Lab 6

Asher Katz 4/10/2022

#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.

```
#TO-DO
pacman::p_load(carData)
X = carData:: GSSvocab
X = na.omit(X)
summary(X)
```

```
##       year           gender      nativeBorn   ageGroup        educGroup
##  2016   : 1856   female:15512   no : 2342   18-29:5691   <12 yrs  :5264
##  1996   : 1855   male  :11848   yes:25018   30-39:6024   12 yrs   :8259
##  1994   : 1842                               40-49:5035   13-15 yrs:6942
##  1982   : 1714                               50-59:4113   16 yrs   :3814
##  1987   : 1659                               60+  :6497   >16 yrs  :3081
##  2014   : 1650
##  (Other):16784
##      vocab          age             educ
##  Min.   : 0   Min.   :18.00   Min.   : 0.00
##  1st Qu.: 5   1st Qu.:31.00   1st Qu.:12.00
##  Median : 6   Median :43.00   Median :13.00
##  Mean   : 6   Mean   :45.75   Mean   :13.16
##  3rd Qu.: 7   3rd Qu.:59.00   3rd Qu.:16.00
##  Max.   :10   Max.   :89.00   Max.   :20.00
##
```

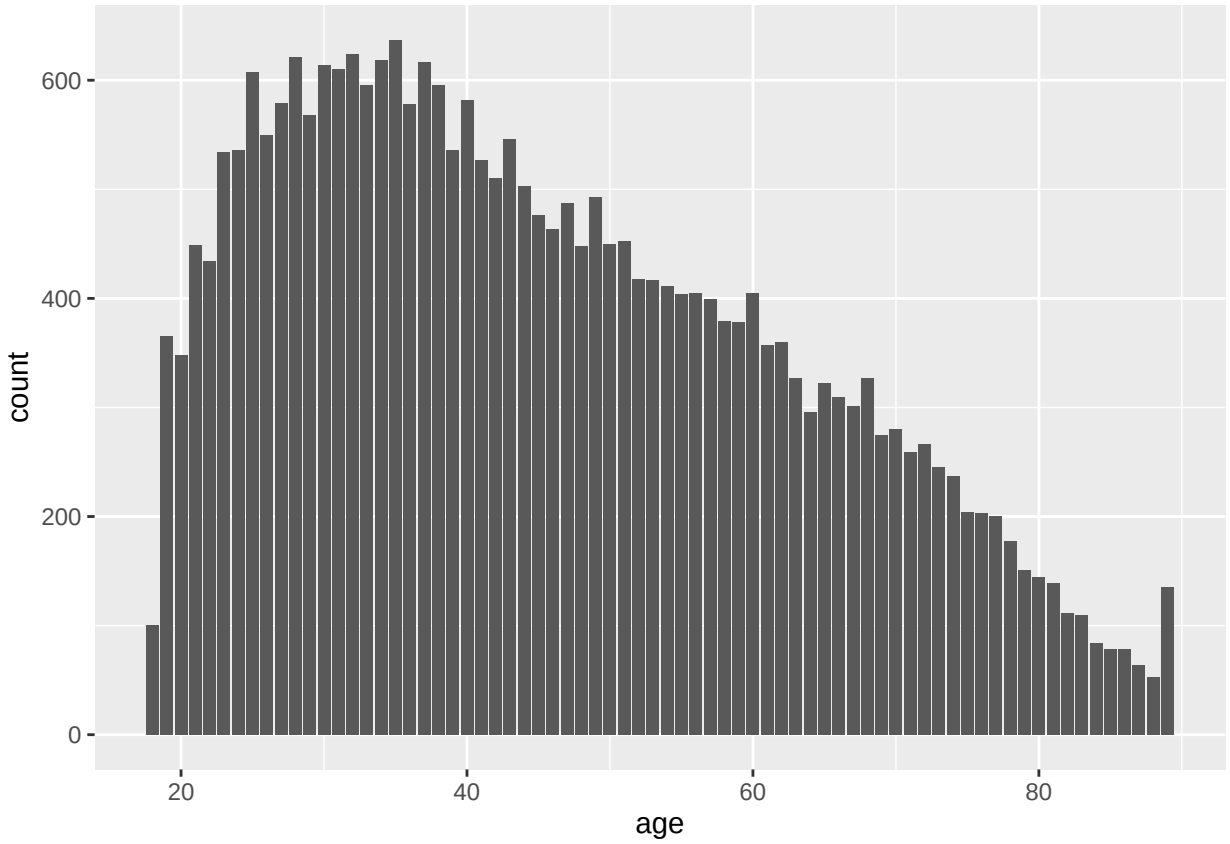
Briefly summarize the documentation on this dataset. What is the data type of each variable? What do you think is the response variable the collectors of this data had in mind?

