Who got the flushot?

Grace Haeun Park

2020-03-04

```
getOption("repos")
```

```
    CRAN
"@CRAN@"
```

Preliminaries

```
library(skimr)
library(modelr)
library(rms)
library(purrr)
library(broom)
library(dplyr)
library(tidyverse)


skim_with(numeric = list(hist = NULL),
          integer = list(hist = NULL))
```

```
function (data, ...)
{
    data_name <- rlang::expr_label(substitute(data))
    if (!is.data.frame(data)) {
        data <- as.data.frame(data)
    }
    stopifnot(is.data.frame(data))
    .vars <- rlang::quos(...)
    cols <- names(data)
    if (length(.vars) == 0) {
        selected <- cols
    }
    else {
        selected <- tidyselect::vars_select(cols, !!!.vars)
    }
    grps <- dplyr::groups(data)
    if (length(grps) > 0) {
        group_variables <- selected %in% as.character(grps)
        selected <- selected[!group_variables]
    }
```

```
    skimmers <- purrr::map(selected, get_final_skimmers, data,
        local_skimmers, append)
    types <- purrr::map_chr(skimmers, "skim_type")
    unique_skimmers <- reduce_skimmers(skimmers, types)
    combined_skimmers <- purrr::map(unique_skimmers, join_with_base,
        base)
    ready_to_skim <- tibble::tibble(skim_type = unique(types),
        skimmers = purrr::map(combined_skimmers, mangle_names,
            names(base$funs)), skim_variable = split(selected,
            types)[unique(types)])
    grouped <- dplyr::group_by(ready_to_skim, .data$skim_type)
    nested <- dplyr::summarize(grouped, skimmed = purrr::map2(.data$skimmers,
        .data$skim_variable, skim_by_type, data))
    structure(tidyr::unnest(nested, .data$skimmed), class = c("skim_df",
        "tbl_df", "tbl", "data.frame"), data_rows = nrow(data),
        data_cols = ncol(data), df_name = data_name, groups = dplyr::groups(data),
        base_skimmers = names(base$funs), skimmers_used = get_skimmers_used(unique_skimmers))
}
<bytecode: 0x000000002724b0c0>
<environment: 0x0000000027250f80>
```

Background

There are a fair number of people who refuse to receive flu vaccination every year. The flu vaccine this year (2019) was brought up as a big issue because it did not match the circulating fl so well this flu season. I got the flu shot as I personally thought yearly vaccination was still worth getting. At the same time, I have grown interested in who tend to receive yearly flu vaccinations.

I also want to see if there is any relationship between "always" wearing seat belts while driving a car and getting yearly flu vaccination–if not always wearing seat belts could be consider as an indicator of people being careless and not taking care of themselves as they should be.

Research Questions

1. Who has received flu vaccination over the past 12 months? Can we predict who has received it based on the information such as age, race, sex, and education level?
2. Considering other information mentioned above, whether the subject "always" wears seat belts while driving has a significant impact on whether the subject has received flu vaccination over the past 12 months?

I am going to try building regression models and visualizations to answer these two questions.

My Data

I have obtained the data from the website of the Behavioral Risk Factor Surveillance System (BRFSS) of the Centers for Disease Control and Prevention (CDC). (https://www.cdc.gov/brfss/annual_data/annual_data.htm) My data set is a combination of subsets of the 2016, 2017, and 2018 BRFSS survey data. The BRFSS is a collaborative project throughout the 50 US states, the District of Columbia, Guam, and Puerto

Rico, supported by the CDC. It is a system of telephone surveys for collecting data on all sorts of health-related issues. The survey data were monthly collected through land-line and cell phone calls. This data set is closely related to my research question as it includes all the factors that I want in this study.

However, survey study is a cross-sectional study which makes it difficult to discuss causation; thus, this study would have to be a preliminary or exploratory analysis. Plus, the cohort effect is often is an issue to cross-sectional studies that may conceal the true associations.

Data Load

```
library(haven)
cdc16 <- read_xpt("LLCP2016.xpt") %>% tbl_df
cdc17 <- read_xpt("LLCP2017.xpt") %>% tbl_df
cdc18 <- read_xpt("LLCP2018.xpt") %>% tbl_df
```

As originally loaded, the `cdc17` data contain `486303` rows and `275` columns; `cdc17` contains `450016` rows and `358` columns; and `cdc18` contains `437436` rows and `275` columns.

Tidying, Data Cleaning and Data Management

Change original variable names to be simpler.

```
cdc16$age <- cdc16$"_AGEG5YR"
cdc16$flu <- cdc16$"FLUSHOT6"
cdc16$edu <- cdc16$"_EDUCAG"
cdc16$sb <- cdc16$"SEATBELT"
cdc16$raceinfo <- cdc16$"_RACEG21"
cdc16$sexinfo <- cdc16$"SEX"
cdc16$interviewmth <- cdc16$"IMONTH"

cdc17$age <- cdc17$"_AGEG5YR"
cdc17$flu <- cdc17$"FLUSHOT6"
cdc17$edu <- cdc17$"_EDUCAG"
cdc17$sb <- cdc17$"SEATBELT"
cdc17$raceinfo <- cdc17$"_RACEG21"
cdc17$sexinfo <- cdc17$"SEX"
cdc17$interviewmth <- cdc17$"IMONTH"

cdc18$age <- cdc18$"_AGEG5YR"
cdc18$flu <- cdc18$"FLUSHOT6"
cdc18$edu <- cdc18$"_EDUCAG"
cdc18$sb <- cdc18$"SEATBELT"
cdc18$raceinfo <- cdc18$"_RACEG21"
cdc18$sexinfo <- cdc18$"SEX1"
cdc18$interviewmth <- cdc18$"IMONTH"
```

Combine subsets of the three datasets to merge as one data set.

```
cdc16a <-cdc16 %>%
  select(age, flu, edu, sb, sexinfo, raceinfo, interviewmth)
```

```
cdc17a <-cdc17 %>%
  select(age, flu, edu, sb, sexinfo, raceinfo, interviewmth)

cdc18a <-cdc18 %>%
  select(age, flu, edu, sb, sexinfo, raceinfo, interviewmth)

bind1 <- rbind(cdc16a, cdc17a)
cdc.data <- rbind(bind1, cdc18a)
```

I am going to remove categories of "don't know", "don't care", or "refused" categories here to reduce the data set. I am also sizing down the data set by only having the data set obtained in July out of the twelve months from January to December every year. A table for interview months is shown below.

```
# removing "don't know", "don't care", and "refused" categories
cdc <- cdc.data%>%
  filter(flu!="7") %>%
  filter(flu!="9") %>%
  filter(edu!="9") %>%
  filter(sb !="7") %>%
  filter(sb !="8") %>%
  filter(sb !="9") %>%
  filter(interviewmth=="07") %>%
  filter(sexinfo!="9") %>%
  filter(sexinfo!="7") %>%
  filter(raceinfo!="9") %>%
  filter(age!="14")
```

```
table(cdc.data$interviewmth)
```

```
    01     02     03     04     05     06     07     08     09     10     11
 91313 113125 114101 104573 113350 121847 122692 125907 115799 114823 118831
    12
117394
```

```
skim(cdc)
```

Table 1: Data summary

Name	cdc
Number of rows	110405
Number of columns	7
Column type frequency:	
character	1
numeric	6
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
interviewmth	0	1	2	2	0	1	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	7.72	3.48	1	5	8	10	13	
flu	0	1	1.56	0.50	1	1	2	2	2	
edu	0	1	2.96	0.97	1	2	3	4	4	
sb	0	1	1.22	0.69	1	1	1	1	5	
sexinfo	0	1	1.56	0.50	1	1	2	2	2	
raceinfo	0	1	1.23	0.42	1	1	1	1	2	

I am converting my four binary outcomes to be in categories of 0 and 1 now.

```
cdc <- cdc %>%
  mutate(flushot = ifelse(flu==1,1,0)) %>%
  mutate(seatbelt = ifelse(sb==1,1,0)) %>%
  mutate(sex = ifelse(sexinfo==1,1,0)) %>%
  mutate(race = ifelse(raceinfo==1,1,0)) %>%
  select(-flu, -sb, -sexinfo, -raceinfo, -interviewmth)
```

I am creating variables that have real names for factor levels for building understandable visualizations later.

```
cdc <- cdc %>%
  mutate(flufact = as.factor(case_when(flushot == 1 ~ "Vaccinated",
                                       flushot == 0 ~ "Not Vaccinated"))) %>%
  mutate(sbfact = as.factor(case_when(seatbelt == 1 ~ "Yes",
                                      seatbelt == 0 ~ "No"))) %>%
  mutate(racefact = as.factor(case_when(race == 1 ~ "Non-Hispanic White",
                                      race == 0 ~ "Non-White or Hispanic"
                                      ))) %>%
  mutate(sexfact = as.factor(case_when(sex == 1 ~ "Male",
                                      sex == 0 ~ "Female"))) %>%
  mutate(edufact = as.factor(case_when(edu == 1 ~ "No highschool",
                                      edu == 2 ~ "Highshool graduate",
                                      edu == 3 ~ "Attended college/technical school",
                                      edu == 4 ~ "College/technical school graduate"))) %>%
  mutate(agefact = as.factor(case_when(age == 1~"18-24",
                                       age == 2~"25-29",
                                       age == 3~"30-34",
                                       age == 4~"35-39",
                                       age == 5~"40-44",
                                       age == 6~"45-49",
                                       age == 7~"50-54",
                                       age == 8~"55-59",
                                       age == 9~"60-64",
                                       age == 10~"65-69",
                                       age == 11~"70-74",
                                       age == 12~"75-79",
```

```
                                        age == 13~"80-99"))) %>%
  mutate(racefact = fct_relevel(racefact, "Non-Hispanic White", "Non-White or Hispanic")) %>%
  mutate(flufact = fct_relevel(flufact, "Vaccinated", "Not Vaccinated")) %>%
  mutate(edufact = fct_relevel(edufact, "No highschool", "Highshool graduate", "Attended college/techni
  mutate(sbfact = fct_relevel(sbfact, "Yes", "No"))
```

