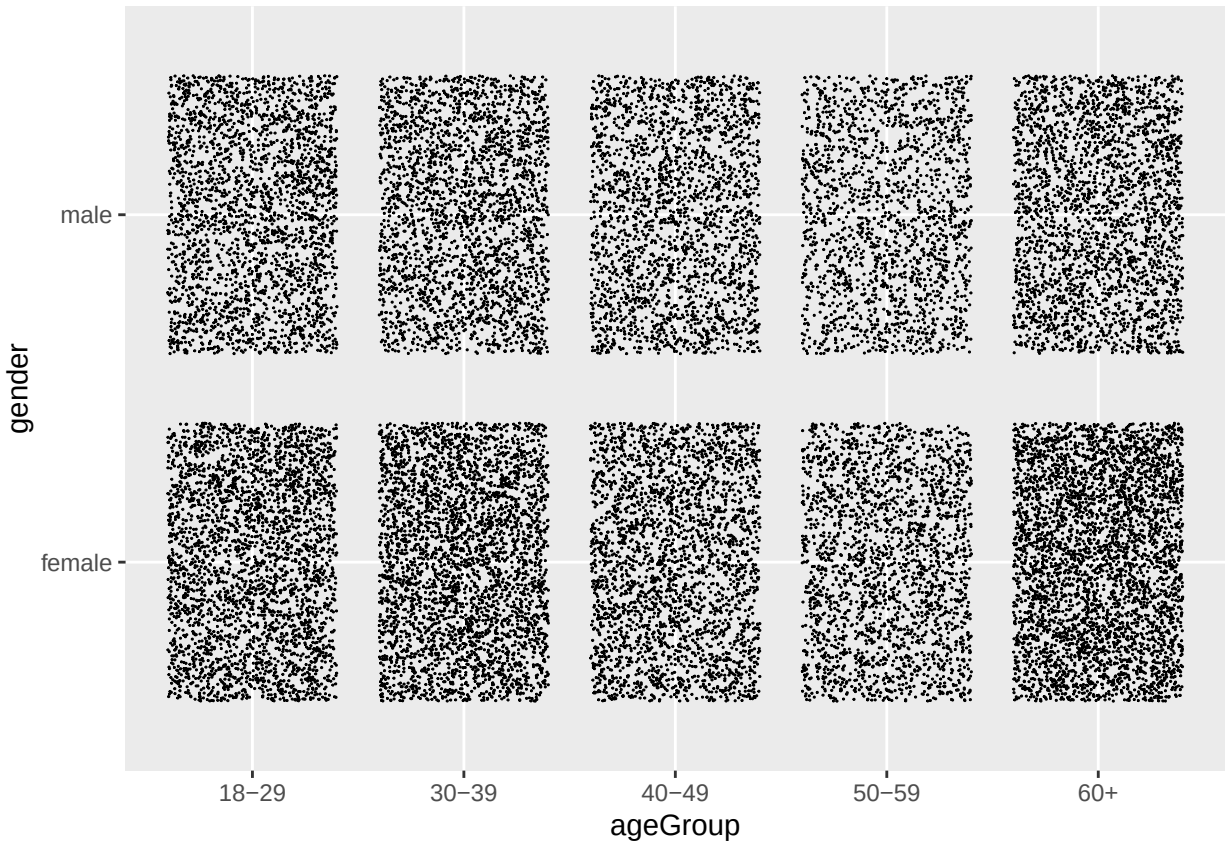


Create the best-looking plot you can to examine the `ageGroup` variable by `gender`. Does there appear to be an association? There are many ways to do this.

```
ggplot(X) +
  geom_count(aes(x = ageGroup, y = gender))
```

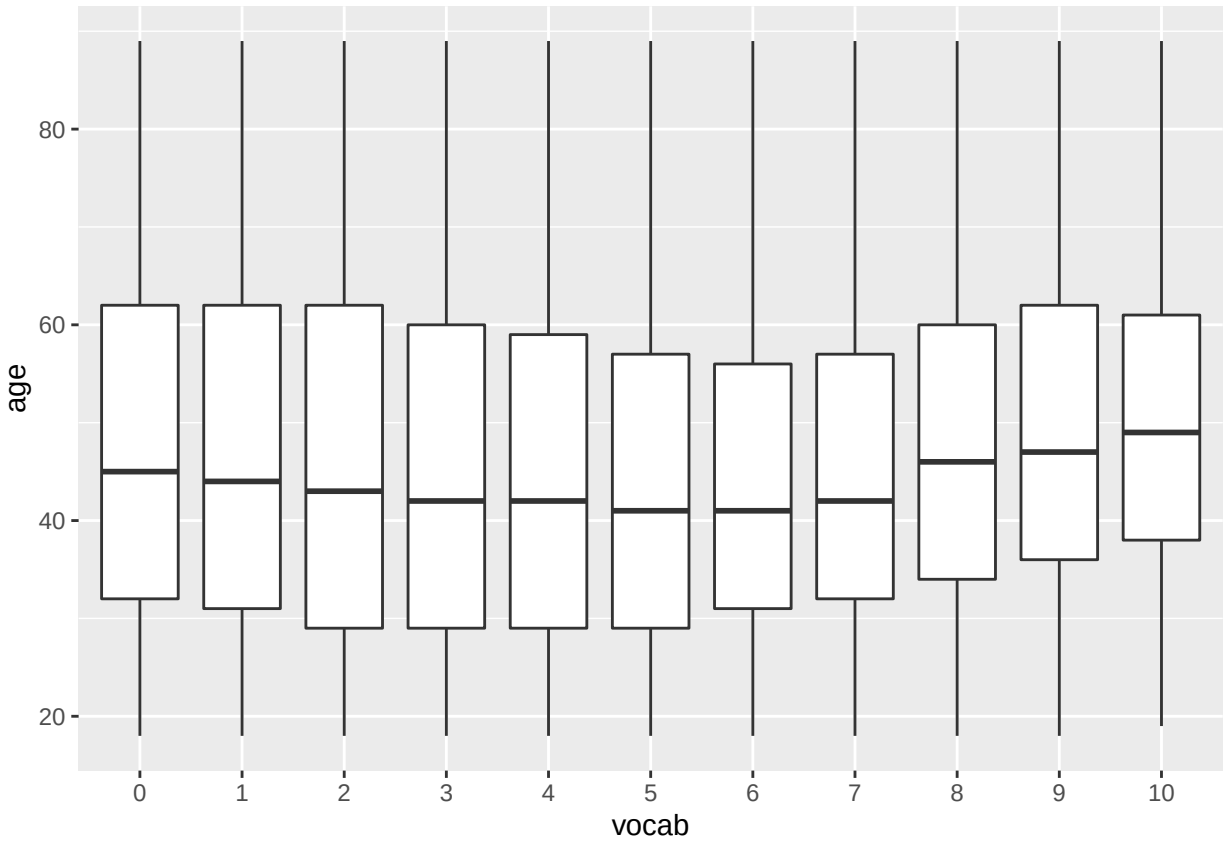



```
ggplot(X) +
  geom_jitter(aes(x = ageGroup, y = gender), size = .000001, shape = 20)
```