This data set illustrates analysis of a multifactor observational study, with response given by subject's score on a vocabulary test, and factors for age group, education level, natality status, gender and year of the survey.

year: ordinal variable with ascending years 1978-2016 gender: nominal catagorical variable nativeBorn: binary catagorical variable ageGroup: ordinal catagorical variable educGroup: ordinal catagorical variable vocab: ordinal catagorical variable age: continuous numeric variable educ:continuous numeric variable

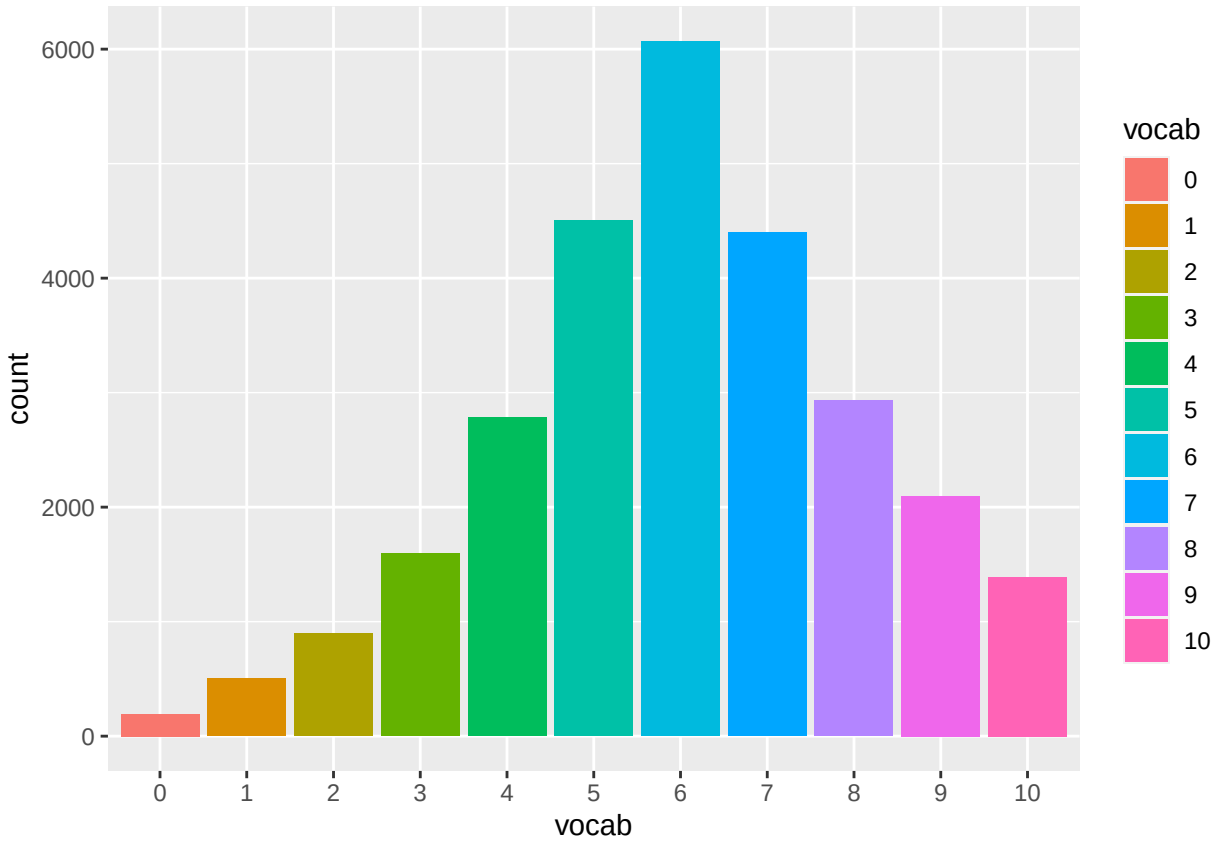
Create two different plots and identify the best-looking plot you can to examine the `age` variable. Save the best looking plot as an appropriately-named PDF.

```
pacman::p_load(ggplot2)
base = ggplot(na.omit(X), aes(age))
base + geom_bar()
```