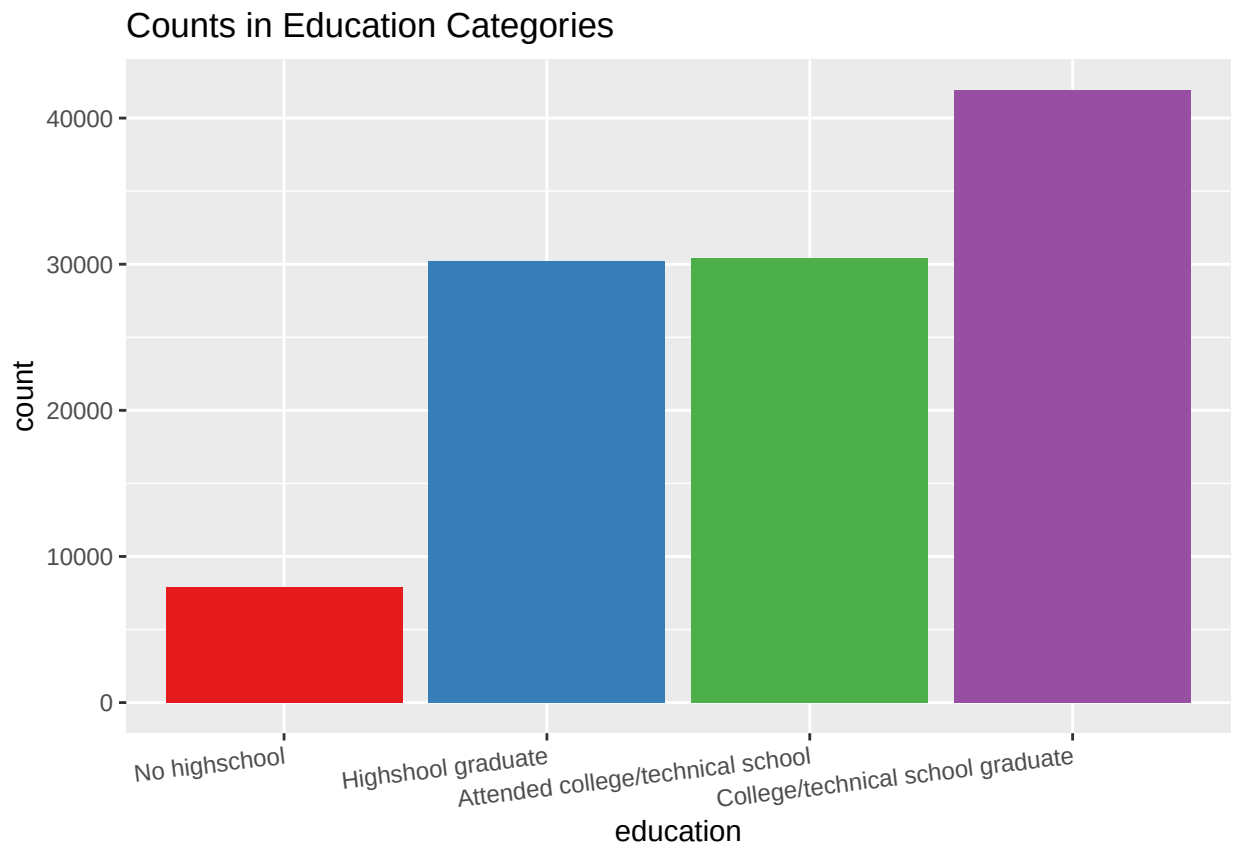
Code Book

Variable	Type	Details
edu	multi-categorical	Completed education level: No high school, Highschool graduate, Attended college or technical school, Graduated college or technical school
race	multi-categorical	Race category: Non-Hispanic white, Other
age	multi-categorical	Fourteen-level age category: 18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-99
flushot	binary	Had flushot or vaccination spray for the past 12 months: Yes (44%) or No (56%)
seatbelt	binary	Whether the subject always wears seatbelt when driving a car: Always (88%) or Not always (12%)
sex	binary	Male (\44%) or Female (\56%)

Variables

Let's take a look at the distribution of categoreis of our variables.d

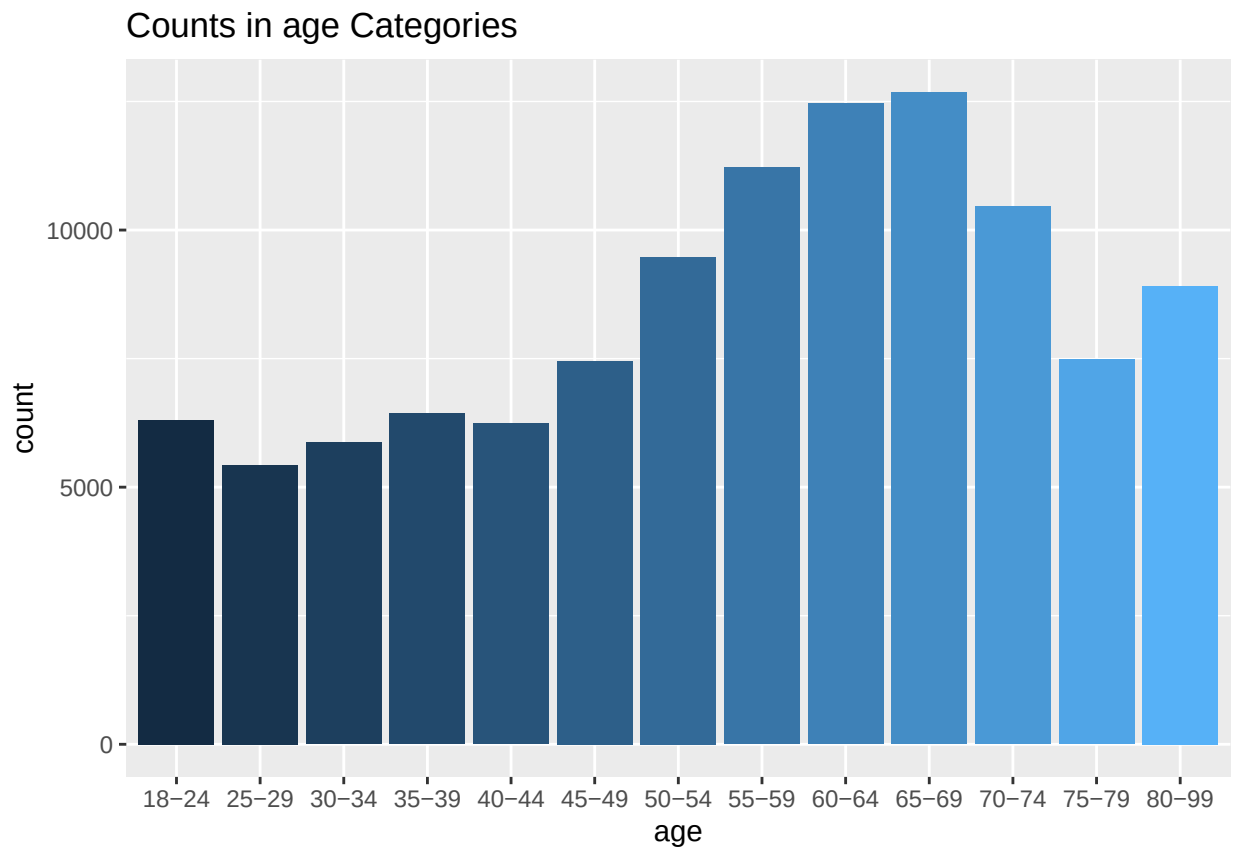
```
ggplot(cdc, aes(x = factor(edufact), fill = edufact)) +
geom_bar()+
  scale_fill_brewer(palette = "Set1") +
  guides(fill=FALSE) +
  theme(axis.text.x = element_text(angle=7, hjust=1)) +
labs(x="education",
     title = "Counts in Education Categories")
```




```
prop.table(table(cdc$edufact))
```

```
                    No highschool                Highshool graduate
                       0.07162719                        0.27342059
Attended college/technical school College/technical school graduate
                       0.27544042                        0.37951180
```