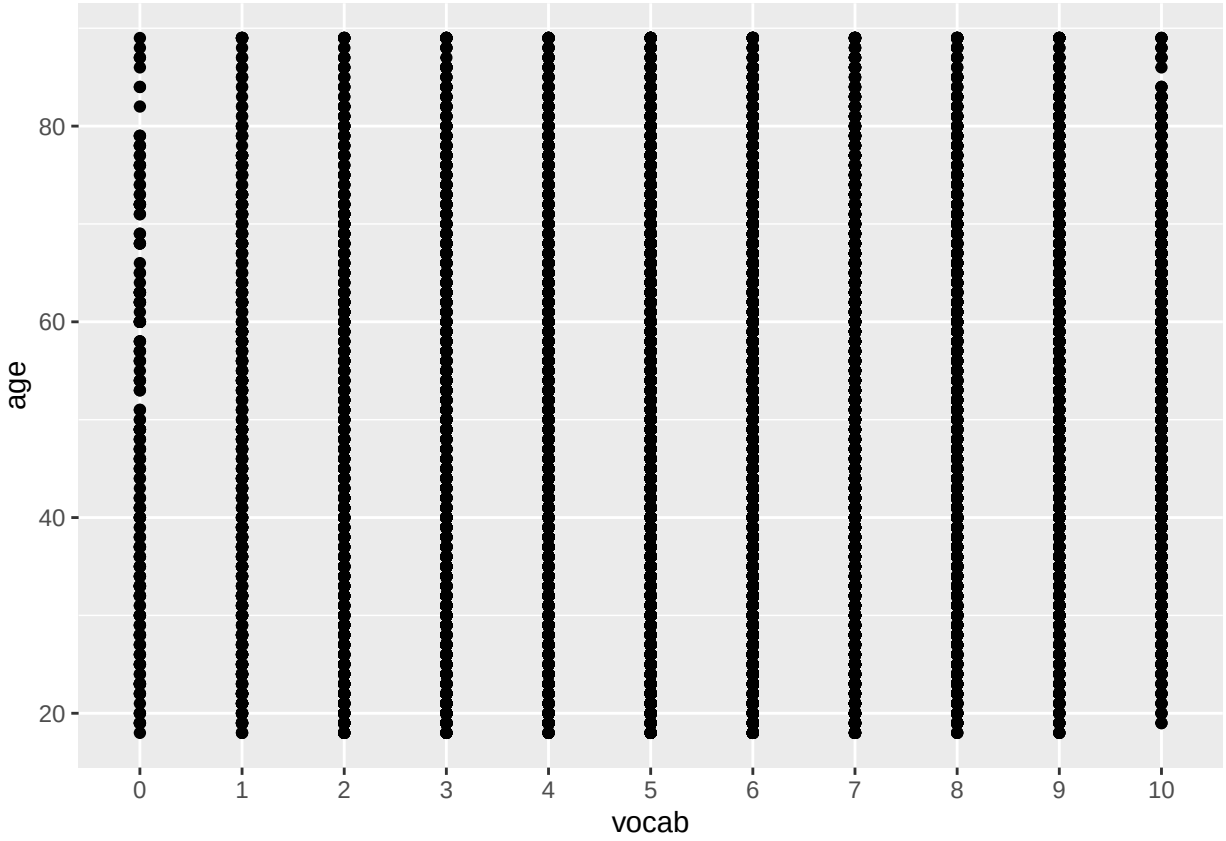


Create the best-looking plot you can to examine the `vocab` variable by `age`. Does there appear to be an association?

```
ggplot(X) +
  geom_boxplot(aes(x = vocab, y = age))
```



