Lab 6

Brendan Gubbins

11:59PM April 15, 2021

#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, 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.

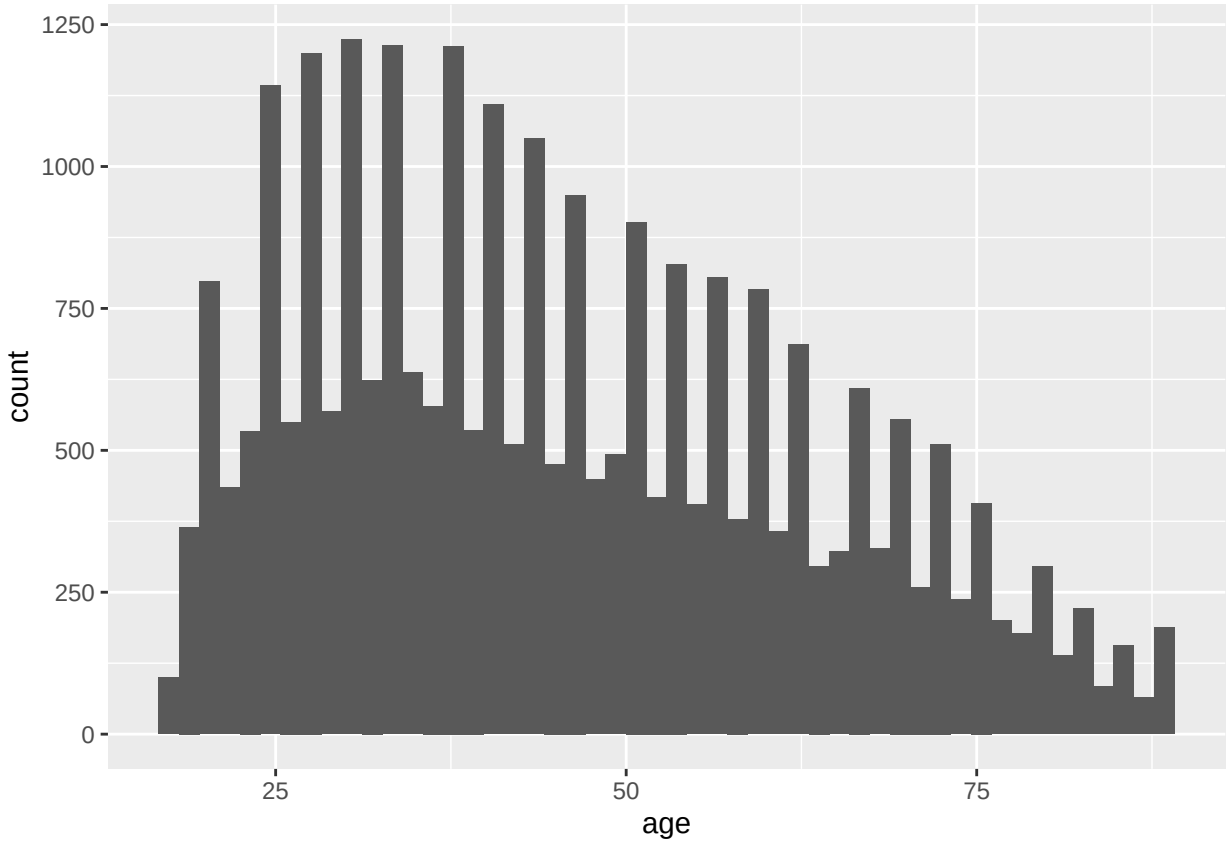
```
pacman::p_load(carData)
data(GSSvocab)
GSSvocab = na.omit(GSSvocab)
```

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?

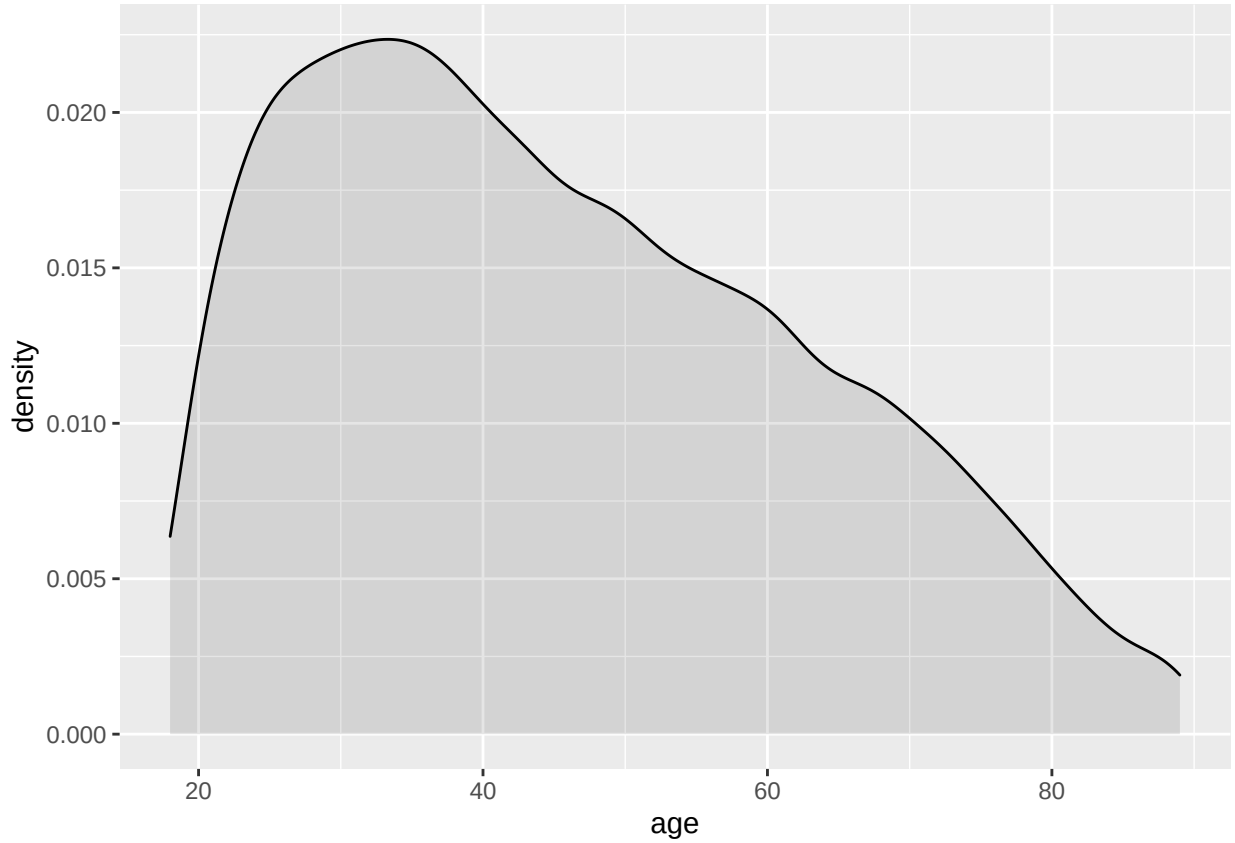
The dataset consists of variables year (factor), gender (factor), nativeBorn (factor), ageGroup (factor), educGroup (factor), vocab (numeric), age (numeric), and educ (numeric). The data aims to use information about the person, their gender, their age, their native status, and their educational experience to collect a response variable vocab, which scores the number of words (out of 10) correct on a vocabulary test.