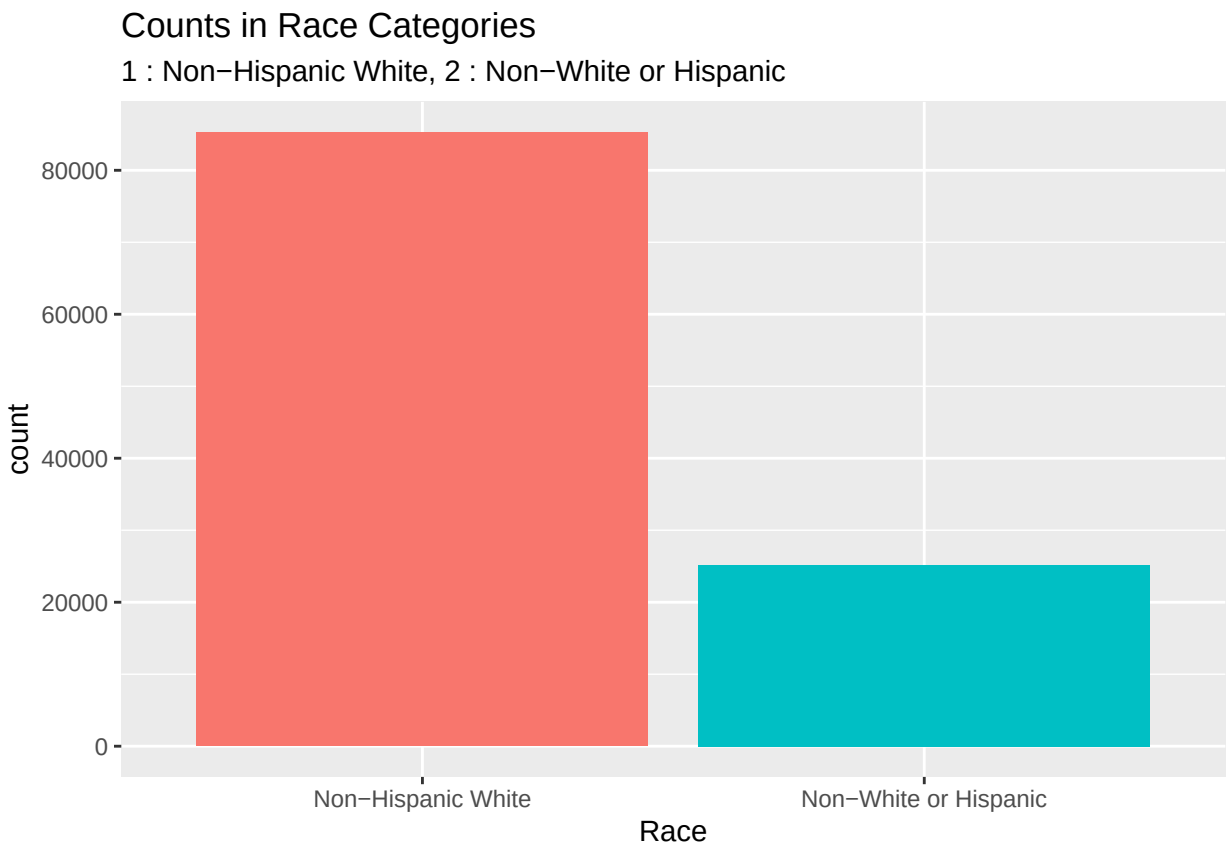
```
ggplot(cdc, aes(x = factor(agefact), fill = age)) +
geom_bar()+
  guides(fill=FALSE) +
labs(x="age",
     title = "Counts in age Categories")
```




```
prop.table(table(cdc$agefact))
```

```
     18-24      25-29      30-34      35-39      40-44      45-49      50-54
0.05708981 0.04918256 0.05312259 0.05823106 0.05651918 0.06748789 0.08582039
     55-59      60-64      65-69      70-74      75-79      80-99
0.10160772 0.11284815 0.11483176 0.09474209 0.06776867 0.08074815
```

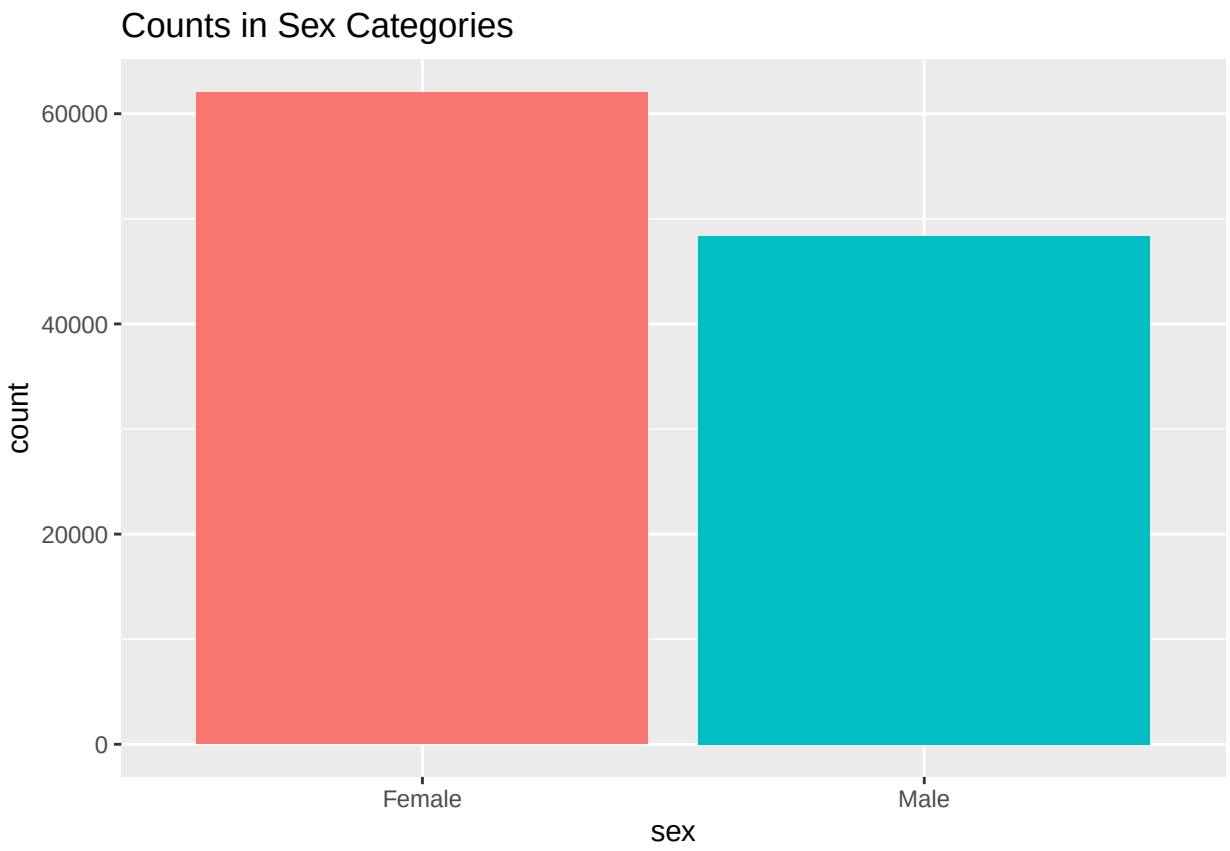
```
ggplot(cdc, aes(x = factor(racefact), fill = racefact)) +
geom_bar()+
  guides(fill=FALSE) +
labs(x="Race",
     title = "Counts in Race Categories",
     subtitle= "1 : Non-Hispanic White, 2 : Non-White or Hispanic")
```




```
prop.table(table(cdc$racefact))
```

```
   Non-Hispanic White Non-White or Hispanic
            0.7720755             0.2279245
```

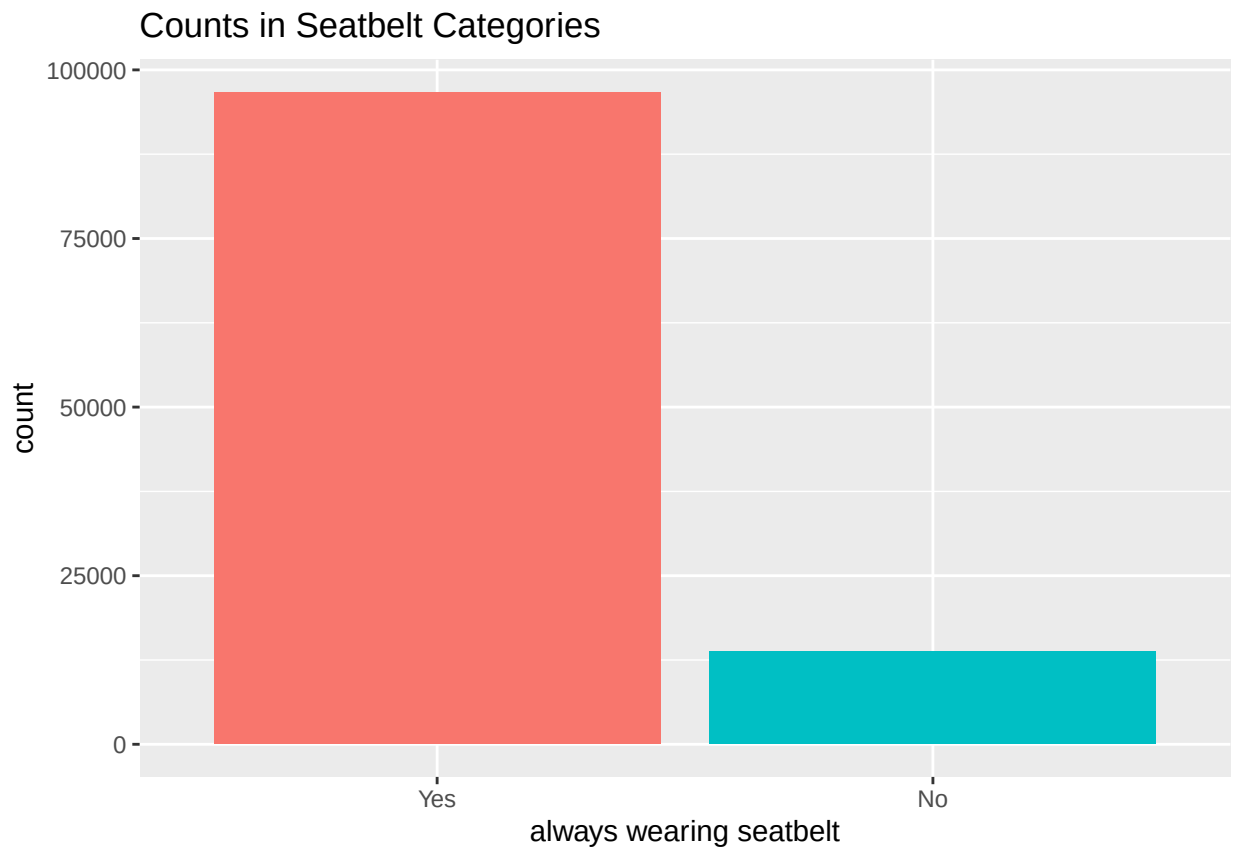
```
ggplot(cdc, aes(x = factor(sexfact), fill = sexfact)) +
geom_bar()+
  guides(fill=FALSE) +
labs(x="sex",
     title = "Counts in Sex Categories")
```




```
prop.table(table(cdc$sexfact))
```

```
   Female      Male
0.5618858 0.4381142
```

