Code

ALA Mode (Group 2): Anja Shahu, Anna Wuest, Ligia Flores

11/4/2020

```
library(tidyverse)
library(RColorBrewer)
```

```
url <- "https://meps.ahrq.gov/mepsweb/data_files/pufs/h209dat.zip"
download.file(url, temp <- tempfile())
meps_path <- unzip(temp, exdir = tempdir())
source("https://meps.ahrq.gov/mepsweb/data_stats/download_data/pufs/h209/h209ru.txt")
unlink(temp)
```

```
# creating a reduced data frame including only the variables that we'll be considering
h209red <- data.frame("pap" = h209$ADPAP42,
                      "region" = h209$REGION18,
                      "race" = h209$RACETHX,
                      "age" = h209$AGE18X,
                      "marital_stat" = h209$MARRY18X, # newly added
                      "educ" = h209$EDUCYR, # newly added
                      "smoke_freq" = h209$OFTSMK53, # newly added
                      "income_indiv" = h209$TTLP18X,
                      "income_fam" = h209$FAMINC18,
                      "income_percpov" = h209$POVLEV18,
                      "hrsworked_rd1" = h209$HOUR31H,
                      "hrsworked_rd2" = h209$HOUR42H,
                      "hrsworked_rd3" = h209$HOUR53H,
                      "limitation" = h209$ACTLIM31, # newly added
                      "menhlth_rd1" = h209$MNHLTH31, # already included
                      "menhlth_rd2" = h209$MNHLTH42, # already included
                      "menhlth_rd3" = h209$MNHLTH53, # already included
                      "genhlth_rd1" = h209$RTHLTH31,
                      "genhlth_rd2" = h209$RTHLTH42,
                      "genhlth_rd3" = h209$RTHLTH53,
                      "totexp" = h209$TOTEXP18,
                      "outofpocket_exp" = h209$TOTSLF18,
                      "afford_care" = h209$AFRDCA42,
                      "have_usc" = h209$HAVEUS42,
                      "dist_from_usc" = h209$TMTKUS42,
                      "rch_usc_byphn" = h209$PHNREG42,
                      "usc_offhrs_nw" = h209$OFFHOU42,
                      "usc_asks_abt_trts" = h209$TREATM42,
                      "usc_asks_hlp_dec" = h209$DECIDE42,
                      "usc_expln_options" = h209$EXPLOP42,
                      "usc_spk_lang" = h209$PRVSPK42,
                      "usc_gender" = h209$GENDRP42,
                      "inscov_gen_2018" = h209$INSCOV18)
```

```r
h209red <- h209red %>%
  as_tibble() %>%
  filter(pap != -1) %>% # filtering out the people who were not asked pap smear question
  filter(age >= 21 & age <= 65) # filtering to women ages 21-65 (note there was 1 inapplicable person t
```

```r
# creating factor versions of our categorical variables

# pap status
h209red <- h209red %>%
  mutate(pap_f = factor(pap,
                        levels = c("1", "2", "-15"))) %>%
  mutate(pap_f = fct_recode(pap_f,
                            "yes" = "1",
                            "no" = "2",
                            "cannot be computed" = "-15"))

# region
h209red <- h209red %>%
  mutate(region_f = factor(region,
                           levels = c("1", "2", "3", "4"))) %>%
  mutate(region_f = fct_recode(region_f,
                               "northeast" = "1",
                               "midwest" = "2",
                               "south" = "3",
                               "west" = "4"))

# race
h209red <- h209red %>%
  mutate(race_f = factor(race,
                         levels = c("2", "1", "3", "4", "5"))) %>%
  mutate(race_f = fct_recode(race_f,
                             "white" = "2",
                             "hispanic" = "1",
                             "black" = "3",
                             "asian" = "4",
                             "other or multiple races" = "5"))

# marital status
h209red <- h209red %>%
  mutate(marital_stat_f = factor(marital_stat,
                                 levels = c("5", "1", "2", "3", "4"))) %>%
  mutate(marital_stat_f = fct_recode(marital_stat_f,
                                     "never married" = "5",
                                     "married" = "1",
                                     "widowed" = "2",
                                     "divorced" = "3",
                                     "seperated" = "4"))
# education
h209red <- h209red %>%
  mutate(educ_f = factor(educ)) %>%
  mutate(educ_f = fct_collapse(educ_f,
                               "none or any elementary" = c("0", "1", "2", "3", "4", "5", "6", "7", "8"
                               "any high school" = c("9", "10", "11", "12"),
```

```
                                "any college" = c("13", "14", "15", "16", "17"),
                                "cannot be computed" = "-15",
                                "did not answer" = c("-8", "-7")))

# smoking frequency
h209red <- h209red %>%
  mutate(smoke_freq_f = factor(smoke_freq,
                               levels = c("3", "2", "1", "-8", "-7", "-1"))) %>%
  mutate(smoke_freq_f = fct_recode(smoke_freq_f,
                                   "never" = "3",
                                   "some days" = "2",
                                   "every day" = "1",
                                   "did not answer" = "-8",
                                   "did not answer" = "-7",
                                   "inapplicable" = "-1"))

# limitation
h209red <- h209red %>%
  mutate(limitation_f = factor(limitation,