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)

# better graph
ggplot(GSSvocab) +
  aes(x = age) +
  geom_histogram(bins = 50)
```

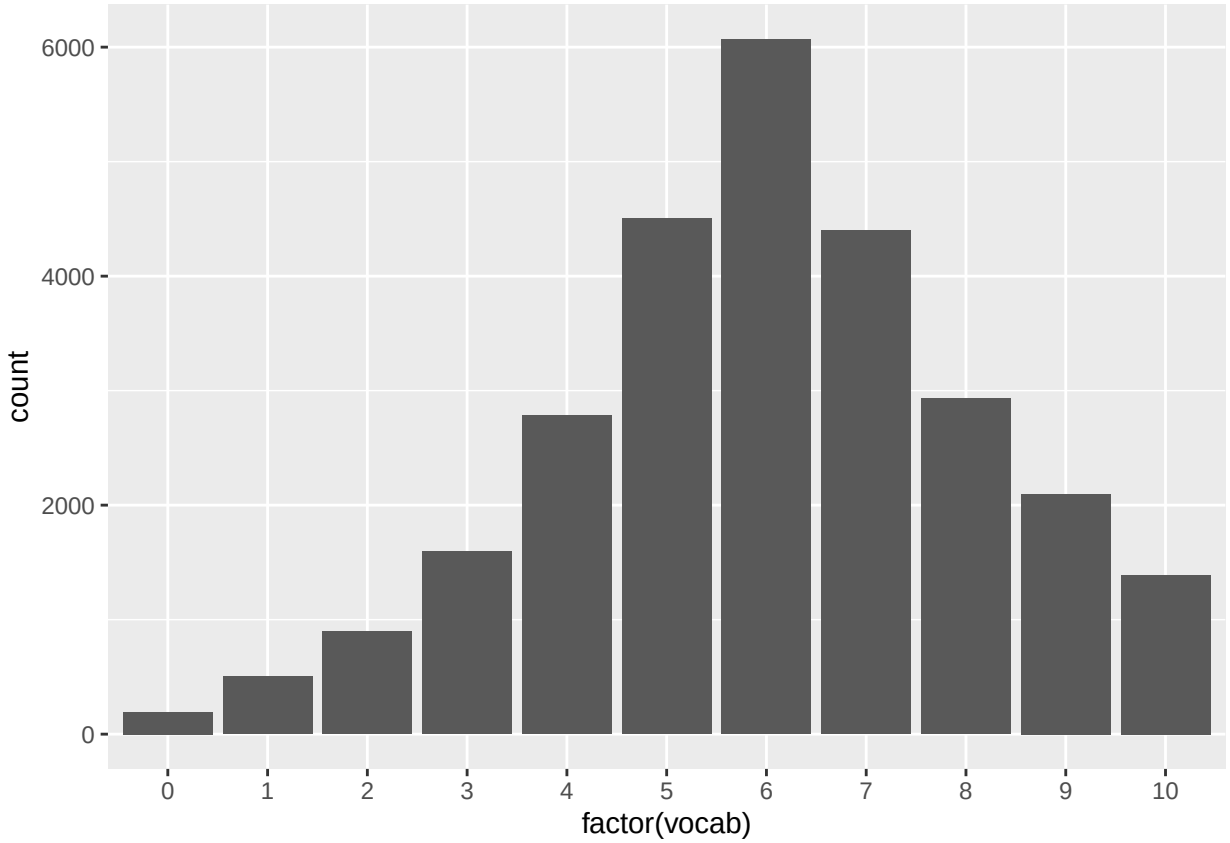



```
ggplot(GSSvocab) +
  aes(x = age) +
  geom_density(fill = "black", alpha = .1)
```

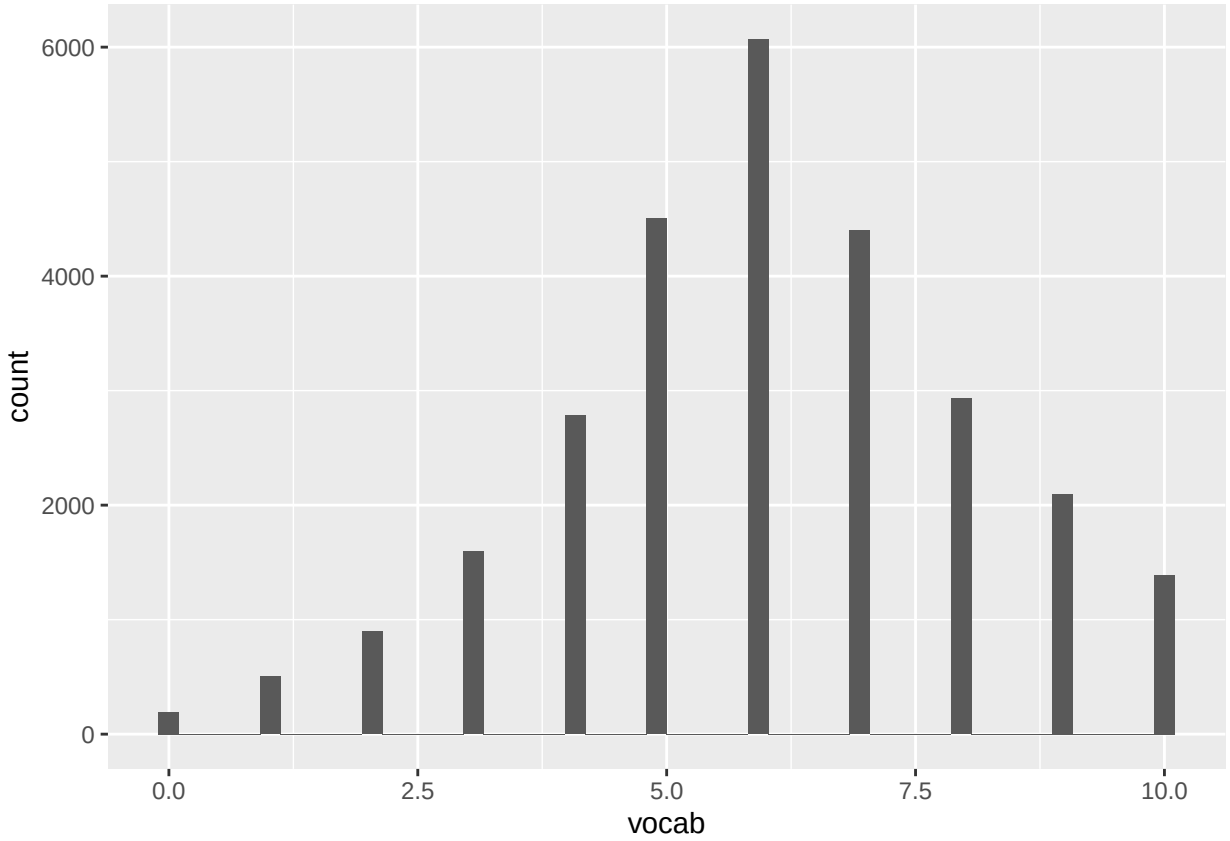


Create two different plots and identify the best looking plot you can to examine the `vocab` variable. Save the best looking plot as an appropriately-named PDF.