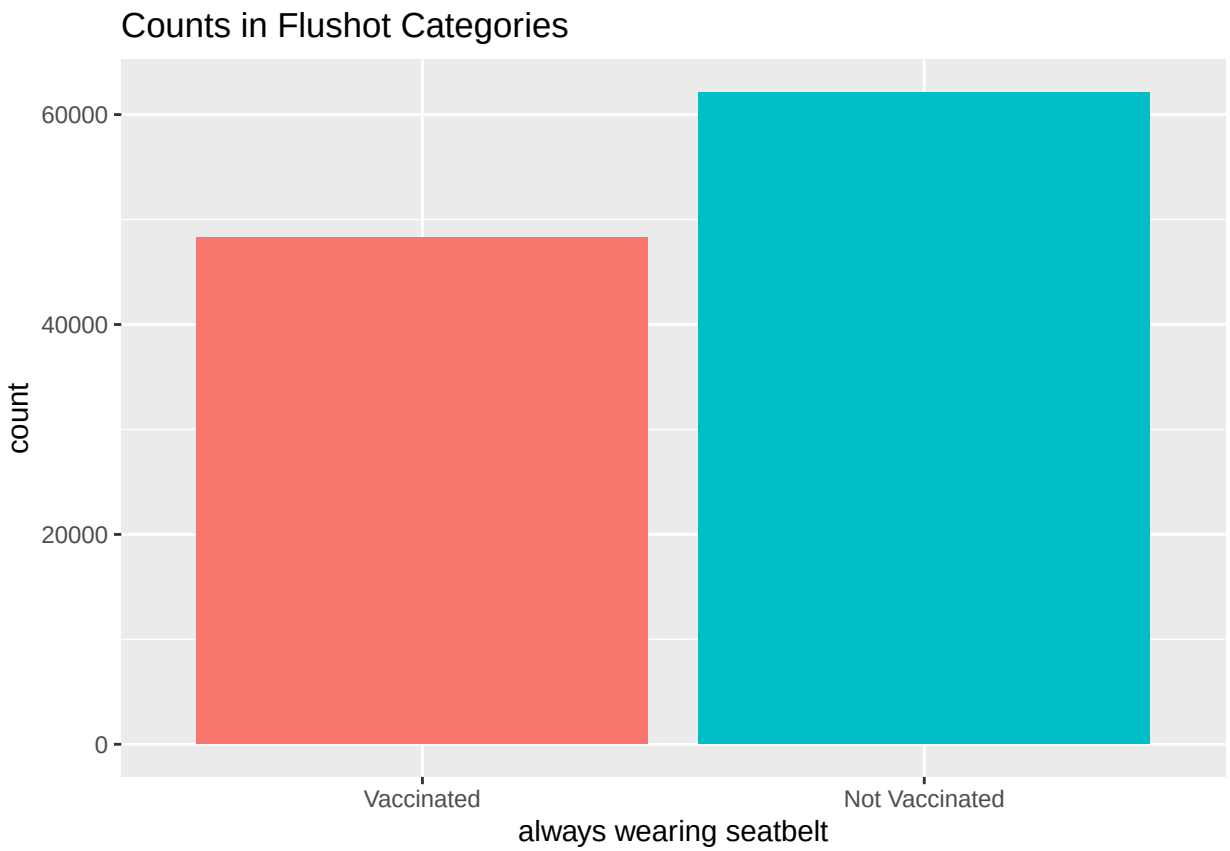
```
ggplot(cdc, aes(x = factor(sbfact), fill = sbfact)) +
geom_bar()+
  guides(fill=FALSE) +
labs(x="always wearing seatbelt",
     title = "Counts in Seatbelt Categories")
```




```
prop.table(table(cdc$sbfact))
```

```
      Yes        No
0.8756125 0.1243875
```

```
ggplot(cdc, aes(x = factor(flufact), fill = flufact)) +
geom_bar()+
  guides(fill=FALSE) +
labs(x="always wearing seatbelt",
     title = "Counts in Flushot Categories")
```




```
prop.table(table(cdc$flufact))
```

```
    Vaccinated Not Vaccinated
     0.4373443      0.5626557
```

Tidied Tibble

Our tibble `cdc` contains 111,629 rows (respondents) and 6 columns (variables). Each variable is contained in a column, and each row represents a single subject. All variables now have appropriate types.

Missingness

```
skim(cdc)
```

Table 5: Data summary

Name	cdc
Number of rows	110405
Number of columns	12

Column type frequency:

Table 5: Data summary

factor	6
numeric	6
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
flufact	0	1	FALSE	2	Not: 62120, Vac: 48285
sbfact	0	1	FALSE	2	Yes: 96672, No: 13733
racefact	0	1	FALSE	2	Non: 85241, Non: 25164
sexfact	0	1	FALSE	2	Fem: 62035, Mal: 48370
edufact	0	1	FALSE	4	Col: 41900, Att: 30410, Hig: 30187, No : 7908
agefact	0	1	FALSE	13	65-: 12678, 60-: 12459, 55-: 11218, 70-: 10460

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	7.72	3.48	1	5	8	10	13	
edu	0	1	2.96	0.97	1	2	3	4	4	
flushot	0	1	0.44	0.50	0	0	0	1	1	
seatbelt	0	1	0.88	0.33	0	1	1	1	1	
sex	0	1	0.44	0.50	0	0	0	1	1	
race	0	1	0.77	0.42	0	1	1	1	1	

We've got no missing data for any of our variables.

Analyses

Let us begin by building a model with all the predictors in, called `model0` from now on.

```
m0_glm <- glm(flushot ~ age + race + sex + edu + seatbelt, data=cdc, family="binomial")
summary(m0_glm)
```

```
Call:
glm(formula = flushot ~ age + race + sex + edu + seatbelt, family = "binomial",
    data = cdc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5444  -1.0618  -0.7241   1.1429   2.1234

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.313617   0.032672  -70.81   <2e-16 ***
```

```
age          0.140645   0.001939   72.54   <2e-16 ***
race         0.161094   0.015736   10.24   <2e-16 ***
sex         -0.173417   0.012813  -13.53   <2e-16 ***
edu          0.202931   0.006651   30.51   <2e-16 ***
seatbelt     0.343395   0.020035   17.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 151316  on 110404  degrees of freedom
Residual deviance: 143175  on 110399  degrees of freedom
AIC: 143187

Number of Fisher Scoring iterations: 4
```

Let us check the collinearity of predictors.

```
vif(m0_glm)
```

```
     age     race      sex      edu seatbelt
1.046931 1.052476 1.014051 1.030155 1.017636
```

None of the predictors have any problem of collinearity.

I am going to build a `lrm` version of the model0 as well for its benefits that on my analyses.

```
m0_lrm <- lrm(flushot ~ age + race + sex + edu + seatbelt, data=cdc, x=T, y=T)
m0_lrm
```

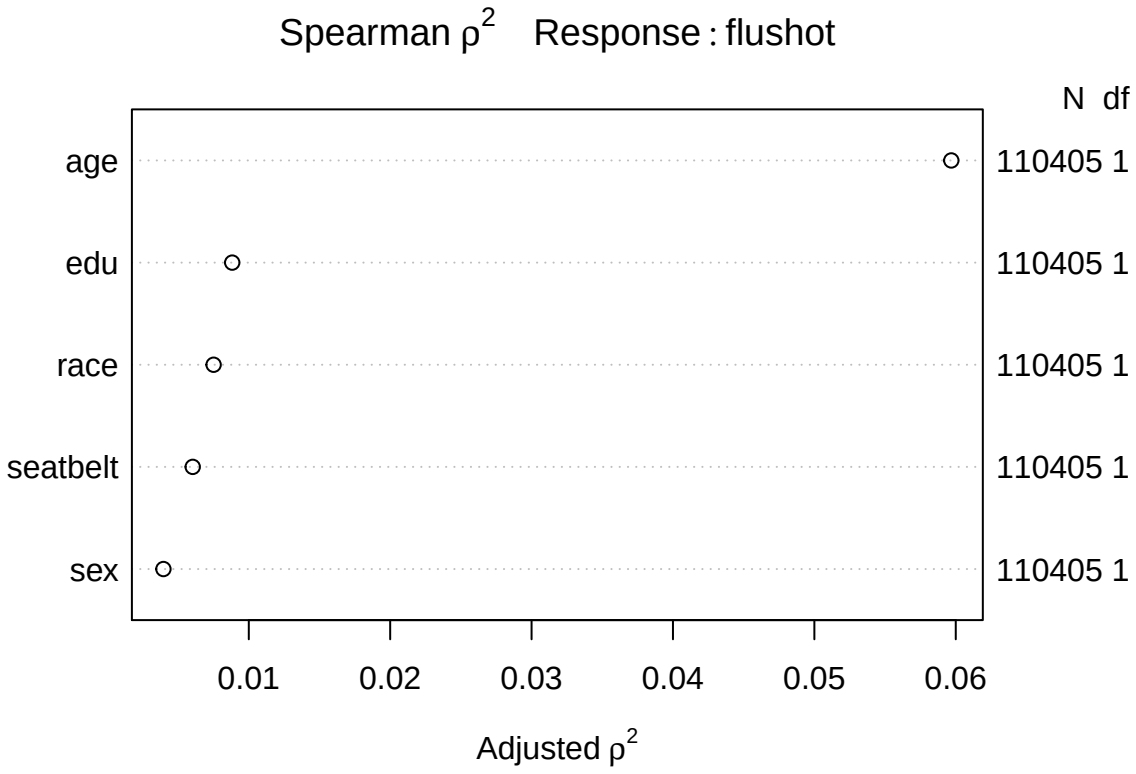
```
Logistic Regression Model

 lrm(formula = flushot ~ age + race + sex + edu + seatbelt, data = cdc,
     x = T, y = T)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs          110405    LR chi2    8140.76    R2       0.095    C       0.657
  0            62120    d.f.             5    g        0.659    Dxy     0.315
  1            48285    Pr(> chi2) <0.0001    gr       1.932    gamma   0.317
 max |deriv|  6e-09                          gp       0.152    tau-a   0.155
                                             Brier    0.228

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -2.3136 0.0327 -70.81 <0.0001
 age        0.1406 0.0019  72.54 <0.0001
 race       0.1611 0.0157  10.24 <0.0001
 sex       -0.1734 0.0128 -13.53 <0.0001
 edu        0.2029 0.0067  30.51 <0.0001
 seatbelt   0.3434 0.0200  17.14 <0.0001
```

The R-squared for this model is 0.095 which means this model can explain only about 10% of observed variation; we cannot call this a good model, but we will see what we can figure out using these variables.