                               levels = c("2", "1", "-8", "-7", "-1"))) %>%
  mutate(limitation_f = fct_recode(limitation_f,
                                   "no" = "2",
                                   "yes" = "1",
                                   "did not answer" = "-8",
                                   "did not answer" = "-7",
                                   "inapplicable" = "-1"))

# perceived mental health (2nd round)
h209red <- h209red %>%
  mutate(menhlth_rd2_f = factor(menhlth_rd2,
                                levels = c("5", "4", "3", "2", "1", "-8", "-7"))) %>%
  mutate(menhlth_rd2_f = fct_recode(menhlth_rd2_f,
                                    "poor" = "5",
                                    "fair" = "4",
                                    "good" = "3",
                                    "very good" = "2",
                                    "excellent" = "1",
                                    "did not answer" = "-8",
                                    "did not answer" = "-7"))


# perceived general heath (2nd round)
h209red <- h209red %>%
  mutate(genhlth_rd2_f = factor(genhlth_rd2,
                                levels = c("5", "4", "3", "2", "1", "-8"))) %>%
  mutate(genhlth_rd2_f = fct_recode(genhlth_rd2_f,
                                    "poor" = "5",
                                    "fair" = "4",
                                    "good" = "3",
                                    "very good" = "2",
                                    "excellent" = "1",
                                    "did not answer" = "-8"))
```

```r
# ability to afford care
h209red <- h209red %>%
  mutate(afford_care_f = factor(afford_care,
                                levels = c("2", "1", "-8", "-7"))) %>%
  mutate(afford_care_f = fct_recode(afford_care_f,
                                    "no" = "2",
                                    "yes" = "1",
                                    "did not answer" = "-8",
                                    "did not answer" = "-7"))


# usual source of care status
h209red <- h209red %>%
  mutate(have_usc_f = factor(have_usc,
                             levels = c("2", "1", "-8", "-7"))) %>%
  mutate(have_usc_f = fct_recode(have_usc_f,
                                 "no" = "2",
                                 "yes" = "1",
                                 "did not answer" = "-8",
                                 "did not answer" = "-7"))

# distance from provider
h209red <- h209red %>%
  mutate(dist_from_usc = ifelse(have_usc_f == "no",
                                -100,
                                dist_from_usc)) %>% # creating level for not having a provider
  mutate(dist_from_usc_f = factor(dist_from_usc,
                                  levels = c("1", "2", "3", "4", "5", "6", "-100", "-8", "-7", "-1"))) %
  mutate(dist_from_usc_f = fct_recode(dist_from_usc_f,
                                      "<15" = "1",
                                      "15 to 30" = "2",
                                      "31 to 60" = "3",
                                      "61 to 90" = "4",
                                      "91 to 120" = "5",
                                      ">120" = "6",
                                      "no usc" = "-100",
                                      "did not answer" = "-8",
                                      "did not answer" = "-7",
                                      "inapplicable" = "-1",))

# ability to reach provider by phone
h209red <- h209red %>%
  mutate(rch_usc_byphn = ifelse(have_usc_f == "no",
                                -100,
                                rch_usc_byphn)) %>% # creating level for not having a provider
  mutate(rch_usc_byphn_f = factor(rch_usc_byphn,
                                  levels = c("4", "3", "2", "1", "-100", "-8", "-7", "-1"))) %>%
  mutate(rch_usc_byphn_f = fct_recode(rch_usc_byphn_f,
                                      "not at all difficult" = "4",
                                      "not too difficult" = "3",
                                      "somewhat difficult" = "2",
                                      "very difficult" = "1",
                                      "no usc" = "-100",
```

```
                                    "did not answer" = "-8",
                                    "did not answer" = "-7",
                                    "inapplicable" = "-1"))

# provider offers office hours during nights/weekends
h209red <- h209red %>%
  mutate(usc_offhrs_nw = ifelse(have_usc_f == "no",
                                -100,
                                usc_offhrs_nw)) %>% # creating level for not having a provider
  mutate(usc_offhrs_nw_f = factor(usc_offhrs_nw,
                                  levels = c("-100", "2", "1", "-8", "-7", "-1"))) %>%
  mutate(usc_offhrs_nw_f = fct_recode(usc_offhrs_nw_f,
                                    "no usc" = "-100",
                                    "no" = "2",
                                    "yes" = "1",
                                    "did not answer" = "-8",
                                    "did not answer" = "-7",
                                    "inapplicable" = "-1"))

# provider asks about treatments
h209red <- h209red %>%
  mutate(usc_asks_abt_trts = ifelse(have_usc_f == "no",