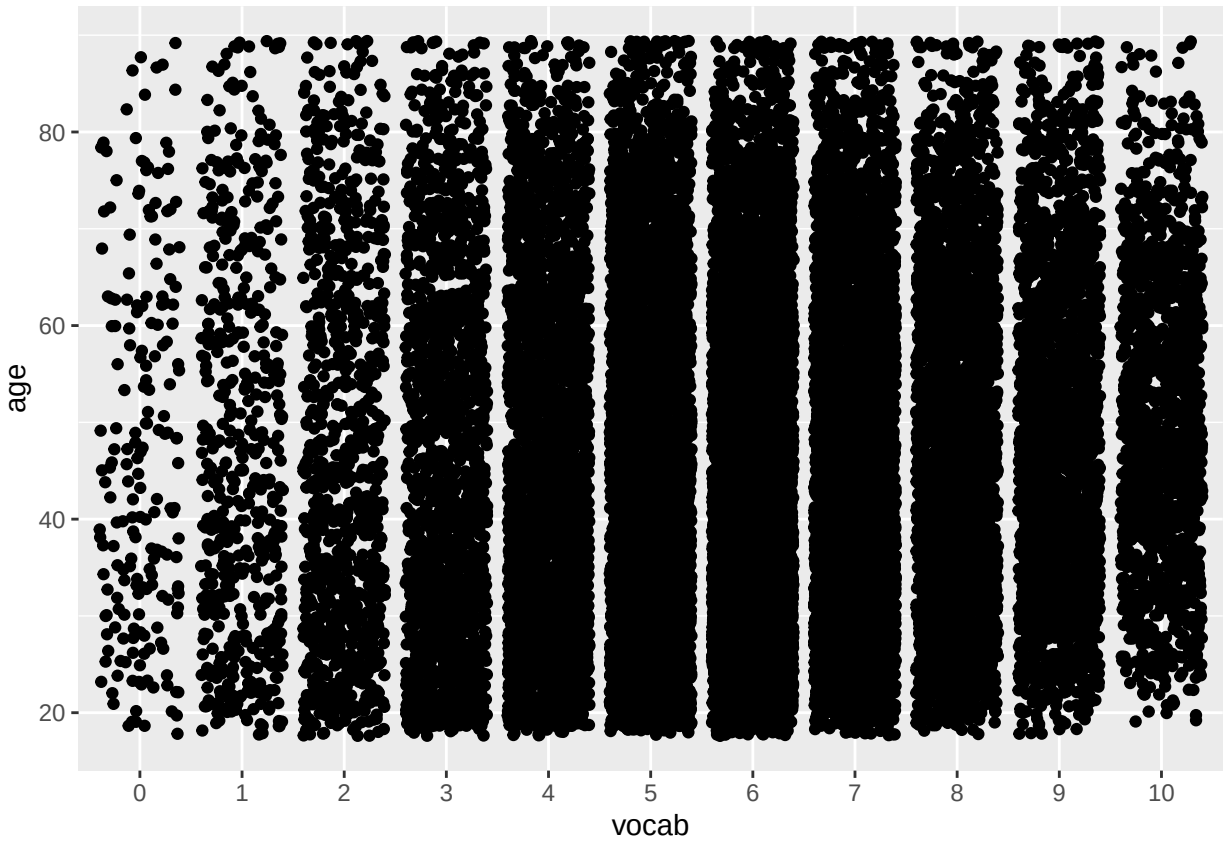

```
#this looks like a hinge, it's kind of flat, but as you age you get more perfect scores.
ggplot(X) +
   geom_point(aes(x = vocab, y = age))
```



Add an estimate of $f(x)$ using the smoothing geometry to the previous plot. Does there appear to be an association now?