Adding an interaction term

Let's take a look at the Spearman's plot to build a model with an interaction term and compare with the original model.

```
plot(spearman2(flushot ~ age + race + sex + edu + seatbelt, data = cdc))
```



The Spearman rho-squared plot suggests `age` and `edu` as the candidate predictors for non-linear terms. I am going to add an interaction term of those two categorical predictors and call this new model `model1`.

```
m1_glm <-glm(flushot ~age*edu + sex + edu + seatbelt, data=cdc, family="binomial")
m1_glm
```

```
Call:  glm(formula = flushot ~ age * edu + sex + edu + seatbelt, family = "binomial",
    data = cdc)

Coefficients:
(Intercept)          age          edu          sex     seatbelt      age:edu
   -2.65108      0.19389      0.34896     -0.16388      0.33383     -0.01669

Degrees of Freedom: 110404 Total (i.e. Null);  110399 Residual
Null Deviance:      151300
Residual Deviance: 143200   AIC: 143200
```

- R-squared for this model with an interaction term: 1 - (Residual Deviance/Null Deviance) = 1-(111623/153000) = 0.027. Let us compare the AIC and BIC of our two models as well.
- Let us compare AIC and BIC of our `model0` and `model1`.

```
glance(m0_glm);
```

```
# A tibble: 1 x 7
  null.deviance df.null  logLik     AIC     BIC deviance df.residual
          <dbl>   <int>   <dbl>   <dbl>   <dbl>    <dbl>       <int>
1       151316.  110404 -71587. 143187. 143244.  143175.      110399
```

```
glance(m1_glm)
```

```
# A tibble: 1 x 7
  null.deviance df.null  logLik     AIC     BIC deviance df.residual
          <dbl>   <int>   <dbl>   <dbl>   <dbl>    <dbl>       <int>
1       151316.  110404 -71605. 143222. 143280.  143210.      110399
```

Model	AIC	BIC
main effects model	145176	145233.7
interaction model	145212.1	145269.9

Based on no improvement of AIC or BIC from building the new model, I prefer the original model. I am not adding the interaction term.

Smaller model

Now, I intend to remove a few predictors from the original model and build a bit more parsimonious model. Let us see the significance of predictors using an ANOVA model.

```
anova(m0_lrm)
```

```
                Wald Statistics          Response: flushot

 Factor     Chi-Square d.f. P
 age        5262.31    1    <.0001
 race        104.80    1    <.0001
 sex         183.17    1    <.0001
 edu         930.95    1    <.0001
 seatbelt    293.78    1    <.0001
 TOTAL      7346.25    5    <.0001
```

Strangely or luckily, all the predictors appear to be statistically significant.

What about stepwise regression?

```
library(MASS)
step(m0_glm)
```

```
Start:  AIC=143186.8
flushot ~ age + race + sex + edu + seatbelt

           Df Deviance    AIC
<none>          143175 143187
- race      1   143280 143290
- sex       1   143358 143368
- seatbelt  1   143475 143485
- edu       1   144117 144127
- age       1   148788 148798


Call:  glm(formula = flushot ~ age + race + sex + edu + seatbelt, family = "binomial",
    data = cdc)

Coefficients:
(Intercept)          age         race          sex          edu     seatbelt
    -2.3136       0.1406       0.1611      -0.1734       0.2029       0.3434

Degrees of Freedom: 110404 Total (i.e. Null);  110399 Residual
Null Deviance:      151300
Residual Deviance: 143200   AIC: 143200
```

Stepwise regression suggests not to remove any of the already existing variables.

Interpretation of our model

```r
summary(m0_glm)
```

```
Call:
glm(formula = flushot ~ age + race + sex + edu + seatbelt, family = "binomial",
    data = cdc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5444  -1.0618  -0.7241   1.1429   2.1234

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.313617   0.032672  -70.81   <2e-16 ***
age          0.140645   0.001939   72.54   <2e-16 ***
race         0.161094   0.015736   10.24   <2e-16 ***
sex         -0.173417   0.012813  -13.53   <2e-16 ***
edu          0.202931   0.006651   30.51   <2e-16 ***
seatbelt     0.343395   0.020035   17.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 151316  on 110404  degrees of freedom
Residual deviance: 143175  on 110399  degrees of freedom
AIC: 143187

Number of Fisher Scoring iterations: 4
```

```r
exp(coef(m0_glm))
```

```
(Intercept)         age        race         sex         edu    seatbelt
 0.09890286  1.15101553  1.17479490  0.84078669  1.22498828  1.40972571
```

```r
exp(confint(m0_glm))
```

```
                2.5 %    97.5 %
(Intercept) 0.09275899 0.1054333
age         1.14665475 1.1554026
race        1.13912943 1.2116088
sex         0.81993235 0.8621675
edu         1.20913053 1.2410690
seatbelt    1.35551160 1.4662599
```

```r
table(cdc$race)
```

```
    0     1
25164 85241
```

```r
table(cdc$sex)
```

```
    0     1
62035 48370
```

Despite its low R-squared, I will interpret the model:

- Moving the age group to the next higher one, or adding 5 years, increases the odds of having received flu vaccination over the past year by the estimated factor of 1.15 (95$ CI: 1.14, 1.16).
- Moving the race group from `non-white or Hispanic` to `non-Hispanic white` increases the odds of having received flu vaccination over the past year to 1.17 (95$ CI: 1.14, 1.21) of the odds ratio.
- The odds of having received flu vaccination over the past year for males is 0.84 times (95% CI: 0.82, 0.86) of the odds for females.
- Moving the education group to the next higher level increases the odds of having received flu vaccination over the past year by the estimated factor of 1.22 times (95% CI: 1.21, 1.24).
- The odds of having received flu vaccination over the past year for those who always wear seat belts while driving is 1.41 times (95% CI: 1.36, 1.47) of the odds for those who do not always wear seat belts while driving.

Let us have one instance of prediction of if two individuals have received flu vaccination over the past 12 months based on our model.

An example of prediction

```
dat <- data.frame(race = c("1", "0"),
                  age=c("5","9"),
                  sex=c("1","0"),
                  seatbelt=c("0","1"),
                  edu=c("1","4"))
pred <- predict(m0_lrm, dat, type = "fitted")
pred
```

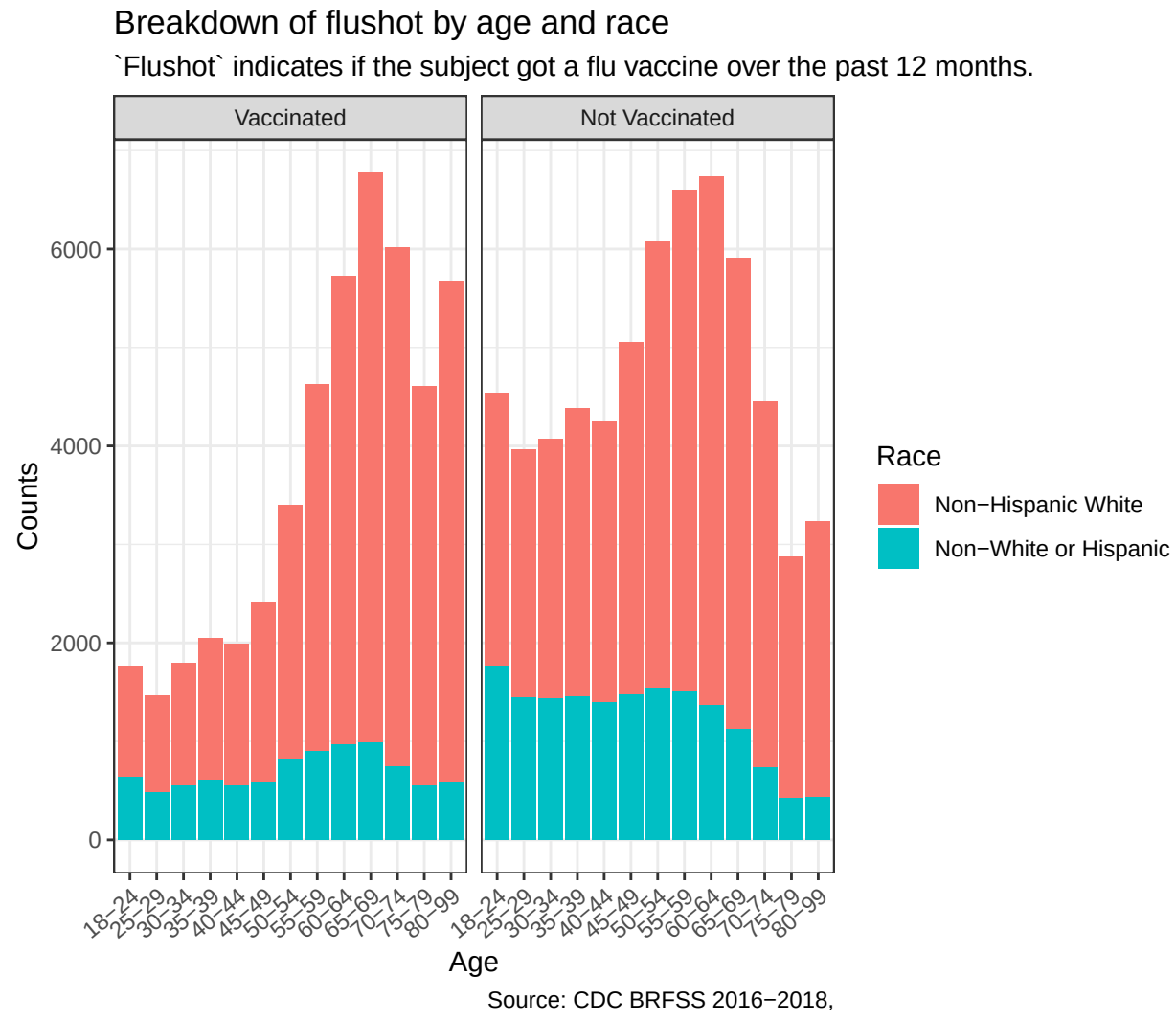
```
         1          2
0.08899722 0.16429033
```

- The first male high school graduate whose race is non-Hispanic white, age is between 40 and 44, who does not always wears seatbelts is predicted to have not received a flu vaccination over the past 12 months.
- The second female college/ technical college graduate whose race is non-White or Hispanic, age is between 60 and 64, who always wears seatbelts while driving is predicted to have received a flu vaccination over the past 12 months.