```
# better graph
ggplot(GSSvocab) +
  aes(x = factor(vocab)) +
  geom_bar()
```

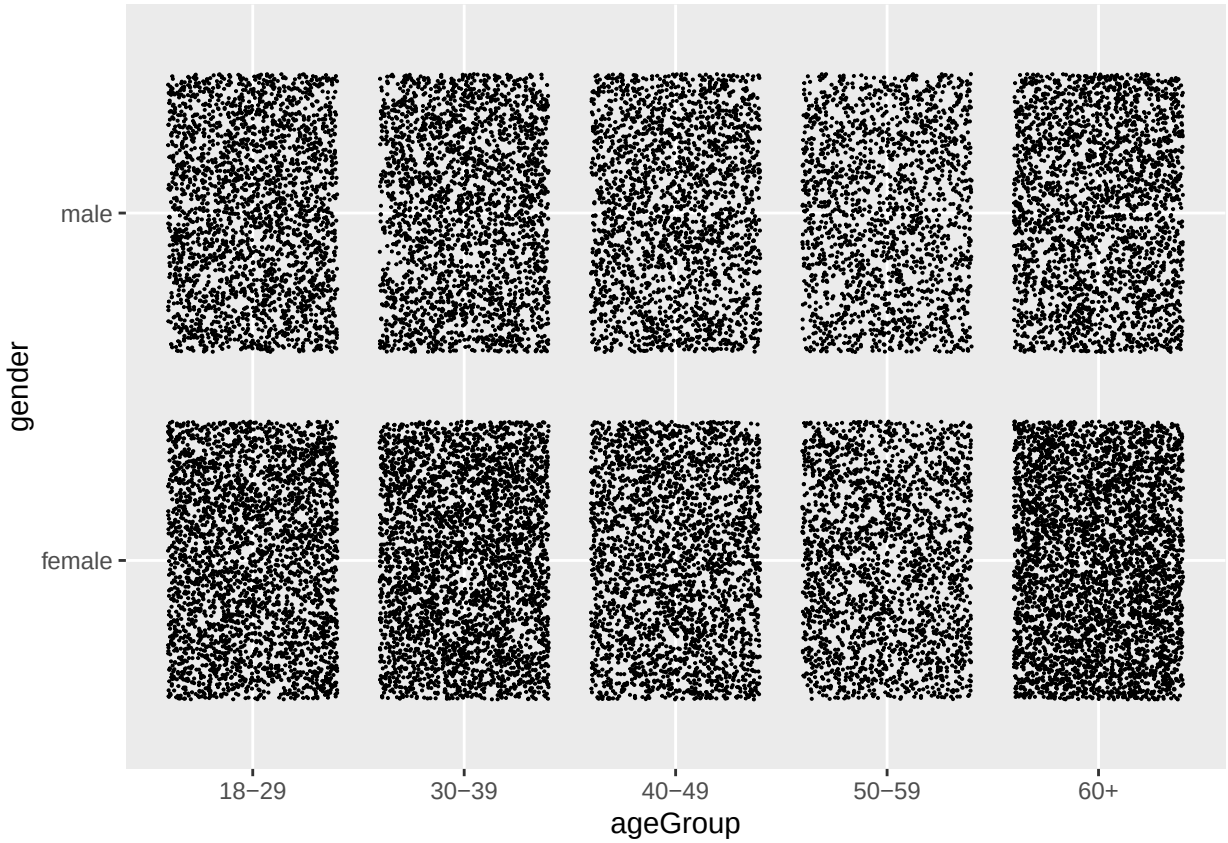



```
ggplot(GSSvocab) +
  aes(x = vocab) +
  geom_histogram(bins = 50)
```



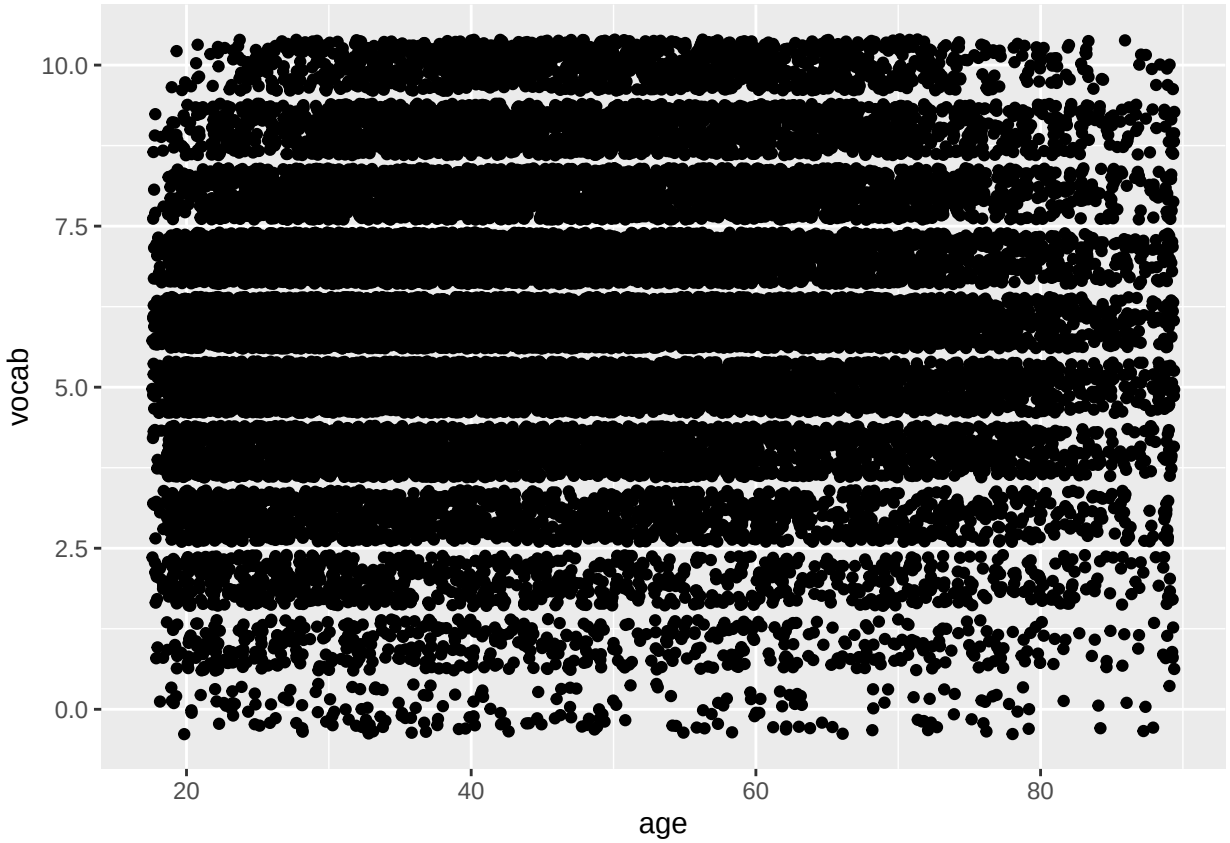
Create the best-looking plot you can to examine the `ageGroup` variable by `gender`. Does there appear to be an association? There are many ways to do this.