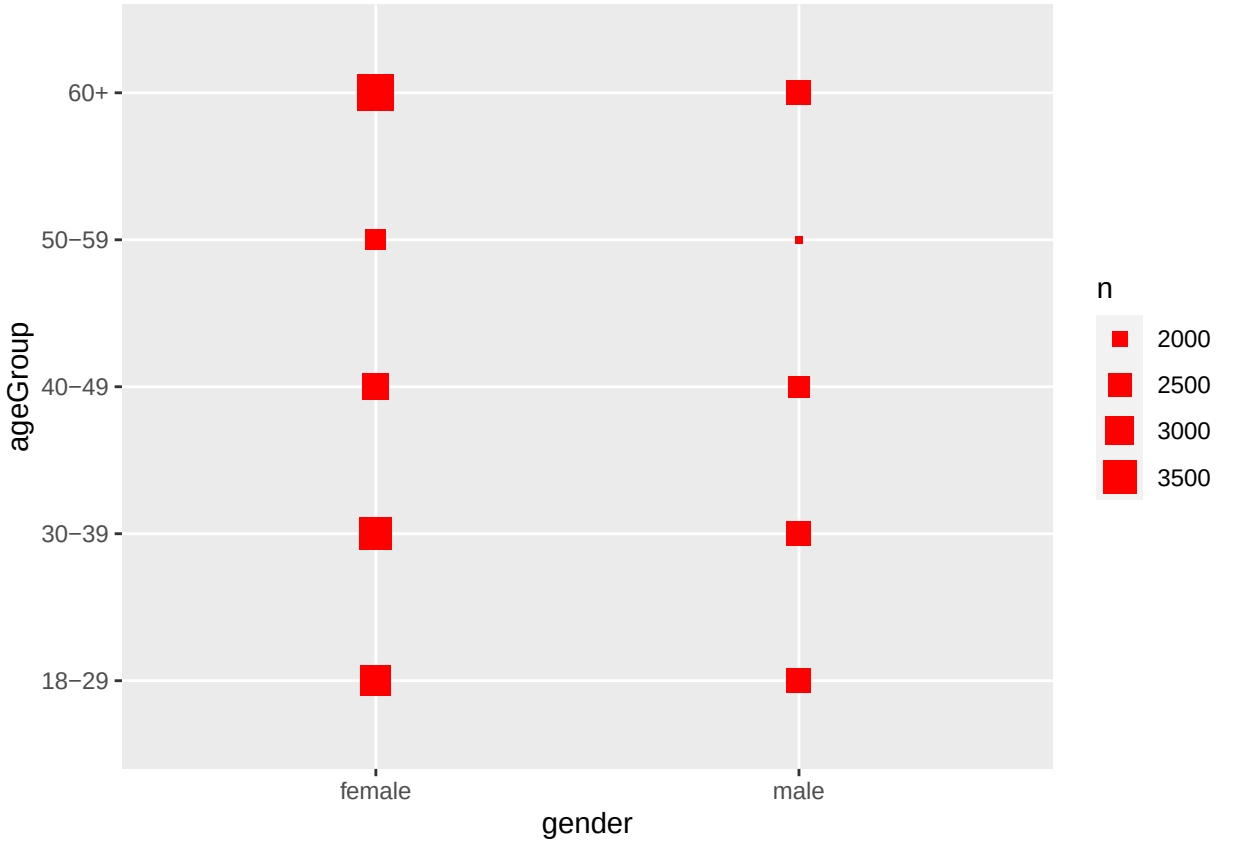
Create two different plots and identify the best looking plot you can to examine the `vocab` variable. Save the best looking plot as an appropriately-named PDF.

```
#TO-DO
X$vocab = factor(X$vocab)
base = ggplot(X, aes(vocab, fill = vocab))
#base + geom_histogram()
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.

```
#TO-DO
base = ggplot(data = X, mapping = aes(x = gender , y = ageGroup ))
base +
  geom_count(color = "red", shape = 15)
```

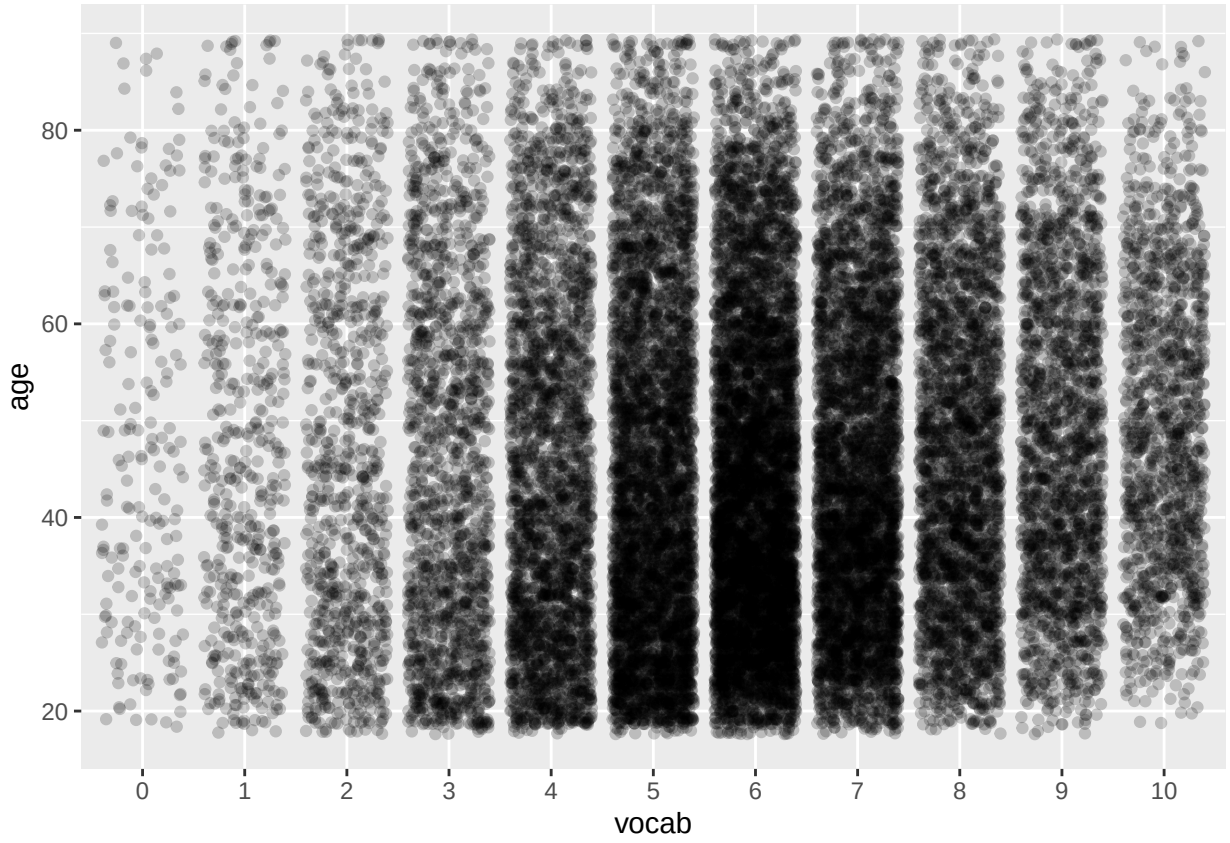