Visualizations

Let's build some meaningful visualizations.

```
ggplot(na.omit(cdc), aes(x = agefact, fill = racefact)) +
geom_bar() +
theme_bw() +
guides(fill=guide_legend(title="Race")) +
facet_grid(~flufact) +
theme(axis.text.x = element_text(angle=40, hjust=1)) +
labs(x="Age",
     y="Counts",
     title="Breakdown of flushot by age and race",
     subtitle="`Flushot` indicates if the subject got a flu vaccine over the past 12 months.",
     caption="Source: CDC BRFSS 2016-2018,

     More population from Non-Hispanic Black and `Other` races are in the not-vaccinated category than i
     The average age is slightly higher in the vaccinated category.")
```



- This plot shows the effects of `age` and `race` on `flushot`.
- There is the higher odds of the `non-white or Hispanic` group in the `not vaccinated` group than in `vaccinated`. There is a larger Non-Hispanic White population on the `vaccinated` side than on the `not vaccinated`.
- The peak is around the age 65-69 for the vaccinated group for both racial groups. But on the side of `not vaccinated`, the younger, the more counts are for the `non-white or Hispanic` group with its peak in the youngest age group, 18-24, and `non-Hispanic white` group has its peak around the age of 60-64. For the `non-Hispanic white` group, the peak on the `vaccinated` side is found at the older age group than its peak on the `not vaccinated` side.
- Let's simplify the bar graphs a bit by taking the 3-group age variable instead.

As my predictor `age` has 13 levels which could be collapsed into fewer levels of factors in order to improve the overall effectivenss of our visual aids. I am trying collapsing the age factors into three now: low, middle and high. The new age variale is called `age3`. `Low` is from 18 years old to 44, `middle` is 45 to 64, and `high` is from 65 to 99.

```
cdc <- cdc %>%
mutate(agegr3 = as.factor(case_when(age==13~3,
                                    age==12~3,
                                    age==11~3,
                                    age==10~3,
                                    age==9~2,
                                    age==8~2,
                                    age==7~2,
                                    age==6~2,
                                    age==5~1,
                                    age==4~1,
                                    age==3~1,
                                    age==2~1,
                                    age==1~1)))

cdc <- cdc %>%
  mutate(age3 = as.factor(case_when(agegr3 == 1 ~ "18-44",
                                        agegr3 == 2 ~ "45-64",
                                        agegr3 == 3 ~ "65-99")))
```

The follwoing are the counts for the old and new age categories.

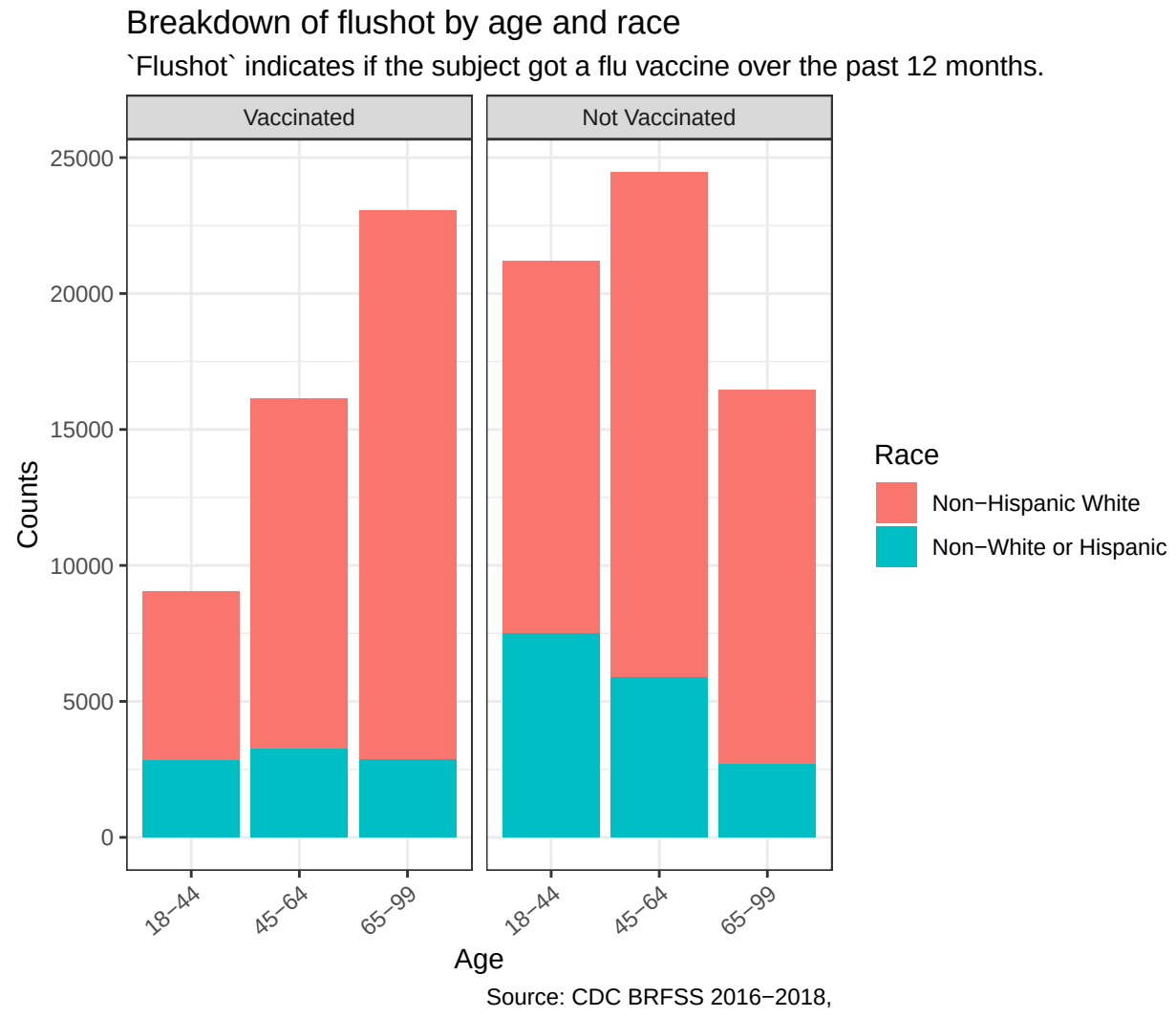
```
table(cdc$age)
```

```
   1    2    3    4    5    6    7    8    9   10   11   12   13
6303 5430 5865 6429 6240 7451 9475 11218 12459 12678 10460 7482 8915
```

```
table(cdc$age3)
```

```
18-44 45-64 65-99
30267 40603 39535
```

```
ggplot(na.omit(cdc), aes(x = age3, fill = racefact)) +
geom_bar() +
theme_bw() +
guides(fill=guide_legend(title="Race")) +
facet_grid(~flufact) +
theme(axis.text.x = element_text(angle=40, hjust=1)) +
labs(x="Age",
     y="Counts",
     title="Breakdown of flushot by age and race",
     subtitle="`Flushot` indicates if the subject got a flu vaccine over the past 12 months.",
     caption="Source: CDC BRFSS 2016-2018,

     More population from Non-Hispanic Black and `Other` races are in the not-vaccinated category than i
     The average age is slightly higher in the vaccinated category.")
```