```
ggplot(GSSvocab) +
  aes(x = ageGroup, y = gender) +
  geom_jitter(size = .05)
```



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

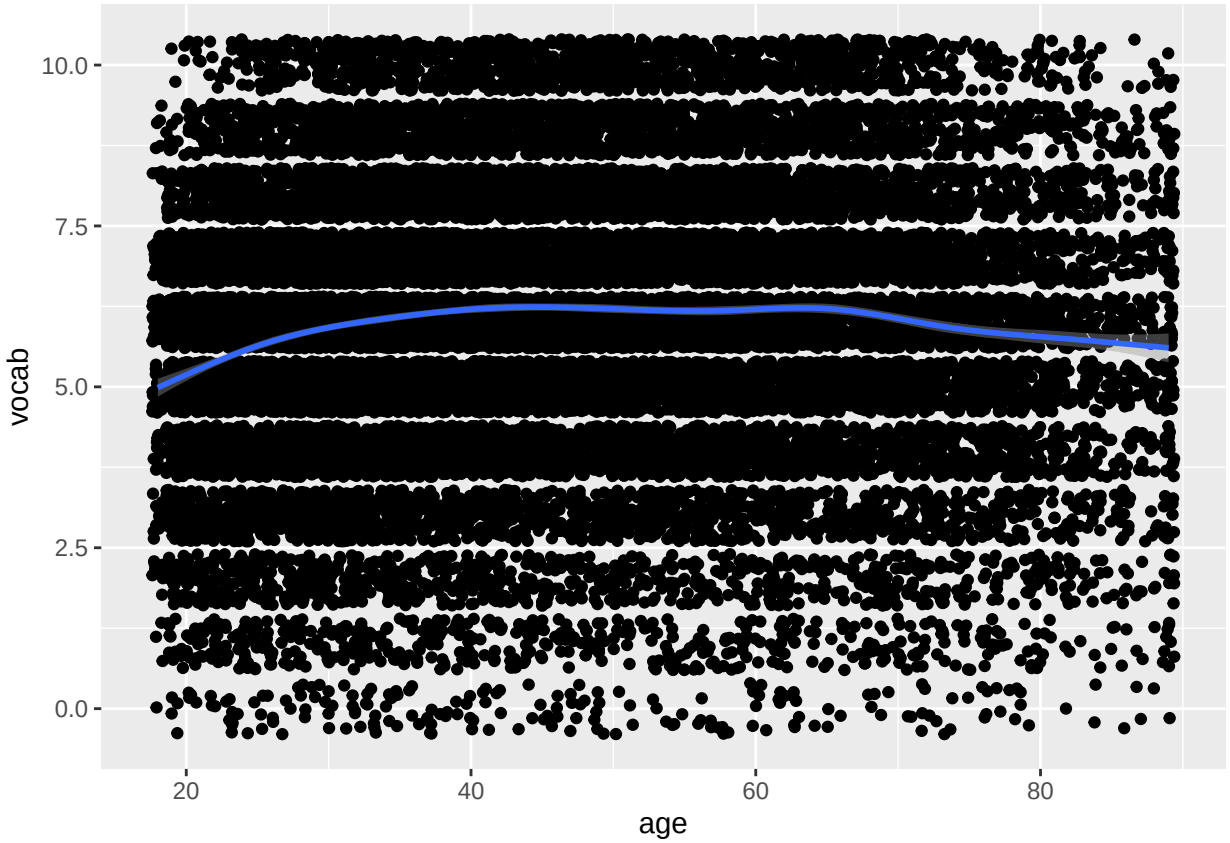
```
ggplot(GSSvocab) +
  aes(x = age, y = vocab) +
  geom_jitter()
```



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

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

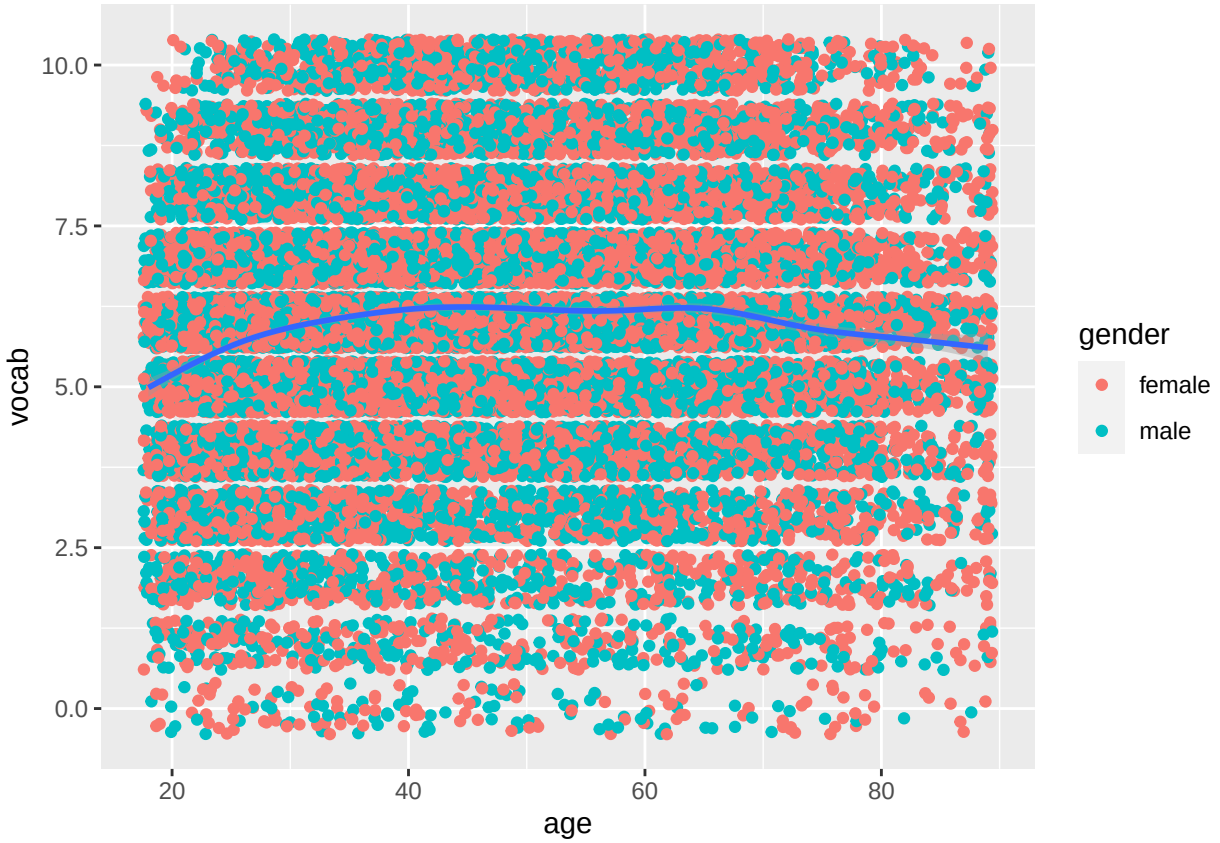
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



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(GSSvocab) +
  aes(x = age, y = vocab) +
  geom_jitter(aes(col = gender)) +
  geom_smooth()
```

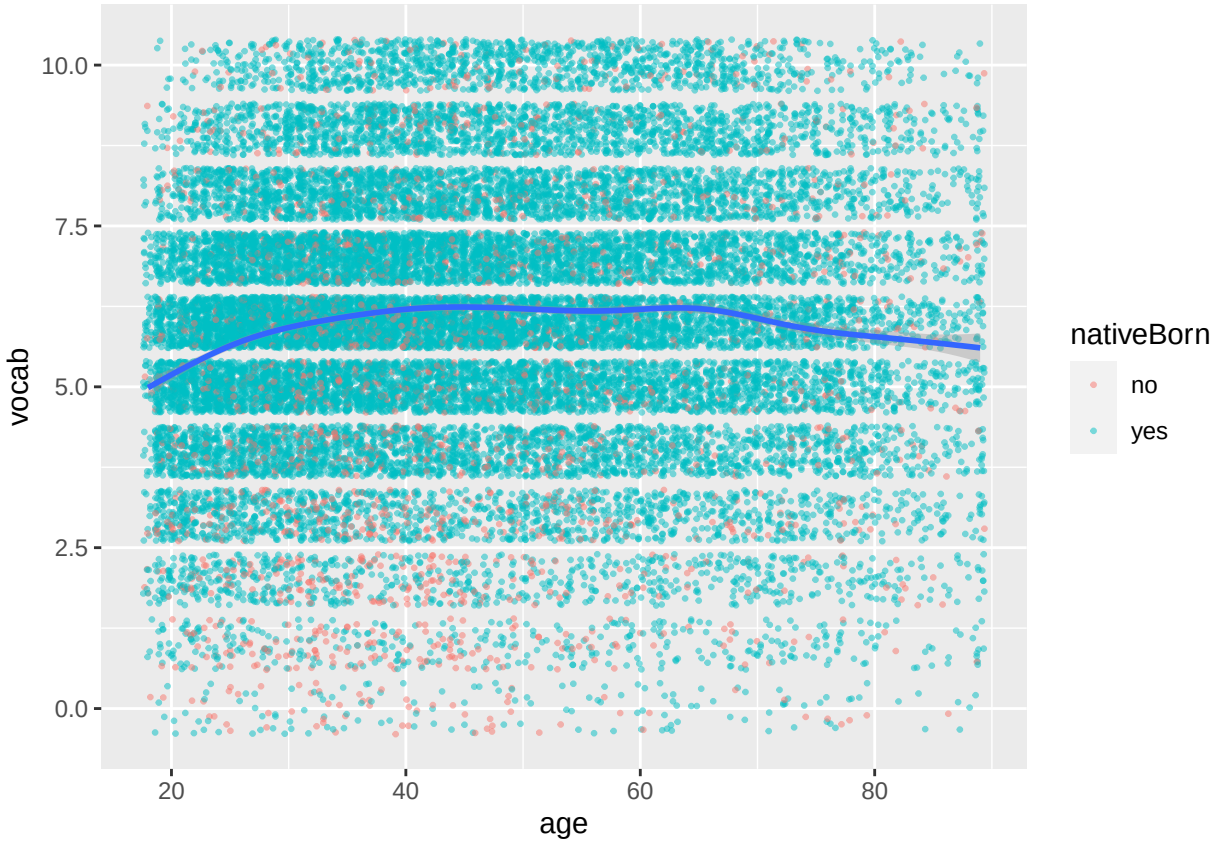
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