                                -100,
                                usc_asks_abt_trts)) %>% # creating level for not having a provider
  mutate(usc_asks_abt_trts_f = factor(usc_asks_abt_trts,
                                    levels = c("-100", "2", "1", "-8", "-7", "-1"))) %>%
  mutate(usc_asks_abt_trts_f = fct_recode(usc_asks_abt_trts_f,
                                        "no usc" = "-100",
                                        "no" = "2",
                                        "yes" = "1",
                                        "did not answer" = "-8",
                                        "did not answer" = "-7",
                                        "inapplicable" = "-1"))

# provider asks person to help make decisions
h209red <- h209red %>%
   mutate(usc_asks_hlp_dec = ifelse(have_usc_f == "no",
                                -100,
                                usc_asks_hlp_dec)) %>% # creating level for not having a provider
  mutate(usc_asks_hlp_dec_f = factor(usc_asks_hlp_dec,
                                    levels = c("-100", "1", "2", "3", "4", "-8", "-7", "-1"))) %>%
  mutate(usc_asks_hlp_dec_f = fct_recode(usc_asks_hlp_dec_f,
                                       "no usc" = "-100",
                                       "never" = "1",
                                       "sometimes" = "2",
                                       "usually" = "3",
                                       "always" = "4",
                                       "did not answer" = "-8",
                                       "did not answer" = "-7",
                                       "inapplicable" = "-1"))

# provider presents and explains all options
h209red <- h209red %>%
```

```
  mutate(usc_expln_options = ifelse(have_usc_f == "no",
                                    -100,
                                    usc_expln_options)) %>% # creating level for not having a provider
  mutate(usc_expln_options_f = factor(usc_expln_options,
                                      levels = c("-100", "2", "1", "-8", "-7", "-1"))) %>%
  mutate(usc_expln_options_f = fct_recode(usc_expln_options_f,
                                          "no usc" = "-100",
                                          "no" = "2",
                                          "yes" = "1",
                                          "did not answer" = "-8",
                                          "did not answer" = "-7",
                                          "inapplicable" = "-1"))

# provider speaks person's language
h209red <- h209red %>%
  mutate(usc_spk_lang = ifelse(have_usc_f == "no",
                               -100,
                               usc_spk_lang)) %>% # creating level for not having a provider
  mutate(usc_spk_lang_f = factor(usc_spk_lang,
                                 levels = c("-100", "2", "1", "-8", "-7", "-1"))) %>%
  mutate(usc_spk_lang_f = fct_recode(usc_spk_lang_f,
                                     "no usc" = "-100",
                                     "no" = "2",
                                     "yes" = "1",
                                     "did not answer" = "-8",
                                     "did not answer" = "-7",
                                     "inapplicable" = "-1"))

# gender of provider
h209red <- h209red %>%
  mutate(usc_gender = ifelse(have_usc_f == "no",
                             -100,
                             usc_gender)) %>% # creating level for not having a provider
  mutate(usc_gender_f = factor(usc_gender,
                               levels = c("1", "2", "-100", "-8", "-1"))) %>%
  mutate(usc_gender_f = fct_recode(usc_gender_f,
                                   "male" = "1",
                                   "female" = "2",
                                   "no usc" = "-100",
                                   "did not answer" = "-8",
                                   "inapplicable" = "-1"))

# insurance indicator in 2018
h209red <- h209red %>%
  mutate(inscov_gen_2018_f = factor(inscov_gen_2018,
                                    levels = c("1", "2", "3"))) %>%
  mutate(inscov_gen_2018_f = fct_recode(inscov_gen_2018_f,
                                        "any private" = "1",
                                        "public only" = "2",
                                        "uninsured" = "3"))
```

```
# checking if any variables will be too problematic (in terms of missing values) to use
```

```r
# pap smear variable
table(h209red$pap_f) # about 600 that weren't able to be computed
```

```
##
##                yes                 no cannot be computed
##               4480               1561                595
```

```r
# demographic variables
# region
table(h209red$region_f) # looks good!
```

```
##
## northeast   midwest     south      west
##      1023      1356      2612      1645
```

```r
# race
table(h209red$race_f) # looks good!
```

```
##
##                   white                hispanic                   black
##                    3293                    1680                    1048
##                   asian other or multiple races
##                     395                     220
```

```r
# age
h209red %>% filter(age < 0) %>% summarise(n = n()) # looks good!
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