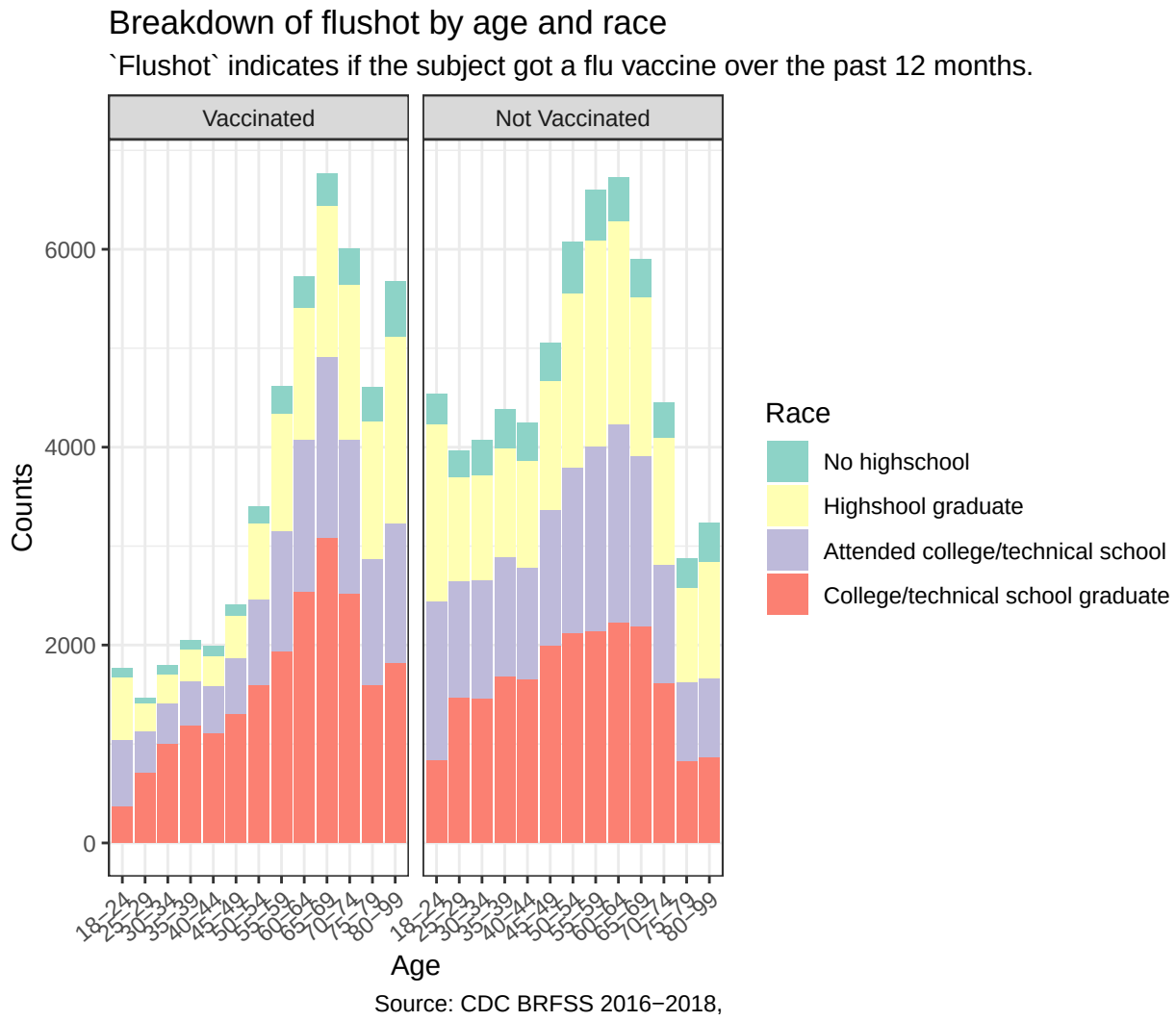


Black and `Other` races are in the not−vaccinated category than in the vaccinated.
The average age is slightly higher in the vaccinated category.

- As explained previously, it is more obvious that the younger, the more counts of "have not received vaccination over the past 12 months" for the "Non-white or Hispanic" racial group. It does not necessarily make those `Non-white or Hispanic` who have received vaccination over the past 12 months have a reverse trend, though.

```
ggplot(na.omit(cdc), aes(x = agefact, fill = edufact)) +
geom_bar() +
theme_bw() +
guides(fill=guide_legend(title="Race")) +
scale_fill_brewer(palette = "Set3") +
facet_grid(~flufact) +
theme(axis.text.x = element_text(angle=40, hjust=1)) +
labs(x="Age",
     y="Counts",
     title="Breakdown of flushot by age and race",
     subtitle="`Flushot` indicates if the subject got a flu vaccine over the past 12 months.",
```

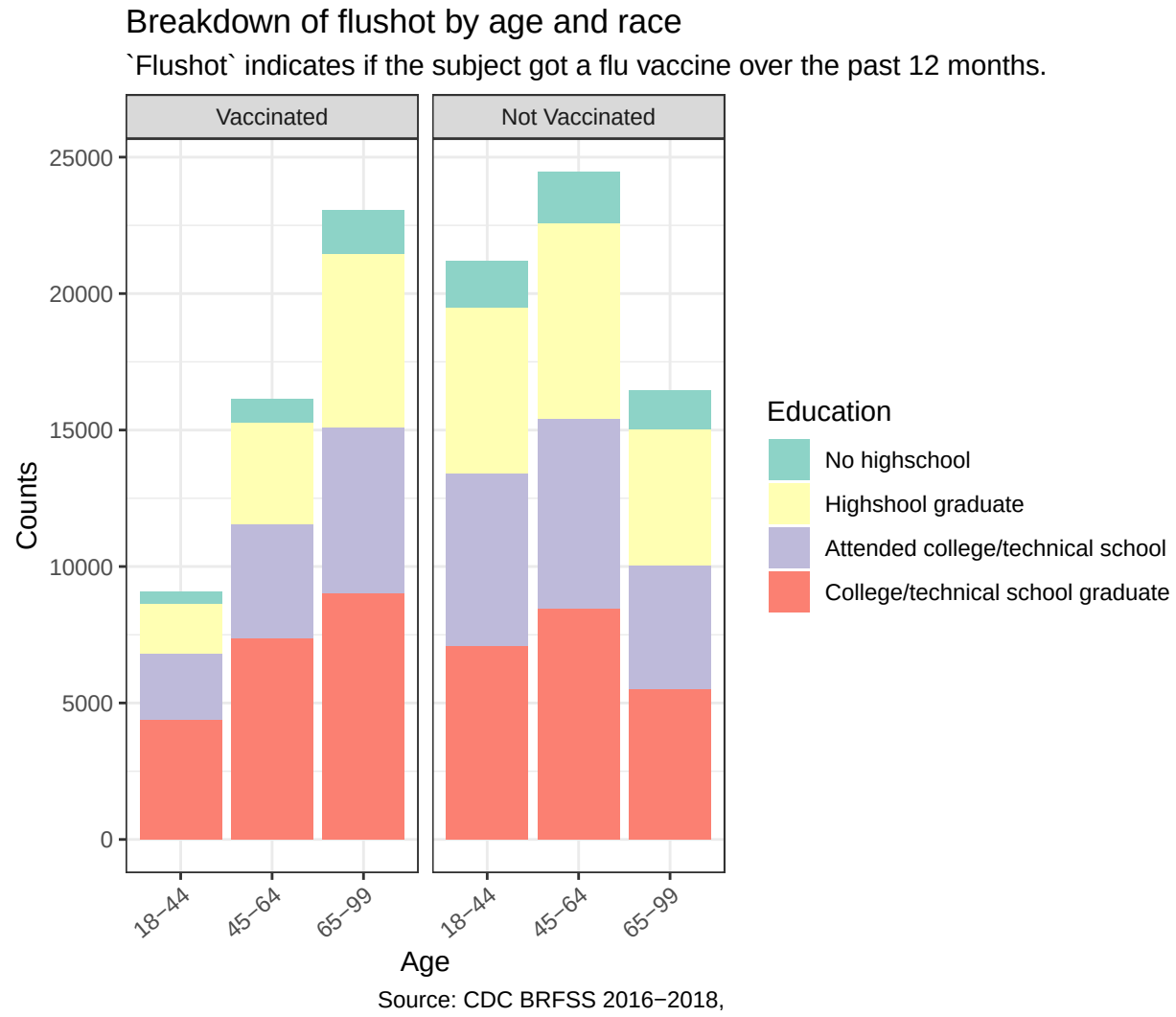
```
     caption="Source: CDC BRFSS 2016-2018,

     More population from Non-Hispanic Black and `Other` races are in the not-vaccinated category than i
     The average age is slightly higher in the vaccinated category.")
```



- Let's simplify the graphs using the 3-group age variable, again.

```
ggplot(na.omit(cdc), aes(x = age3, fill = edufact)) +
geom_bar() +
theme_bw() +
guides(fill=guide_legend(title="Education")) +
scale_fill_brewer(palette = "Set3") +
facet_grid(~flufact) +
theme(axis.text.x = element_text(angle=40, hjust=1)) +
labs(x="Age",
     y="Counts",
```