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

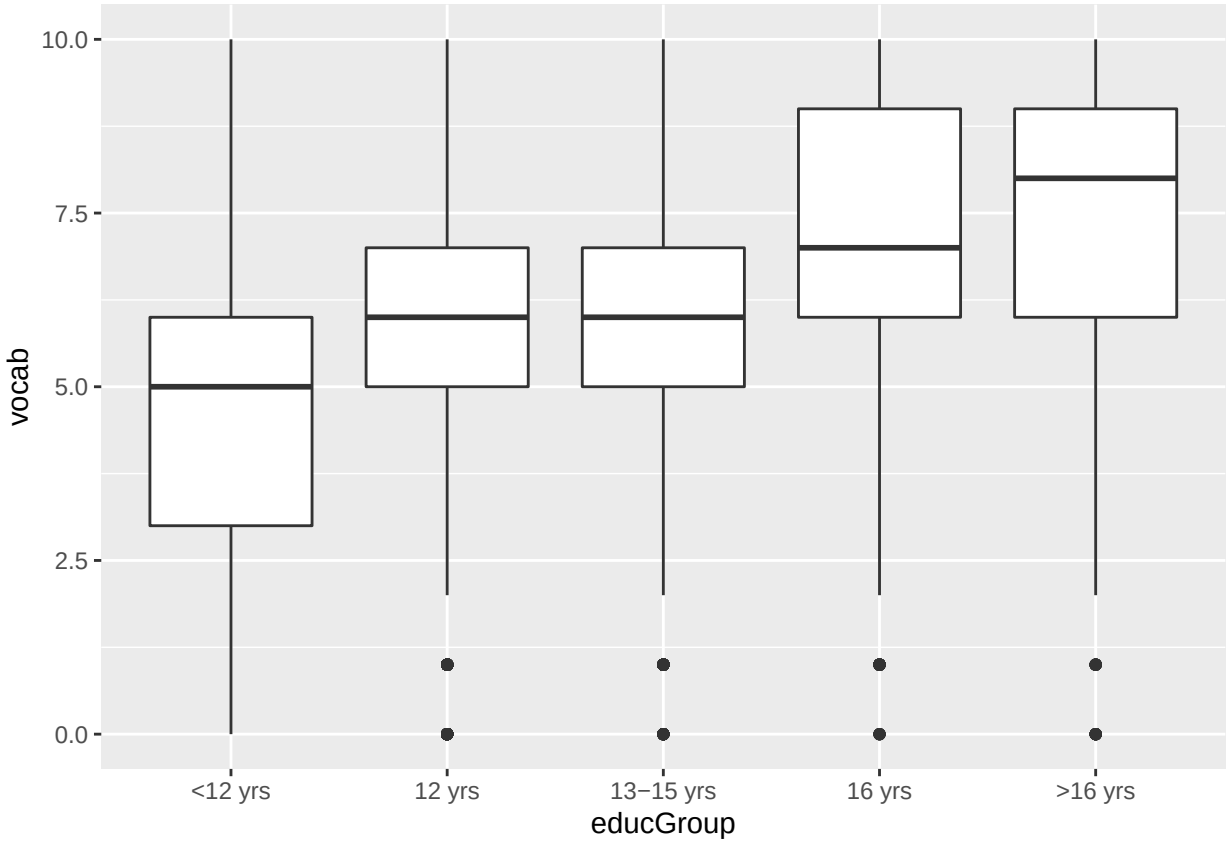
```
ggplot(GSSvocab) +
  aes(x = age, y = vocab) +
  geom_jitter(aes(col = nativeBorn), size = .5, alpha = .5) +
  geom_smooth()
```

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

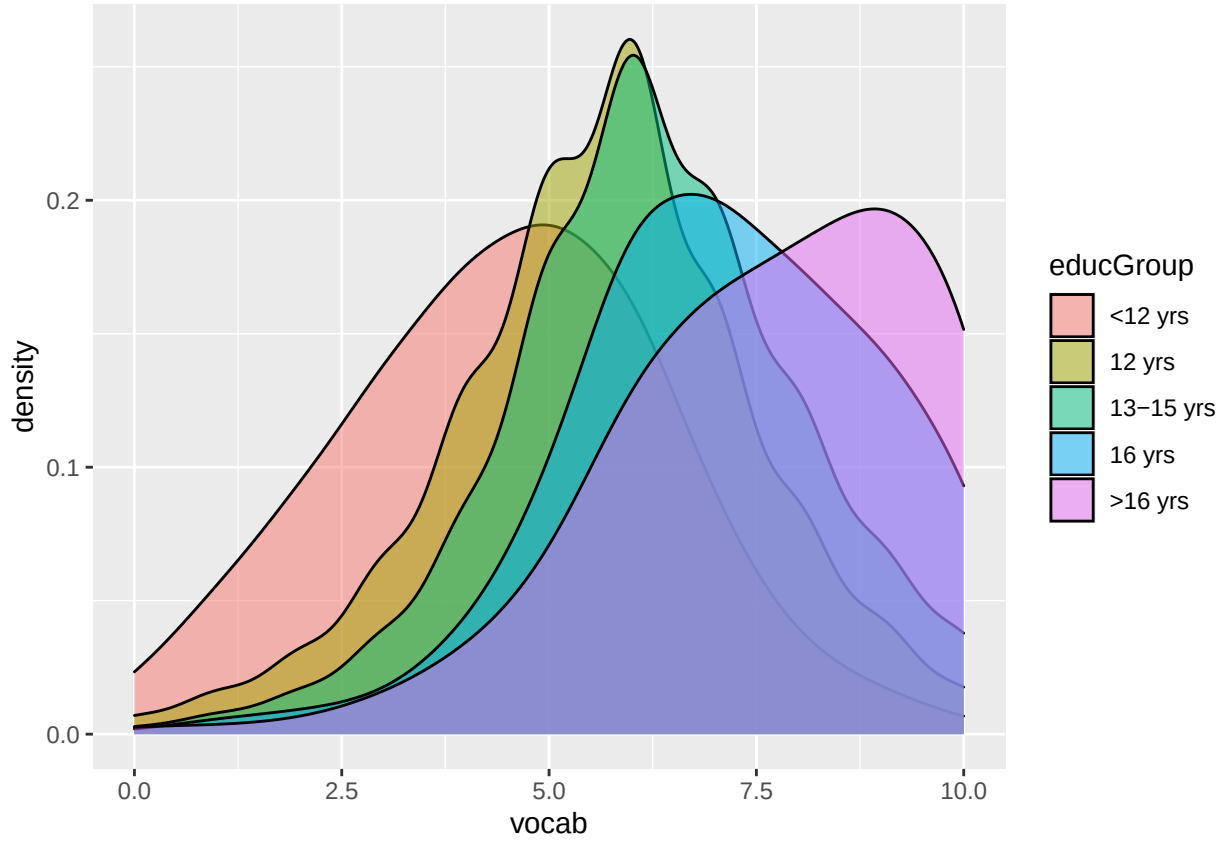


Create two different plots and identify the best-looking plot you can to examine the `vocab` variable by `educGroup`. Does there appear to be an association?

```
ggplot(GSSvocab) +
  aes(x = educGroup, y = vocab) +
  geom_boxplot()
```