```
  #scale_fill_brewer(palette = "Blues")
  #geom_smooth(method = lm, x = X$gender, y = X$ageGroup, size = 0.00001)
  #geom_jitter(height = 2, width = 2, shape = 1, size = 0.00001)
```

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

Yes

```
#TO-DO
ggplot(X)+
  geom_jitter(alpha = .2, aes(x=vocab, y = age))
```

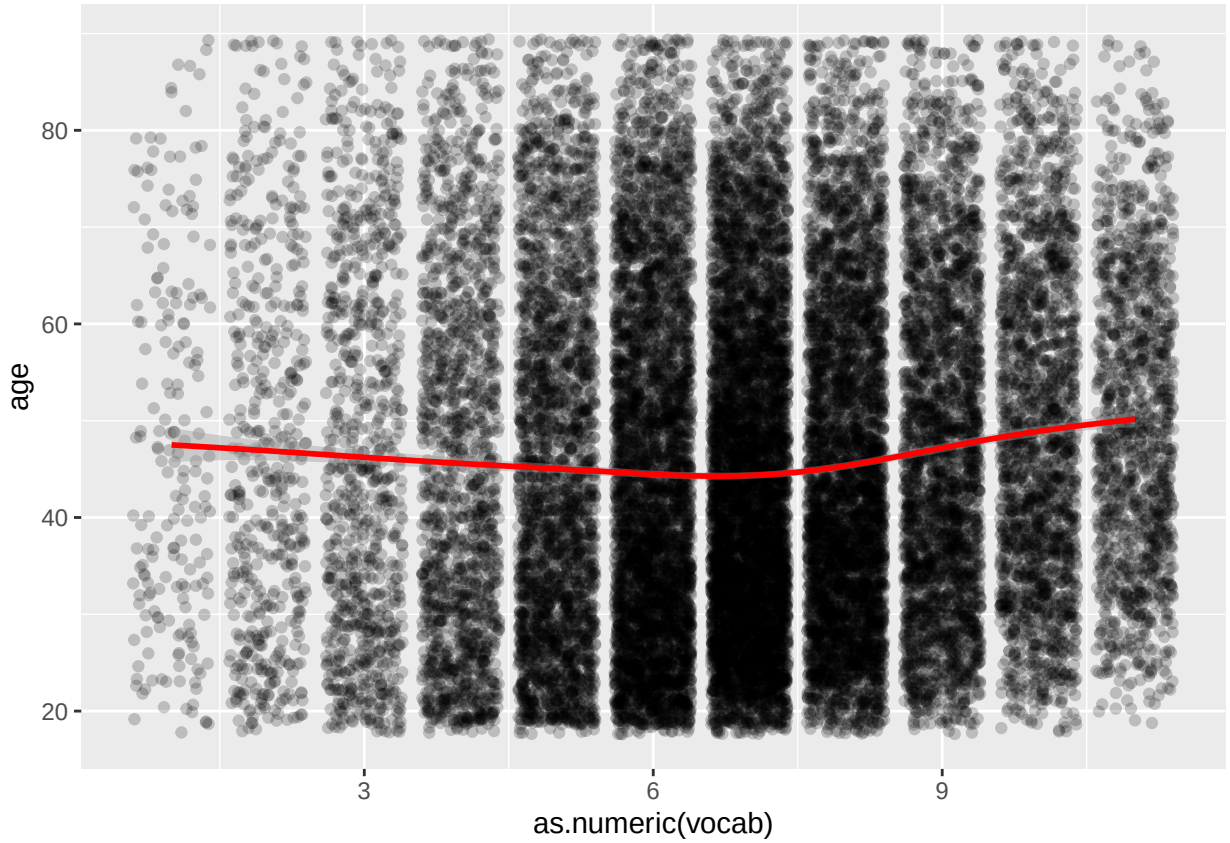



```
  #geom_boxplot(aes(x=vocab, y = age))
```

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