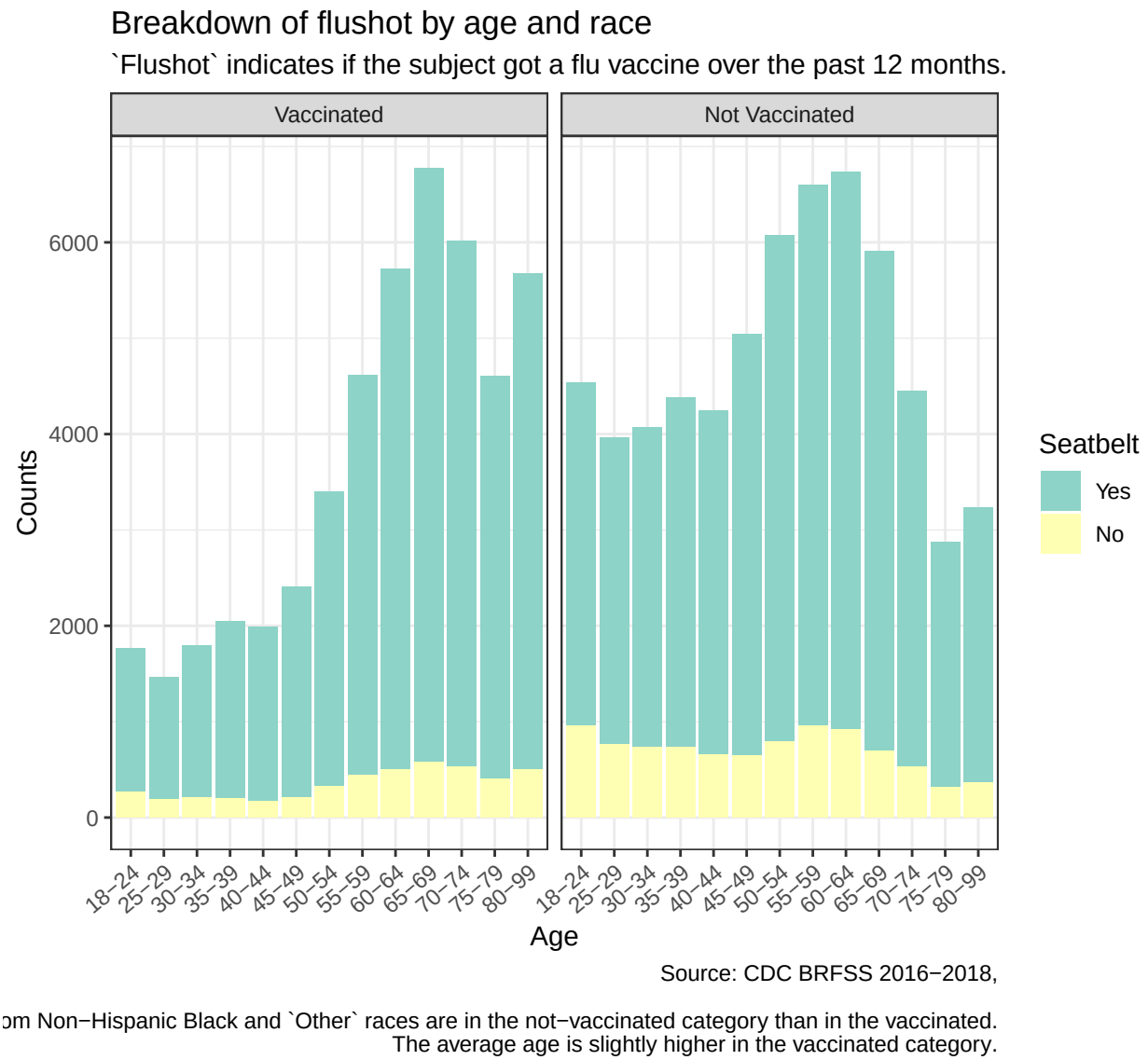
```
    title="Breakdown of flushot by age and race",
    subtitle="`Flushot` indicates if the subject got a flu vaccine over the past 12 months.",
    caption="Source: CDC BRFSS 2016-2018,

    More population from Non-Hispanic Black and `Other` races are in the not-vaccinated category than i
    The average age is slightly higher in the vaccinated category.")
```



- This plot shows the effects of `age` and `education` on `flushot`.
- The every educational group on the `vaccinated` side appears to have a positive relationship with `vaccinated`. On the other hand, on the `not vaccinated` side, the second age group seems to have the most counts in `not vaccinated` for almost all educational groups.

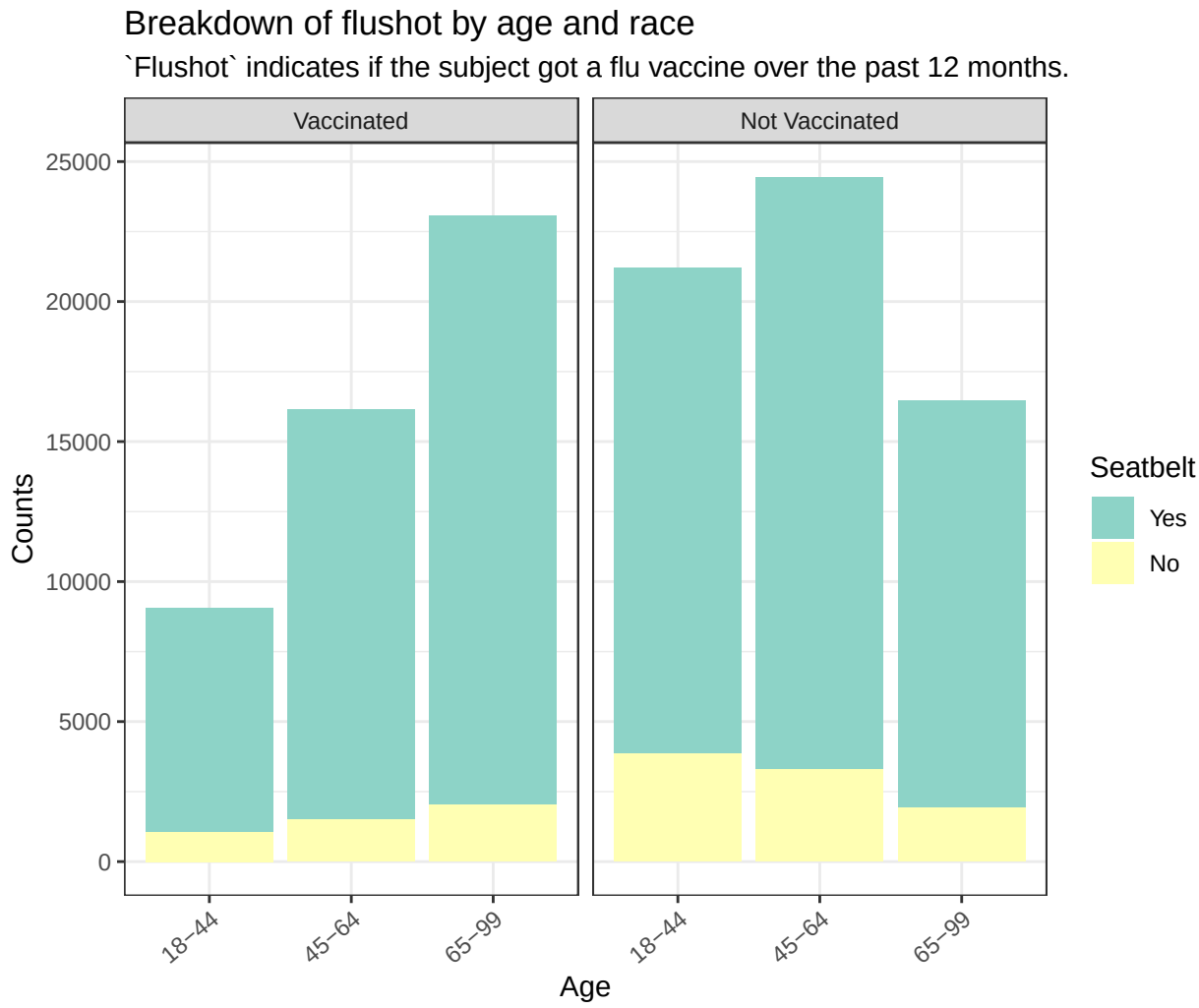
```
ggplot(na.omit(cdc), aes(x = agefact, fill = sbfact)) +
geom_bar() +
theme_bw() +
guides(fill=guide_legend(title="Seatbelt")) +
```

```
scale_fill_brewer(palette = "Set3") +
facet_grid(~flufact) +
theme(axis.text.x = element_text(angle=40, hjust=1)) +
labs(x="Age",
     y="Counts",
     title="Breakdown of flushot by age and race",
     subtitle="`Flushot` indicates if the subject got a flu vaccine over the past 12 months.",
     caption="Source: CDC BRFSS 2016-2018,

     More population from Non-Hispanic Black and `Other` races are in the not-vaccinated category than i
     The average age is slightly higher in the vaccinated category.")
```



- Now, this plot shows the effects of `age` and `seatbelt` on `flushot`.
- The odds of `not always wearing seatbelt` appears substantially higher in the `not vaccinated` category than in the `vaccinated`.
- Build a simpler graph once again.

```r
ggplot(na.omit(cdc), aes(x = age3, fill = sbfact)) +
geom_bar() +
theme_bw() +
guides(fill=guide_legend(title="Seatbelt")) +
scale_fill_brewer(palette = "Set3") +
facet_grid(~flufact) +
theme(axis.text.x = element_text(angle=40, hjust=1)) +
labs(x="Age",
     y="Counts",
     title="Breakdown of flushot by age and race",
     subtitle="`Flushot` indicates if the subject got a flu vaccine over the past 12 months.",
     caption="Source: CDC BRFSS 2016-2018,

     More population from Non-Hispanic Black and `Other` races are in the not-vaccinated category than i
     The average age is slightly higher in the vaccinated category.")
```



Source: CDC BRFSS 2016−2018,

ɔm Non−Hispanic Black and `Other` races are in the not−vaccinated category than in the vaccinated.
The average age is slightly higher in the vaccinated category.

- It shows the more odds of `not always wearing seatbelt` group on the `not vaccinated` side clearer here.

Discussion

My analyses were on a model with a binary outcome, `flushot`, whether or not the subject has got flu vaccination over the past 12 months. It is an extremely important public health question since if an increasing number of people refuse to get yearly vaccinated for seasonal flu, then the herd immunity does not hold any more, and the danger of flu will increase. Hence, knowing who tend not to have received flu vaccination is important. Using basic personal information such as age, race, sex, and education level from the survey data set from BRFSS, I could build a logistic regression model that might help answer my two questions as introduced at the beginning:

1. Who has received flu vaccination over the past 12 months? Can we predict who has received it based on the information such as age, race, sex, and education level?
2. Considering other information mentioned above, whether the subject "always" wears seat belts while driving has a significant impact on whether the subject has received flu vaccination over the past 12 months?

Regarding the first question, all originally included predictors had significant effects on the outcome. Understanding "why" certain groups of people do not receive flu vaccination would be the next step after this cross-sectional study.

Regarding the second question, I used the same regression model I built. Whether or not always wearing seat belts while driving was also significantly associated with our outcome, adjusting for the basic information from the previous question. The association is positive. Studying the reason behind this positive association would be not only interesting but also worth the time and effort because not always wearing seat belts and not getting flu vaccination both could be very harmful to the society as a whole.

While working on this project, I was grateful for the existence of logistic regression and different methods for building it that R provides such as `lrm` and `glm`, making my anlyses a delightful experience. I wish I had found a dataset that had many quantitative predictors. Only with categorical predictors, the visualization part turned out to be a bit too boring; I leaned on barplots much.