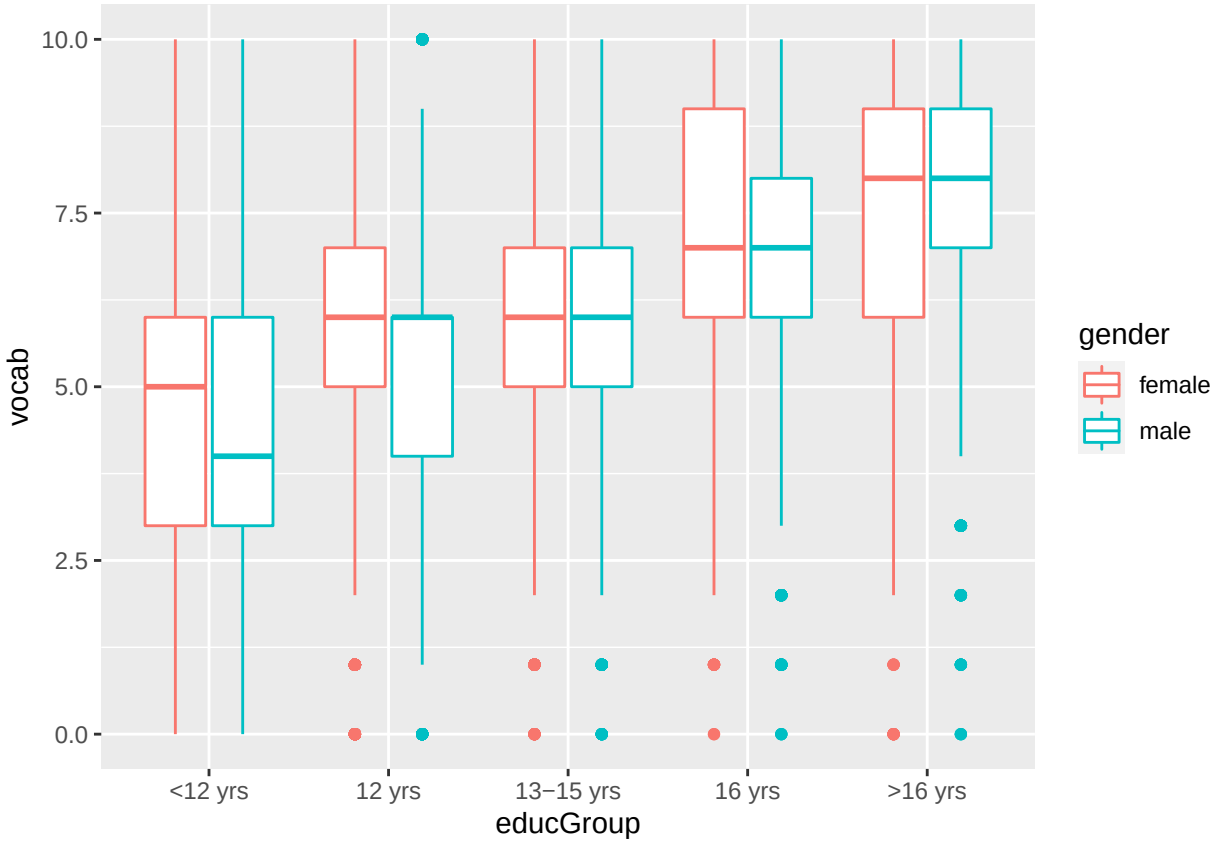



```
# better graph
ggplot(GSSvocab) +
  aes(x = vocab) +
  geom_density(aes(fill = educGroup), adjust = 2, alpha = .5)
```



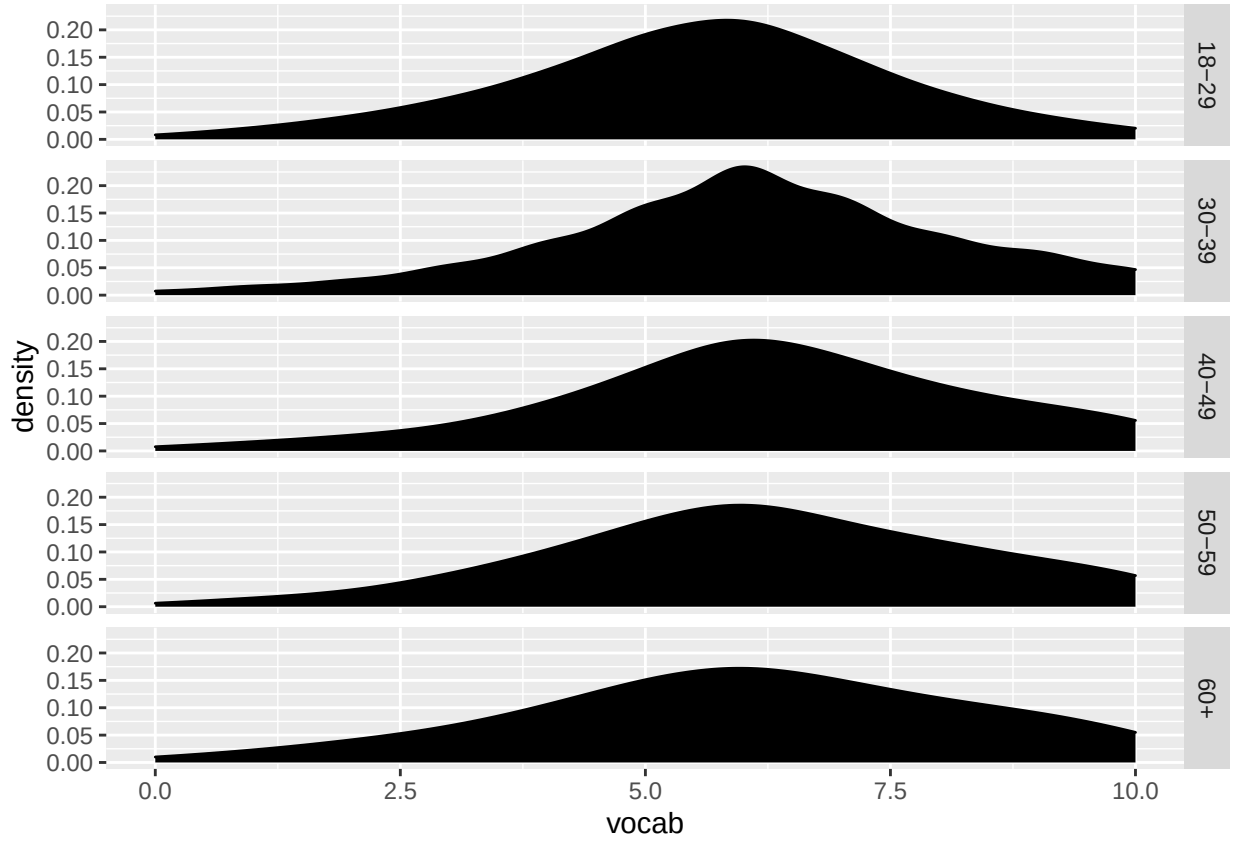
Using the best-looking plot from the previous question, create the best looking overloading with variable `gender`. Does there appear to be an interaction of `gender` and `educGroup`?

```
ggplot(GSSvocab) +
  aes(x = educGroup, y = vocab) +
  geom_boxplot(aes(col = gender))
```



Using facets, examine the relationship between `vocab` and `ageGroup`. Are we getting dumber?

```
ggplot(GSSvocab) +
  aes(x = vocab) +
  geom_density(adjust = 2, fill = "black") +
  facet_grid(ageGroup~.)
```



Probability Estimation and Model Selection

Load up the `adult` in the package `ucidata` dataset and remove missingness and the variable `fnlwgt`:

```
pacman::p_load_gh("coatless/ucidata")
data(adult)
adult = na.omit(adult) #kill any observations with missingness
adult$fnlwgt = NULL
```

Cast income to binary where 1 is the `>50K` level.

```
adult$income = ifelse(adult$income == ">50K", 1, 0)
```

We are going to do some dataset cleanup now. But in every cleanup job, there's always more to clean! So don't expect this cleanup to be perfect.