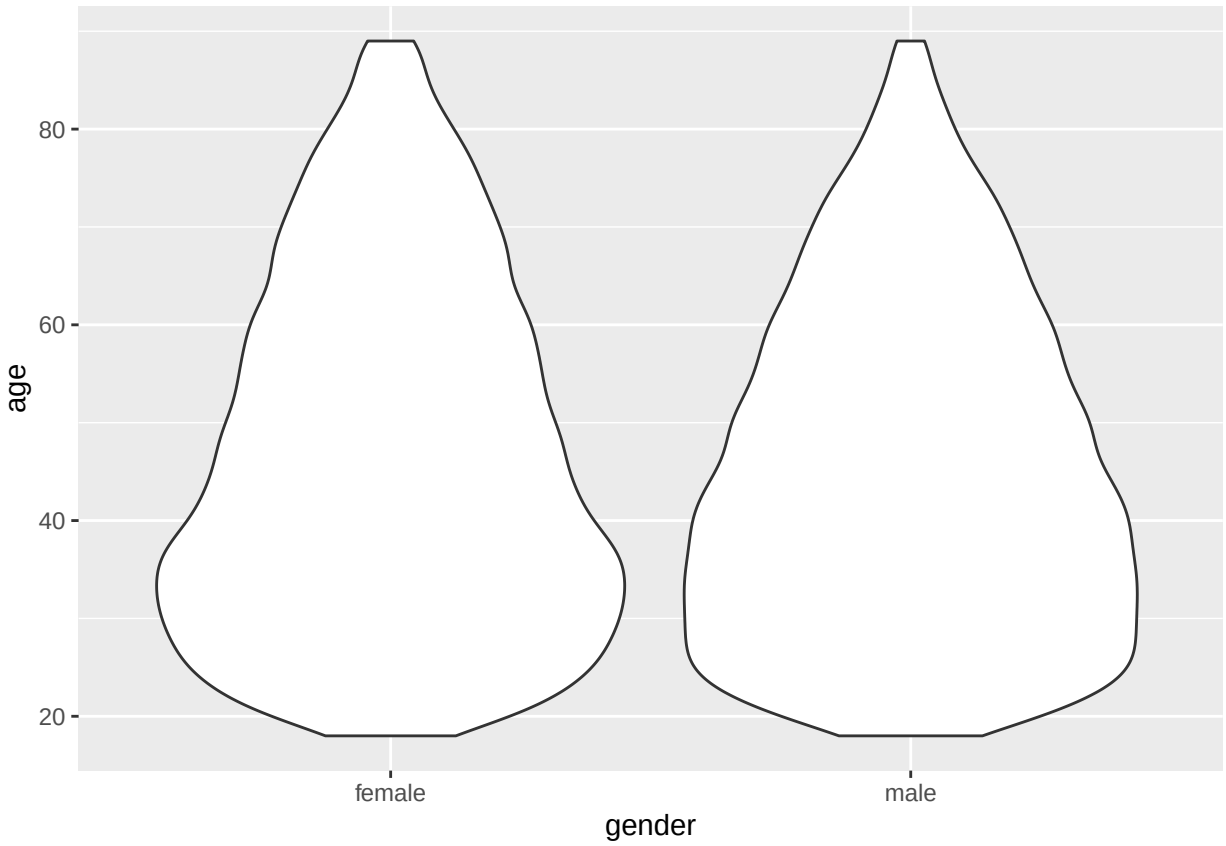
```
ggplot(X) +
  geom_jitter(aes(x = vocab, y = age)) +
  geom_smooth(aes(x = vocab, y = age))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

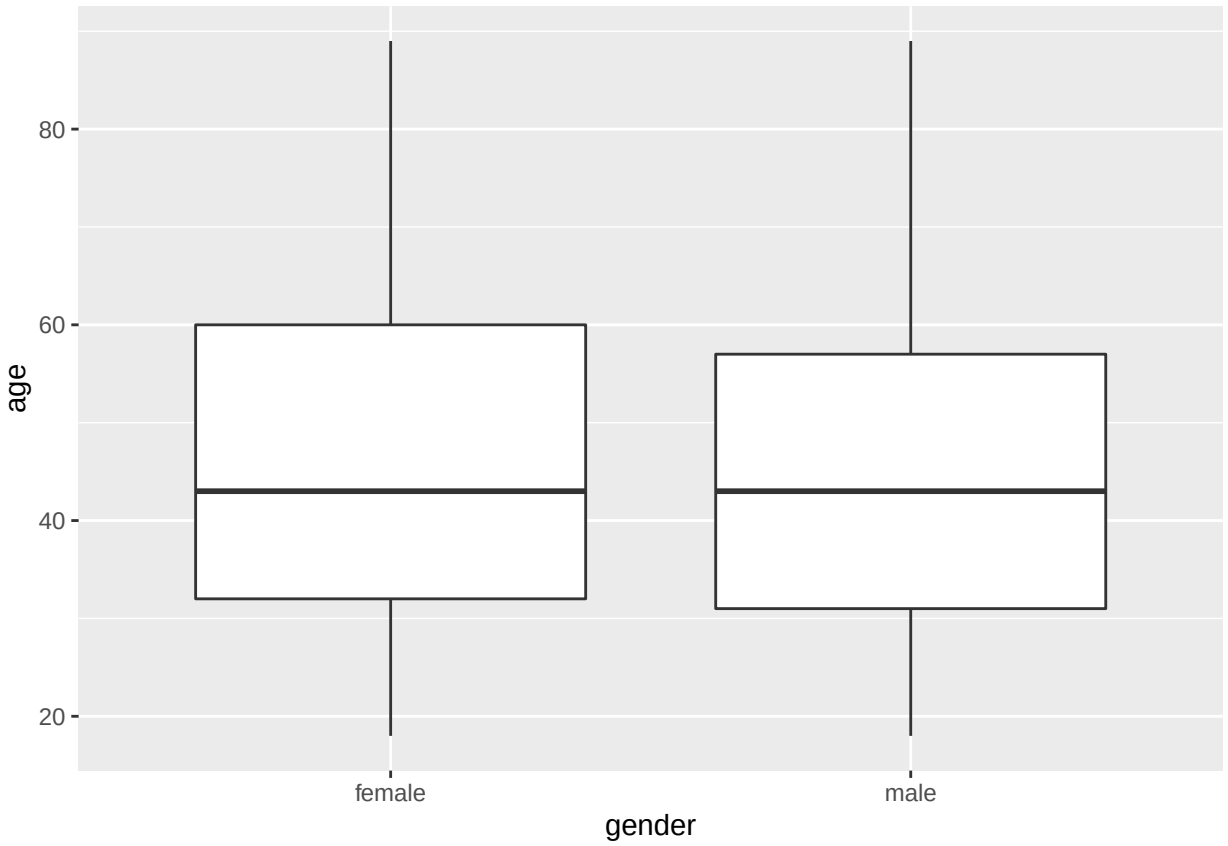


Using the plot from the previous question, create the best looking plot overloading with variable `gender`. Does there appear to be an interaction of `gender` and `age`?