```
#TO-DO
ggplot(X, aes(x= as.numeric(vocab), y = age)) +
  geom_jitter(alpha = .2) +
  geom_smooth(color = "red")
```

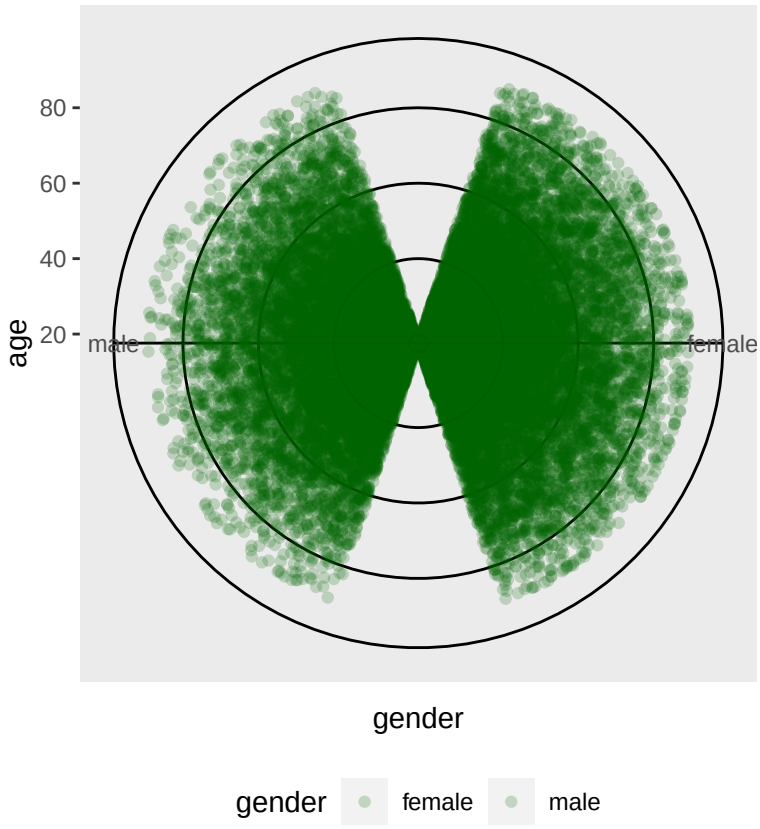
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



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`?

```
#TO-DO
ggplot(X, aes(x=gender, y = age, fill = gender, label = age))+
  coord_polar(theta = "x", direction = 1) +
  geom_jitter(alpha = .2, scale = "area", color = "darkgreen", )+
  theme(panel.grid.major=element_line(colour="black"), panel.grid.minor=element_line(colour="black")) +
  theme(legend.position = "bottom")
```

```
## Warning: Ignoring unknown parameters: scale
```

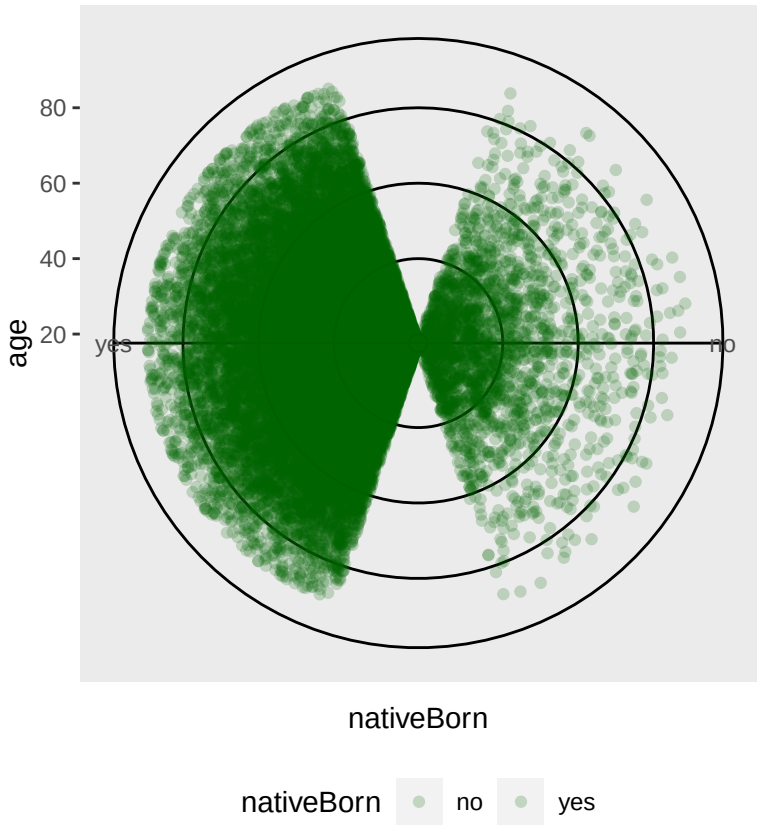


Using the plot from the previous question, create the best looking plot overloading with variable `nativeBorn`. Does there appear to be an interaction of `nativeBorn` and `age`?

yes