Firstly, a couple of small things. In variable `marital_status` collapse the levels `Married-AF-spouse` (armed force marriage) and `Married-civ-spouse` (civilian marriage) together into one level called `Married`. Then in variable `education` collapse the levels `1st-4th` and `Preschool` together into a level called `<=4th`.

```
adult$marital_status = as.character(adult$marital_status)

adult$marital_status = ifelse(adult$marital_status == "Married-AF-spouse" | adult$marital_status == "Ma
```

```
adult$marital_status = as.factor(adult$marital_status)

adult$education = as.character(adult$education)

adult$education = ifelse(adult$education == "1st-4th" | adult$education == "Preschool", "<=4th", adult$

adult$education = as.factor(adult$education)
```

Create a model matrix `Xmm` (for this prediction task on just the raw features) and show that it is *not* full rank (i.e. the result of `ncol` is greater than the result of `Matrix::rankMatrix`).

```
Xmm = model.matrix(income~., adult)
ncol(Xmm)
```

```
## [1] 95
```

```
Matrix::rankMatrix(Xmm)
```

```
## [1] 94
## attr(,"method")
## [1] "tolNorm2"
## attr(,"useGrad")
## [1] FALSE
## attr(,"tol")
## [1] 6.697087e-12
```

Now tabulate and sort the variable `native_country`.

```
tab = sort(table(adult$native_country))
```

Do you see rare levels in this variable? Explain why this may be a problem.

Yes we see many such as Holand, Scotland, Honduras, ... We're not going to be able to tell the difference between these rare levels. If we cut the data into train/test, we likely will not see many of them in the training set, and the predictions will be bad.

Collapse all levels that have less than 50 observations into a new level called `other`. This is a very common data science trick that will make your life much easier. If you can't hope to model rare levels, just give up and do something practical! I would recommend first casting the variable to type "character" and then do the level reduction and then recasting back to type `factor`. Tabulate and sort the variable `native_country` to make sure you did it right.

```
adult$native_country = as.character(adult$native_country)

adult$native_country = ifelse(adult$native_country %in% names(tab[tab<50]), "other", adult$native_counti

adult$native_country = as.factor(adult$native_country)
```

We're still not done getting this data down to full rank. Take a look at the model matrix just for `workclass` and `occupation`. Is it full rank?

```
Xmm = model.matrix(income ~ workclass + occupation, adult)
ncol(Xmm)
```

```
## [1] 21
```

```
Matrix::rankMatrix(Xmm)
```

```
## [1] 20
## attr(,"method")
## [1] "tolNorm2"
## attr(,"useGrad")
## [1] FALSE
## attr(,"tol")
## [1] 6.697087e-12
```

These variables are similar and they probably should be interacted anyway eventually. Let's combine them into one factor. Create a character variable named `worktype` that is the result of concatenating `occupation` and `workclass` together with a ":" in between. Use the `paste` function with the `sep` argument (this casts automatically to type `character`). Then tabulate its levels and sort.

```
adult$worktype = paste(adult$occupation, adult$workclass, sep = ":")
tabulate = sort(table(adult$worktype))

adult = subset(adult, select = -c(occupation, workclass))
```

Like the `native_country` exercise, there are a lot of rare levels. Collapse levels with less than 100 observations to type `other` and then cast this variable `worktype` as type `factor`. Recheck the tabulation to ensure you did this correct.

```
adult$worktype = ifelse(adult$worktype %in% names(tabulate[tabulate<100]), "other", adult$worktype)

adult$worktype = as.factor(adult$worktype)

sort(table(adult$worktype))
```

```
##
##        Transport-moving:Local-gov         Protective-serv:State-gov
##                                115                               116
## Transport-moving:Self-emp-not-inc           Other-service:State-gov
##                                118                               123
##            Craft-repair:Local-gov           Priv-house-serv:Private
##                                143                               143
##       Prof-specialty:Self-emp-inc        Prof-specialty:Federal-gov
##                                157                               167
##    Other-service:Self-emp-not-inc       Exec-managerial:Federal-gov
##                                173                               179
##         Exec-managerial:State-gov           Protective-serv:Private
##                                186                               186
##           Other-service:Local-gov         Exec-managerial:Local-gov
##                                189                               212
##            Adm-clerical:State-gov            Adm-clerical:Local-gov
```

```
##                               250                               281
##               Sales:Self-emp-inc        Protective-serv:Local-gov
##                               281                               304
##         Adm-clerical:Federal-gov  Prof-specialty:Self-emp-not-inc
##                               316                               365
##           Sales:Self-emp-not-inc Exec-managerial:Self-emp-not-inc
##                               376                               383
##     Exec-managerial:Self-emp-inc         Prof-specialty:State-gov
##                               385                               403
## Farming-fishing:Self-emp-not-inc          Farming-fishing:Private
##                               430                               450
##    Craft-repair:Self-emp-not-inc         Prof-specialty:Local-gov
##                               523                               692
##             Tech-support:Private                             other
##                               723                              1008
##          Transport-moving:Private        Handlers-cleaners:Private
##                              1247                              1255
##         Machine-op-inspct:Private           Prof-specialty:Private
##                              1882                              2254
##           Exec-managerial:Private            Other-service:Private
##                              2647                              2665
##              Adm-clerical:Private                     Sales:Private
##                              2793                              2895
##              Craft-repair:Private
##                              3146
```

To do at home: merge the two variables `relationship` and `marital_status` together in a similar way to what we did here.

```
adult$relationship_status = paste(adult$relationship, adult$marital_status, sep = ":")

adult$relationship_status = as.factor(adult$relationship_status)

adult = subset(adult, select = -c(relationship, marital_status))
```

We are finally ready to fit some probability estimation models for `income`! In lecture 16 we spoke about model selection using a cross-validation procedure. Let's build this up step by step. First, split the dataset into `Xtrain`, `ytrain`, `Xtest`, `ytest` using K=5.

```
K = 5
test_prop = 1 / K
train_indices = sample(1 : nrow(adult), round((1 - test_prop) * nrow(adult)))
adult_train = adult[train_indices, ]
y_train = adult_train$income
X_train = adult_train
X_train$income = NULL
test_indices = setdiff(1 : nrow(adult), train_indices)
adult_test = adult[test_indices, ]
y_test = adult_test$income
X_test = adult_test
X_test$income = NULL
```

Create the following four models on the training data in a `list` objected named `prob_est_mods`: logit, probit, cloglog and cauchit (which we didn't do in class but might as well). For the linear component within

the link function, just use the vanilla raw features using the `formula` object `vanilla`. Each model's key in the list is its link function name + "-vanilla". One for loop should do the trick here.