```
ggplot(X) +
  geom_violin(aes(x = gender, y = age))
```

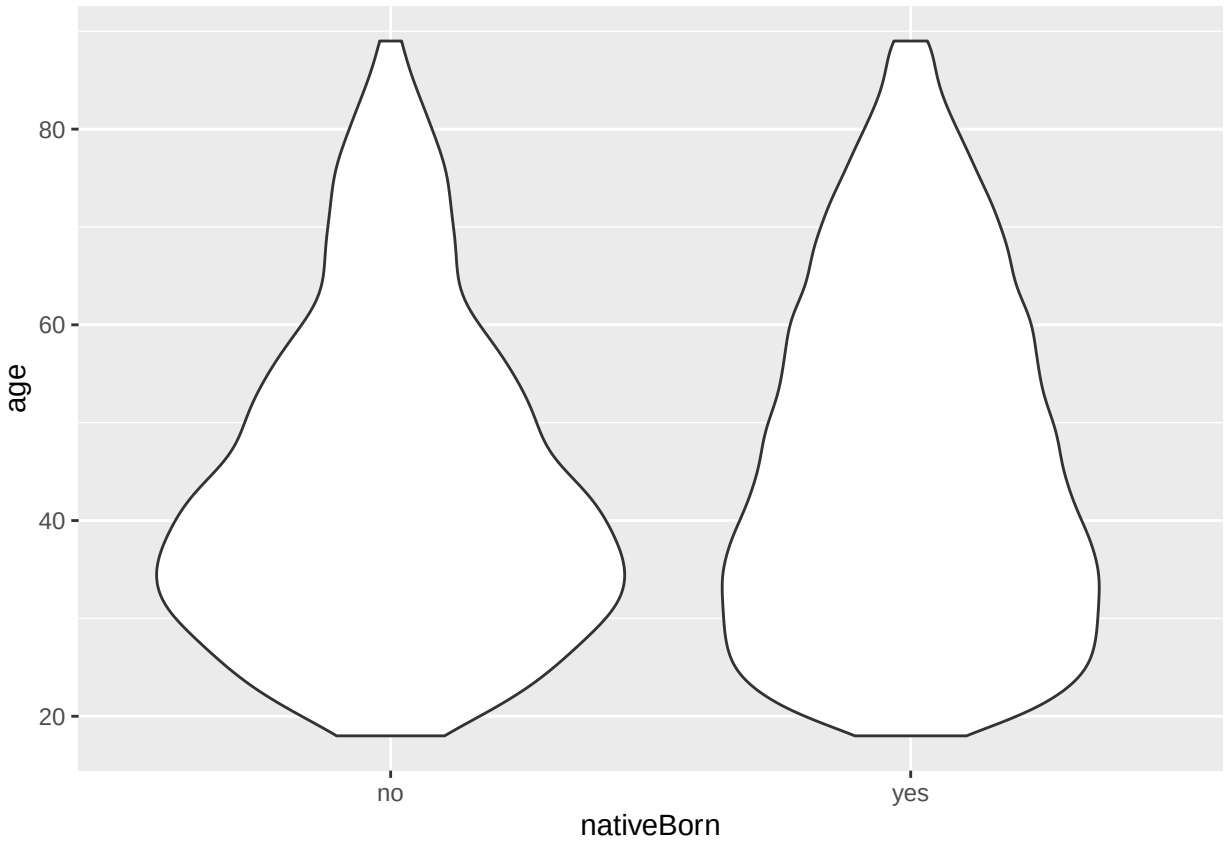



```
ggplot(X) +
  geom_boxplot(aes(x = gender, y = age))
```

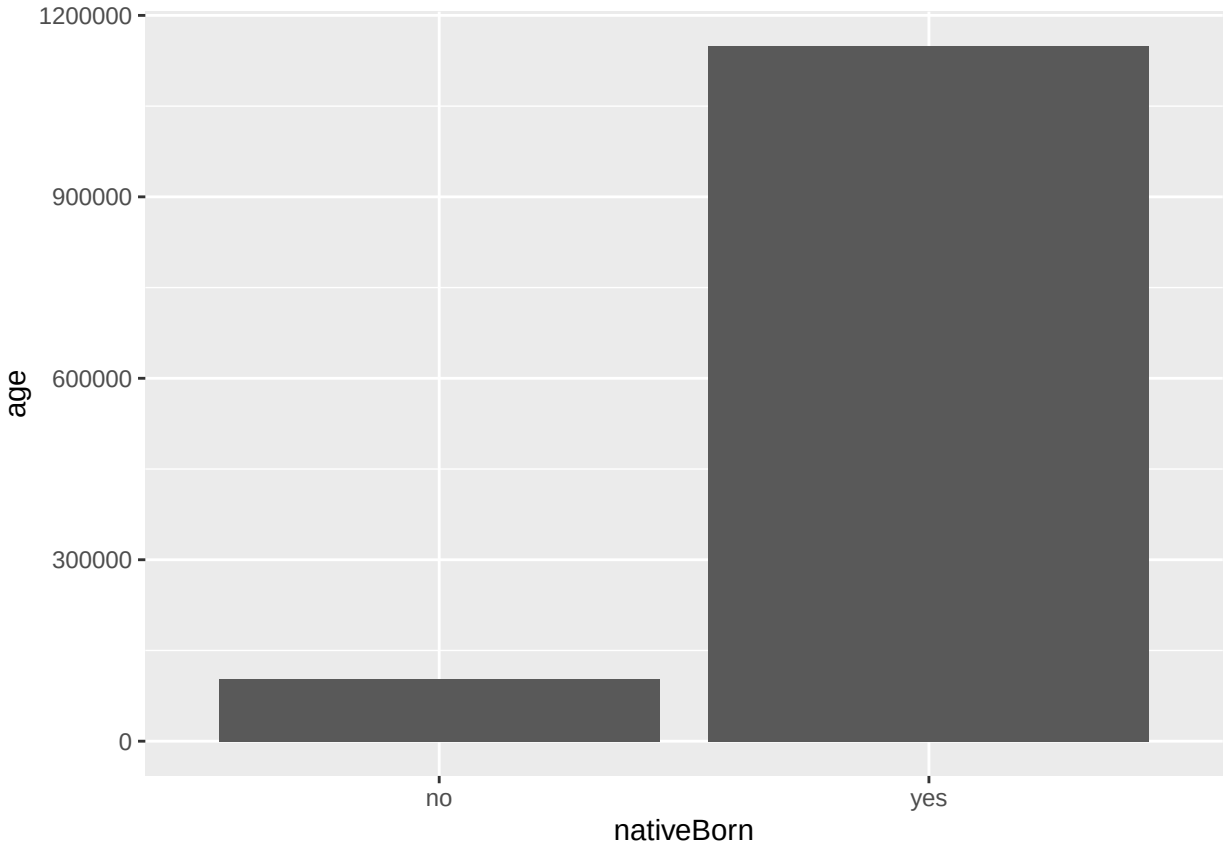


Using the plot from the previous question, create the best looking plot overloading with variable `nativeBorn`. Does there appear to be an interaction of `nativeBorn` and `age`?