```
#TO-DO
ggplot(X, aes(x=nativeBorn, y = age, fill = nativeBorn, label = age))+
  coord_polar(theta = "x", direction = 1) +
  geom_jitter(alpha = .2, scale = "area", color = "darkgreen", )+
  theme(panel.grid.major=element_line(colour="black"), panel.grid.minor=element_line(colour="black")) +
  theme(legend.position = "bottom")
```

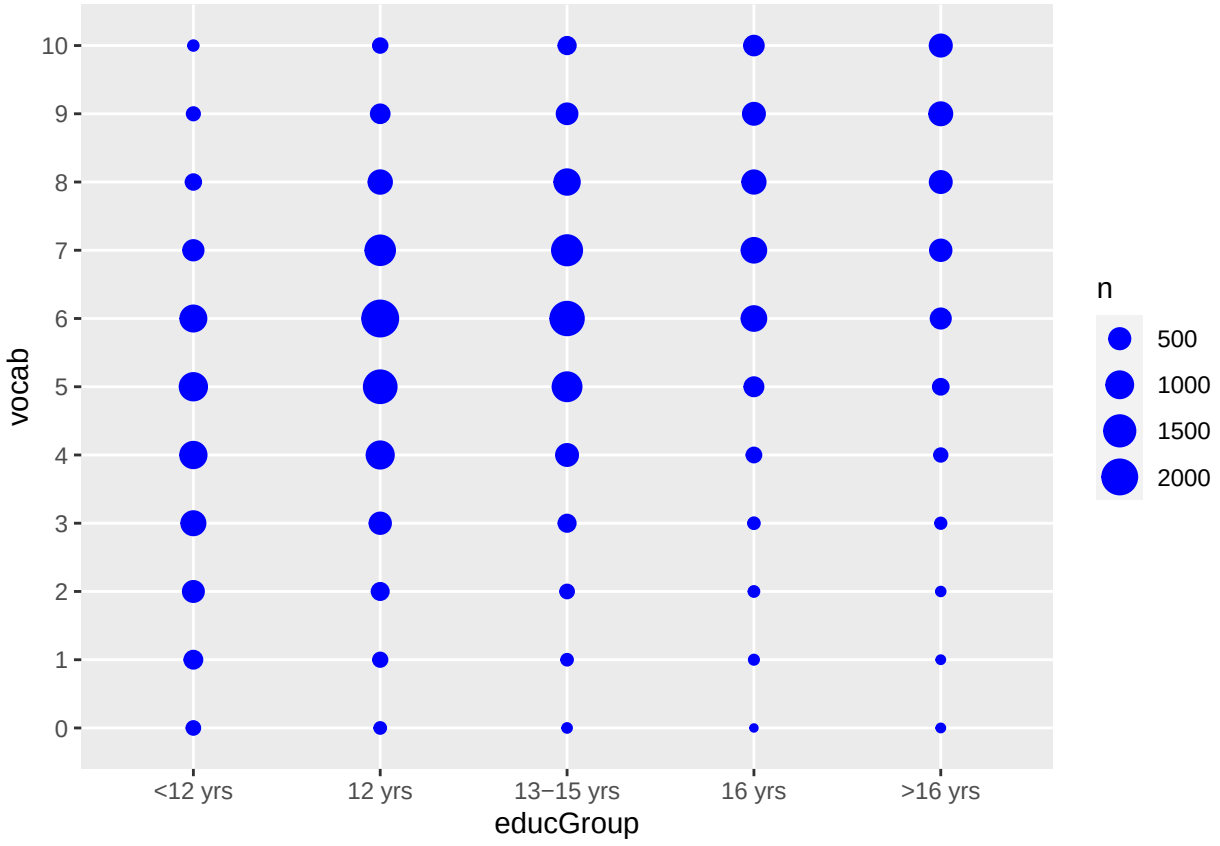
```
## Warning: Ignoring unknown parameters: scale
```



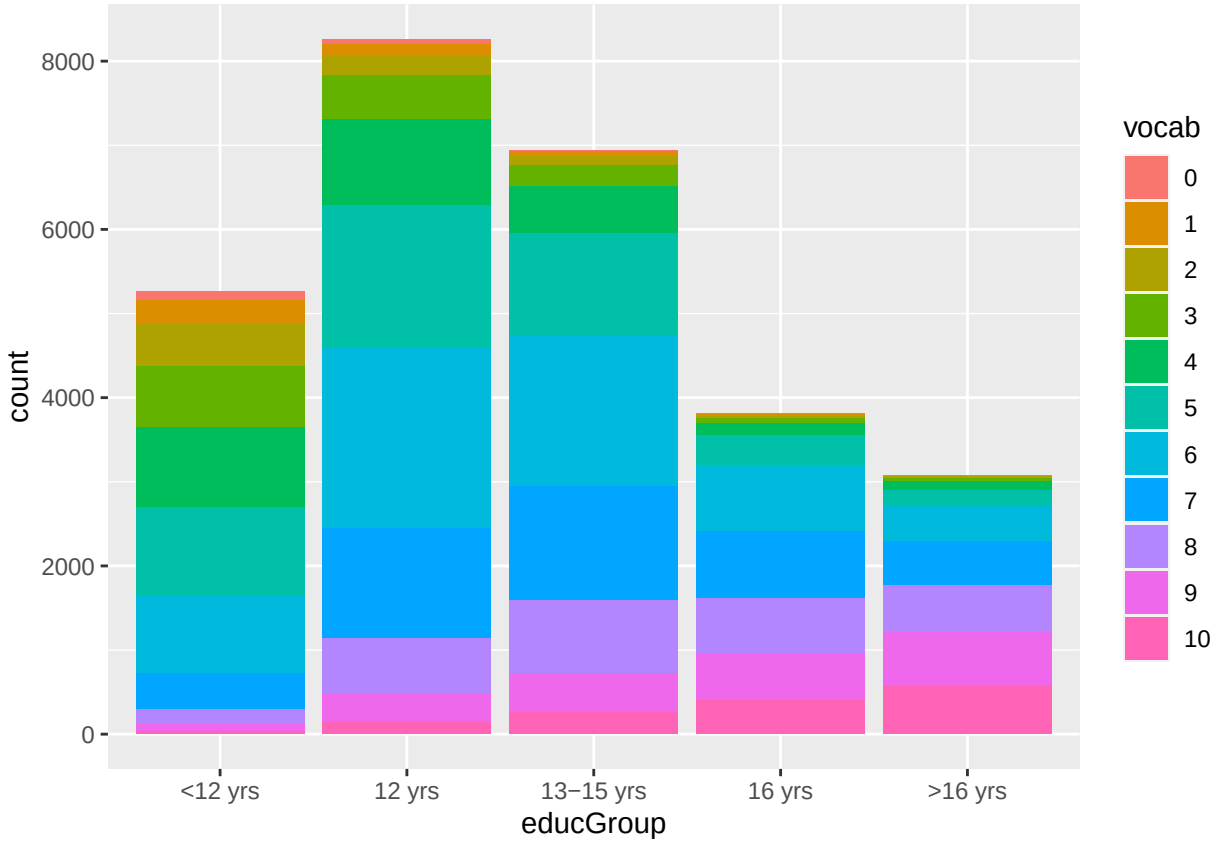
Create two different plots and identify the best-looking plot you can to examine the `vocab` variable by `educGroup`. Does there appear to be an association?

yes