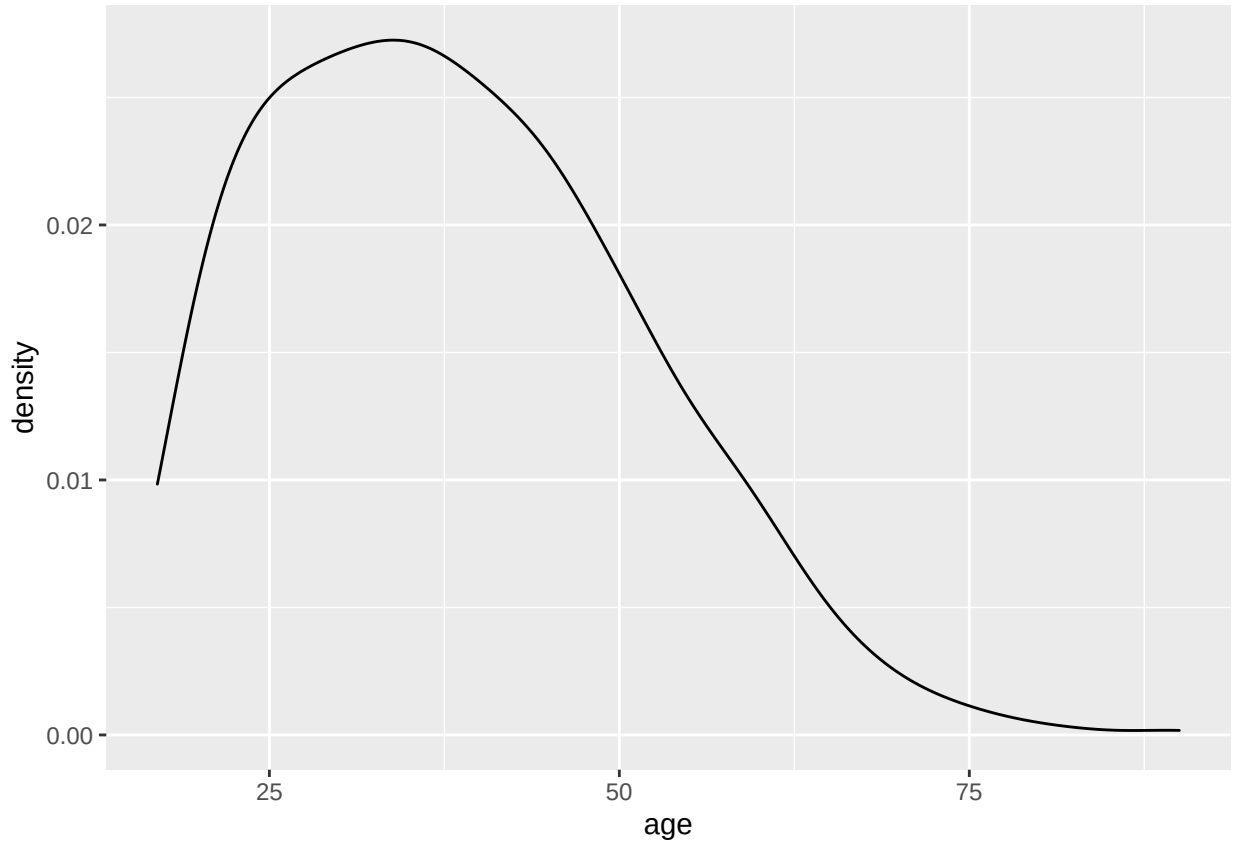
```
link_functions = c("logit", "probit", "cloglog", "cauchit")
vanilla = income ~ .
prob_est_mods = list()

for (link_function in link_functions) {
  prob_est_mods[[paste(link_function, "vanilla", sep = "-")]] = glm(vanilla, adult, family = binomial(li
}
```

Now let's get fancier. Let's do some variable transforms. Add `log_capital_loss` derived from `capital_loss` and `log_capital_gain` derived from `capital_gain`. Since there are zeroes here, use log_x = log(1 + x) instead of log_x = log(x). That's always a neat trick. Just add them directly to the data frame so they'll be picked up with the `.` inside of a formula.

```
adult$log_capital_loss = log(adult$capital_loss + 1)
adult$log_capital_gain = log(adult$capital_gain + 1)
```

Create a density plot that shows the age distribution by `income`.

```
ggplot(adult) +
  aes(x = age) +
  geom_density(y = "income", adjust = 2)
```



What do you see? Is this expected using common sense?

The graph shows that density of income is lower for younger adults and increases/peaks towards middle age, then declines at around retirement age. This makes sense as there is less of a density for older/retired people making > \$50k. It makes sense that the density is higher for middle aged adults where their careers are peaking.

Now let's fit the same models with all link functions on a formula called `age_interactions` that uses interactions for `age` with all of the variables. Add all these models to the `prob_est_mods` list.

```
K = 5
test_prop = 1 / K
train_indices = sample(1 : nrow(adult), round((1 - test_prop) * nrow(adult)))
adult_train = adult[train_indices, ]
y_train = adult_train$income
X_train = adult_train
X_train$income = NULL
test_indices = setdiff(1 : nrow(adult), train_indices)
adult_test = adult[test_indices, ]
y_test = adult_test$income
X_test = adult_test
X_test$income = NULL

age_interactions = income ~ age * .

link_functions = c("logit", "probit", "cloglog", "cauchit")

for (link_function in link_functions) {
  prob_est_mods[[paste(link_function, "age_interactions", sep = "-")]] = glm(age_interactions, adult, fa
}
```

Create a function called `brier_score` that takes in a probability estimation model, a dataframe `X` and its responses `y` and then calculates the brier score.

```
brier_score = function(prob_est_mod, X, y){
  phat = predict(prob_est_mod, X, type = "response")
  mean(-(y - phat)^2)
}
```

Now, calculate the in-sample Brier scores for all models. You can use the function `lapply` to iterate over the list and pass in in the function `brier_score`.

```
lapply(prob_est_mods, brier_score, X_train, y_train)
```

```
## $`logit-vanilla`
## [1] -0.1034692
##
## $`probit-vanilla`
## [1] -0.1035025
##
## $`cloglog-vanilla`
## [1] -0.1043748
##
## $`cauchit-vanilla`
```

```
## [1] -0.1051033
##
## $`logit-age_interactions`
## [1] -0.09999441
##
## $`probit-age_interactions`
## [1] -0.1004585
##
## $`cloglog-age_interactions`
## [1] -0.1007105
##
## $`cauchit-age_interactions`
## [1] -0.1011924
```

Now, calculate the out-of-sample Brier scores for all models. You can use the function `lapply` to iterate over the list and pass in the function `brier_score`.

```
lapply(prob_est_mods, brier_score, X_test, y_test)
```

```
## $`logit-vanilla`
## [1] -0.1015174
##
## $`probit-vanilla`
## [1] -0.1016746
##
## $`cloglog-vanilla`
## [1] -0.102732
##
## $`cauchit-vanilla`
## [1] -0.1028721
##
## $`logit-age_interactions`
## [1] -0.0982463
##
## $`probit-age_interactions`
## [1] -0.09872681
##
## $`cloglog-age_interactions`
## [1] -0.09908798
##
## $`cauchit-age_interactions`
## [1] -0.09877957
```

Which model wins in sample and which wins out of sample? Do you expect these results? Explain.

The interacted logit tends to win both in-sample and out-of-sample (closest to 0). It makes sense because the logistic regression is used for binary outputs, in our case we have 1 if income $>$ \$50k and 0 otherwise.

What is wrong with this model selection procedure? There are a few things wrong.

There is no validation over the training-test sets through a procedure like K-fold Cross Validation. Secondly, we are not doing "nested resampling" on the models to help pick the best one.

Run all the models again. This time do three splits: subtrain, select and test. After selecting the best model, provide a true oos Brier score for the winning model.

```
K = 5
n = nrow(adult)
test_indices = sample(1 : n, size = n * 1 / K)
master_train_indices = setdiff(1 : n, test_indices)
select_indices = sample(master_train_indices, size = n * 1 / K)
subtrain_indices = setdiff(master_train_indices, select_indices)

adult_subtrain = adult[subtrain_indices, ]
adult_select = adult[select_indices, ]
adult_test = adult[test_indices, ]

y_subtrain = adult_subtrain$income
y_select = adult_select$income
y_test = adult_test$income

link_functions = c("logit", "probit", "cloglog", "cauchit")
reran_mods = list()

for (link_function in link_functions) {
  reran_mods[[paste(link_function, "vanilla", sep = "-")]] = glm(vanilla, adult_subtrain, family = bino
}

for (link_function in link_functions) {
  reran_mods[[paste(link_function, "age_interactions", sep = "-")]] = glm(age_interactions, adult_subtra
}

mods = lapply(reran_mods, brier_score, adult_select, y_select)
which.max(mods) # logit vanilla is best
```

```
## logit-vanilla
##             1
```

```
best_model = glm(income ~ ., adult_train, family = binomial(link = "logit"))
true_oos_brier_score = brier_score(best_model, adult_test, y_test)
true_oos_brier_score
```

```
## [1] -0.1018281
```