```
ggplot(X) +
  geom_violin(aes(x = nativeBorn, y = age))
```

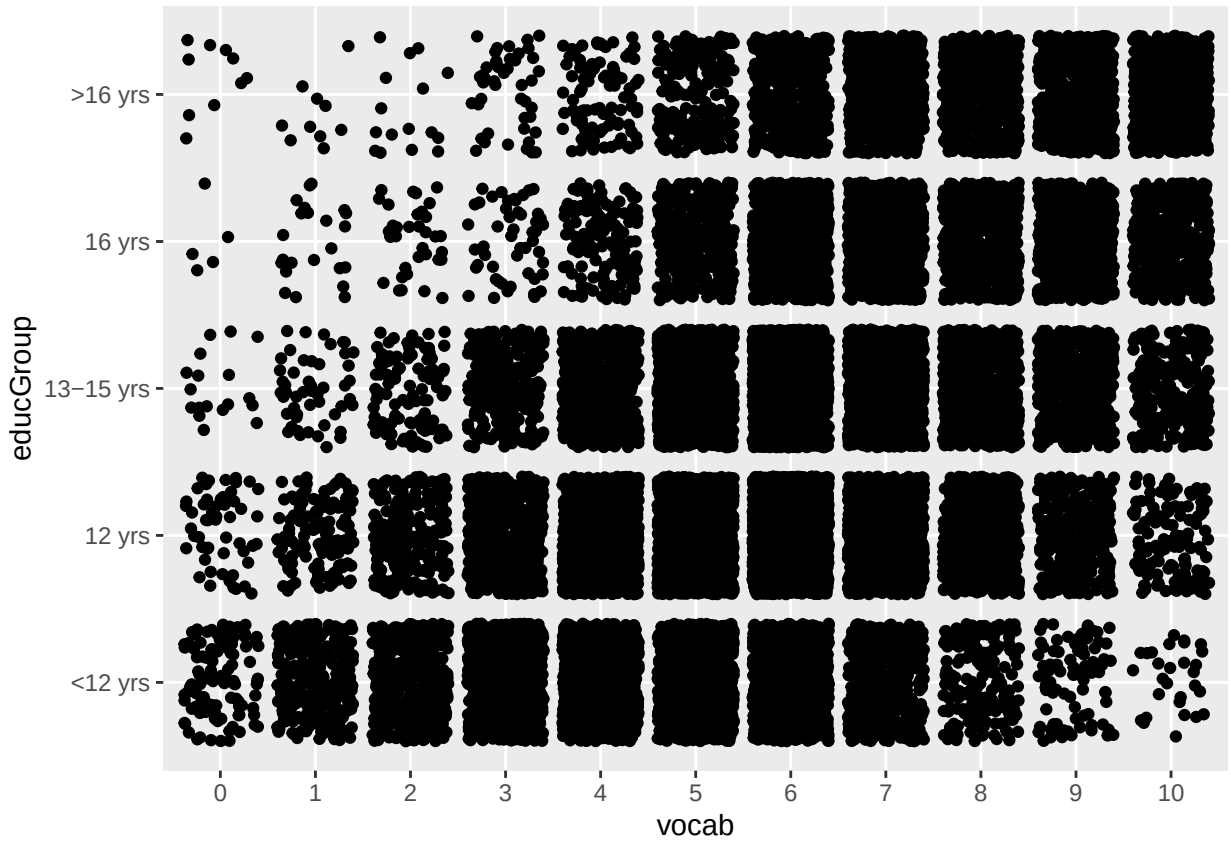