```
#TO-DO
ggplot(X, aes(x = educGroup, y = vocab)) +
  geom_count(color = "blue")
```




```
ggplot(X, aes(x = educGroup, fill = vocab)) +
  geom_bar()
```

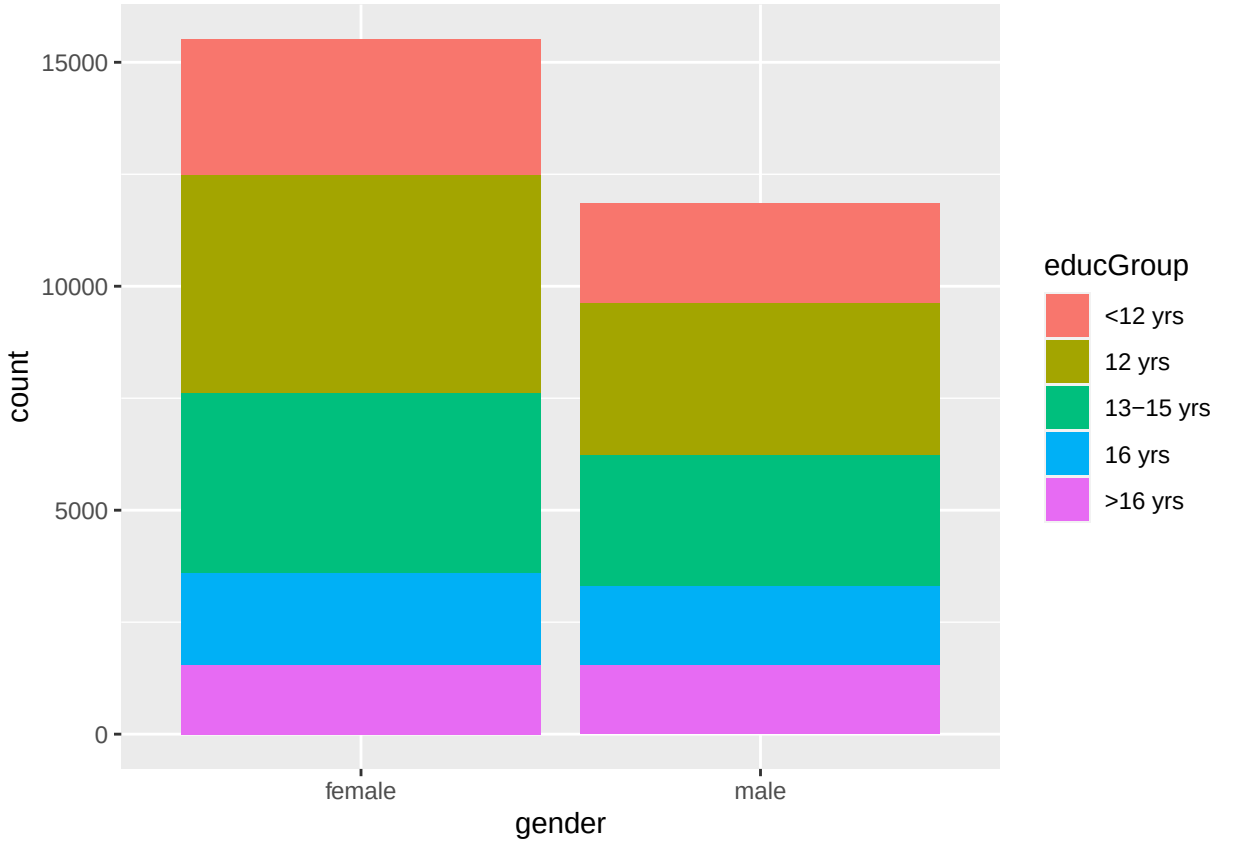


Using the best-looking plot from the previous question, create the best looking overloading with variable `gender`. Does there appear to be an interaction of `gender` and `educGroup`?

yes