```
ggplot(X) +
  geom_col(aes(x = nativeBorn, y = age))
```

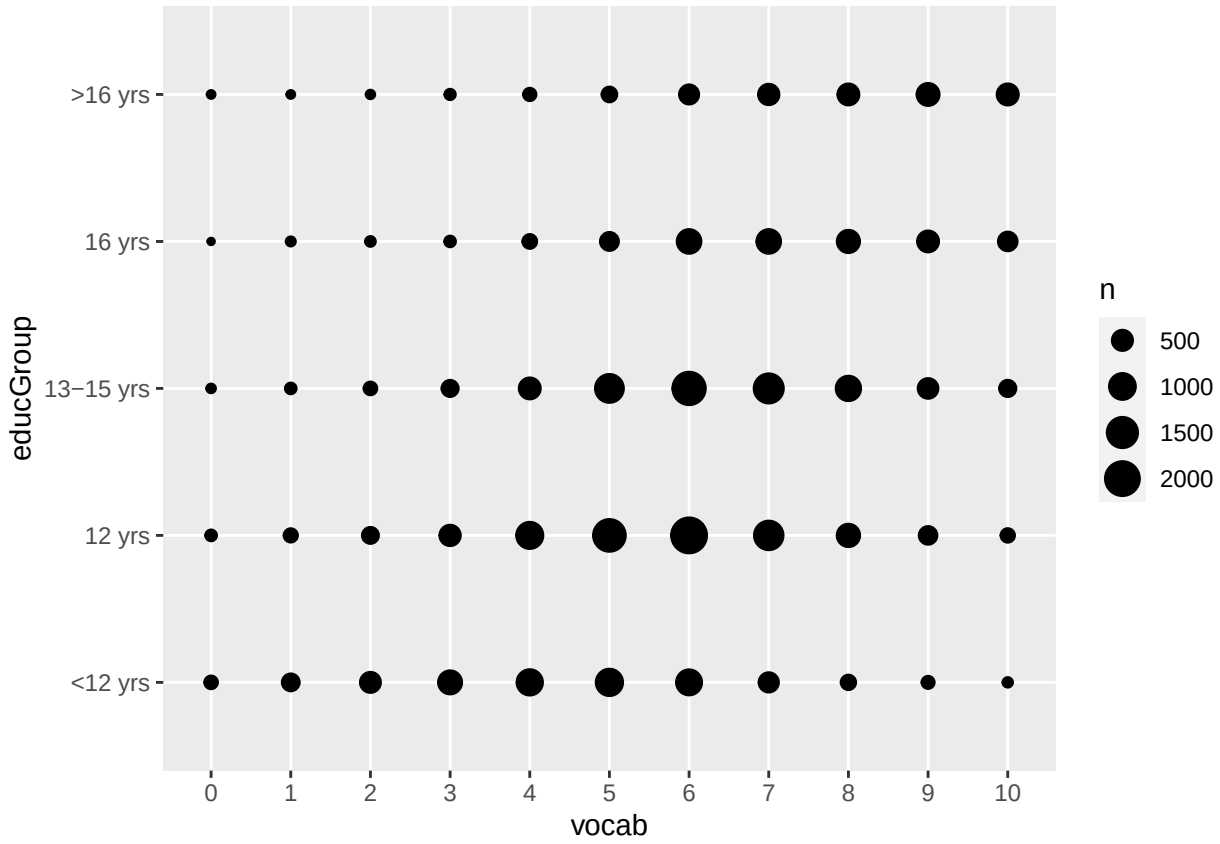


Create two different plots and identify the best-looking plot you can to examine the `vocab` variable by `educGroup`. Does there appear to be an association?

```
ggplot(X) +
  geom_jitter(aes(x = vocab, y = educGroup))
```

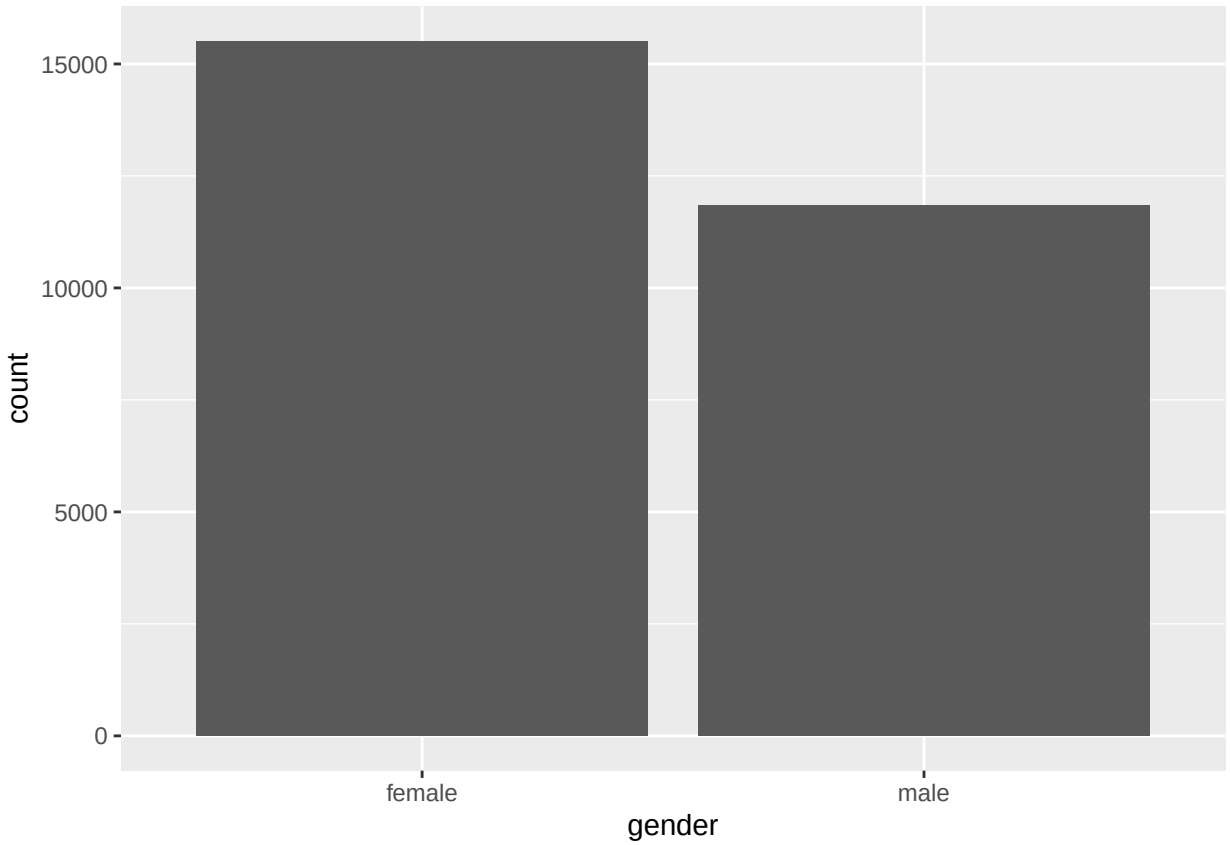