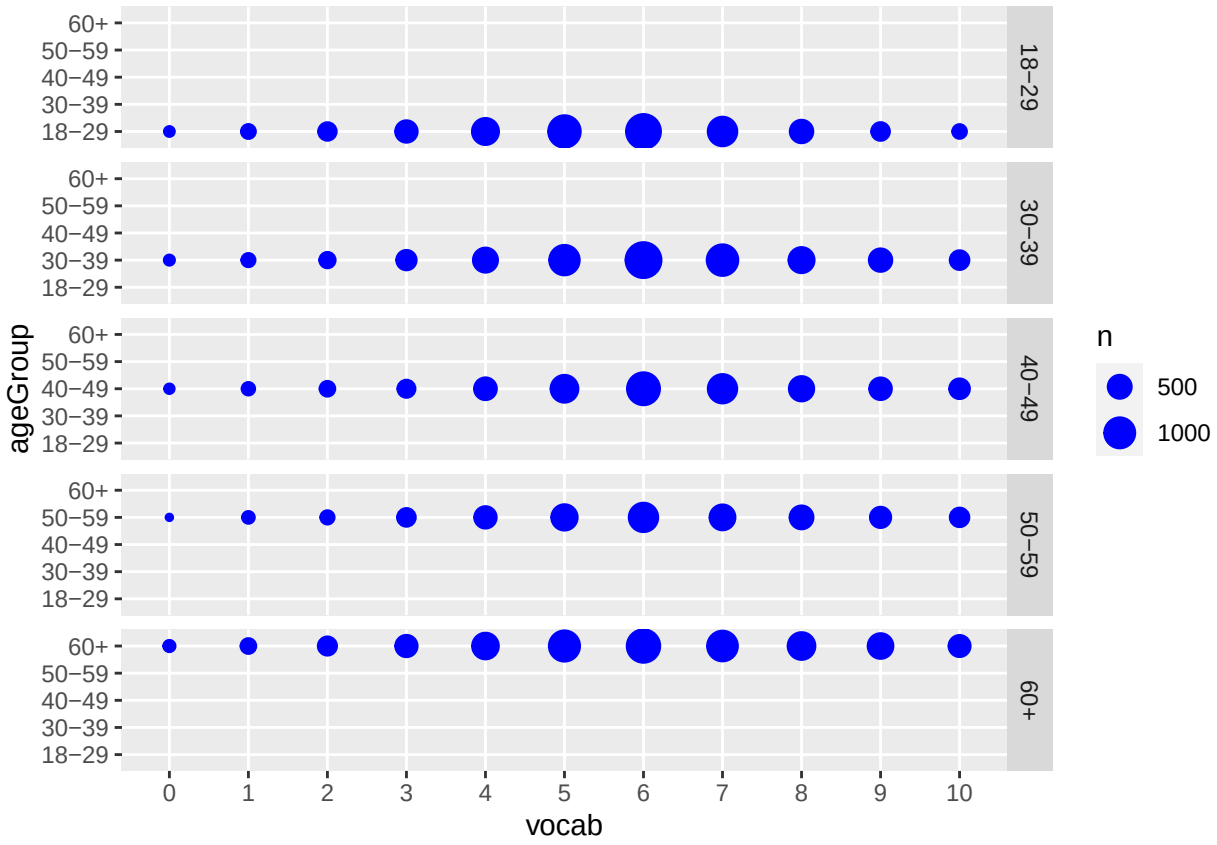
```
#TO-DO

ggplot(X, aes(x = gender, fill = educGroup)) +
  geom_bar()
```



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

```
#TO-DO
ggplot(X, aes(x = vocab, y = ageGroup)) +
  geom_count( color = "blue")+
  facet_grid(rows = vars(ageGroup))
```


ageGroup
60+
50–59
40–49
30–39
18–29
18–29
30–39
40–49
50–59
60+
0
1
2
3
4
5
6
7
8
9
10
vocab
n
500
1000