```
ggplot(X) +
  geom_count(aes(x = vocab, y = educGroup))
```



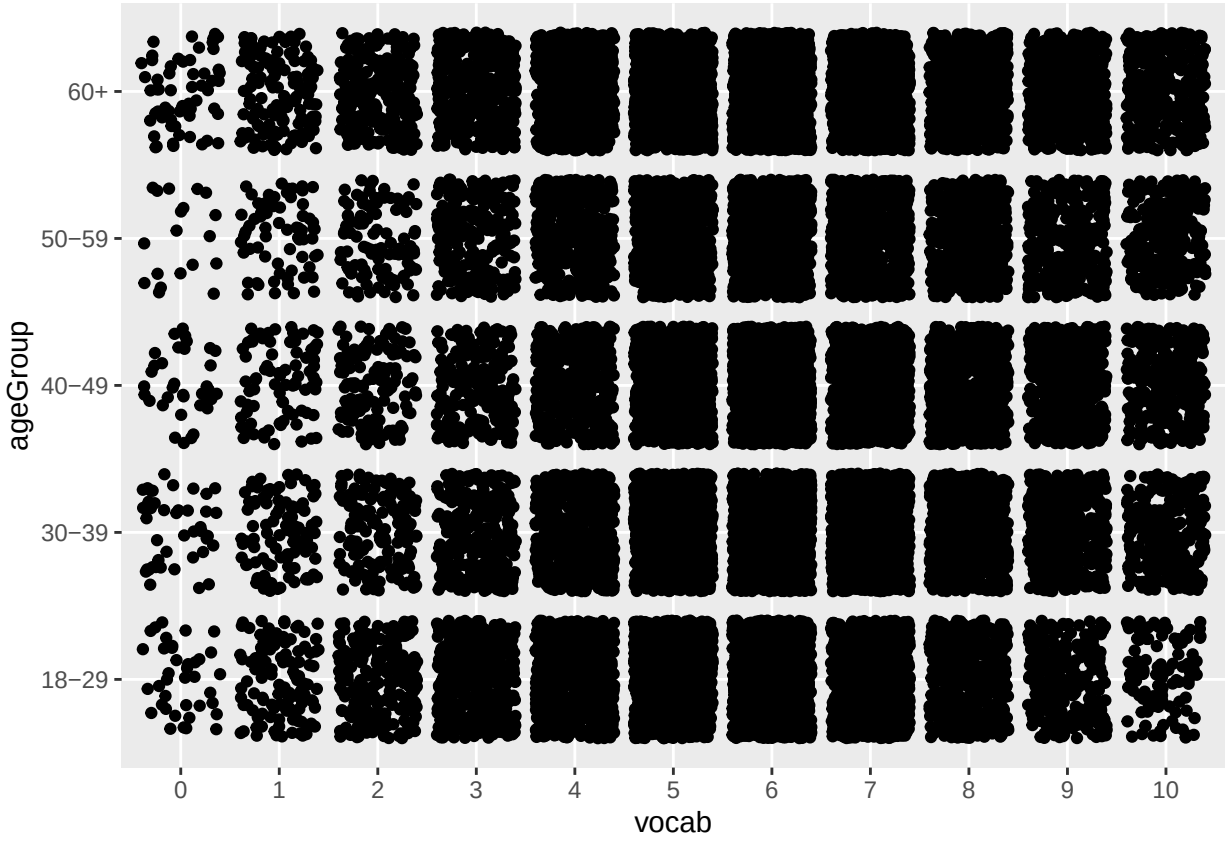
Using the best-looking plot from the previous question, create the best looking overloading with variable `gender`. Does there appear to be an interaction of `gender` and `educGroup`?

```
ggplot(X) +
  geom_bar(aes(x = gender))
```



Using facets, examine the relationship between `vocab` and `ageGroup`. You can drop year level `(Other)`. Are we getting dumber?

```
ggplot(X) +
  geom_jitter(aes(x = vocab, y = ageGroup))
```


60+
50–59
40–49
30–39
18–29
ageGroup
0
1
2
3
4
5
6
7
8
9
10
vocab