```r
# marital status
table(h209red$marital_stat_f) # looks good
```

```
##
## never married       married       widowed      divorced     seperated
##          1972          3440           194           805           225
```

```r
# education level
table(h209red$educ_f)
```

```
##
##     cannot be computed         did not answer none or any elementary
##                      1                     46                    314
##        any high school            any college
##                   2379                   3896
```

```r
# smoking variable
table(h209red$smoke_freq_f) # only a few missing
```

```
##
##          never      some days      every day did not answer   inapplicable
##           5642            293            676             16              9
```

```r
# income variables
# individual income
h209red %>% filter(income_indiv < 0) %>% summarise(n = n())
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     9
```

```
# family income
h209red %>% filter(income_fam < 0) %>% summarise(n = n())
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     3
```

```
# family income as percent of poverty line
h209red %>% filter(income_percpov < 0) %>% summarise(n = n())
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     3
```

```
# in all of these we have some negative incomes but no missing ones!


# hours worked variable
h209red$hrsworked_rd1[h209red$hrsworked_rd1 == -1] <- NA
h209red$hrsworked_rd2[h209red$hrsworked_rd2 == -1] <- NA
h209red$hrsworked_rd3[h209red$hrsworked_rd3 == -1] <- NA


# limitation variable
table(h209red$limitation_f) # only a few missing
```

```
##
##             no            yes did not answer   inapplicable
##           6075            545             12              4
```

```
# health variables
# mental health
table(h209red$menhlth_rd2_f) # use second round analysis
```

```
##
##           poor           fair           good      very good      excellent
##            106            424           1847           2128           2128
## did not answer
##              3
```

```
# general health
table(h209red$genhlth_rd2_f) # use second round in analysis
```

```
##
##           poor           fair           good      very good      excellent
##            149            679           2006           2325           1476
## did not answer
##              1
```

```
# expenditure variables
# total exp
```

```r
h209red %>% filter(totexp < 0) %>% summarise(n = n()) # no missing
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

```r
# out of pocket
h209red %>% filter(outofpocket_exp < 0) %>% summarise(n = n()) # no missing
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

```r
# ability to afford care
table(h209red$afford_care_f) # only a few missing
```

```
##
##             no            yes did not answer
##           6080            542             14
```

```r
# access and provider satisfaction variables
table(h209red$have_usc_f)
```

```
##
##             no            yes did not answer
##           1823           4695            118
```

```r
table(h209red$dist_from_usc_f)
```

```
##
##            <15       15 to 30       31 to 60       61 to 90      91 to 120
##           2720           1495            396             46             15
##           >120        no usc did not answer   inapplicable
##             15           1823              8            118
```

```r
table(h209red$rch_usc_byphn_f)
```

```
##
## not at all difficult    not too difficult   somewhat difficult
##                 2296                 1371                  541
##       very difficult              no usc       did not answer
##                  253                1823                  234
##         inapplicable
##                  118
```

```r
table(h209red$usc_offhrs_nw_f)
```

```
##
##         no usc             no            yes did not answer   inapplicable
##           1823           2687           1466            516            144
```

```r
table(h209red$usc_asks_abt_trts_f)
```

```
##
##         no usc             no            yes did not answer   inapplicable
##           1823            873           3670            152            118
```

```r
table(h209red$usc_asks_hlp_dec_f)
```

```
##
##         no usc          never      sometimes        usually         always
##           1823            422            653            939           2366
## did not answer   inapplicable
##            315            118
```

```r
table(h209red$usc_expln_options_f)
```

```
##
##         no usc             no            yes did not answer   inapplicable
##           1823            196           4370            129            118
```

```r
# these have a few missing values (some variables more than others)


# variable for if provider speaks person's language
table(h209red$usc_spk_lang_f) # we see a lot of inapplicables
```

```
##
##         no usc             no            yes did not answer   inapplicable
##           1823             18            324              5           4466
```

```r
# according to code book "PRVSPK42 is set to a value other than '-1' (Inapplicable) for persons eligibl
# Access to Care supplement, who had a usual source of care, and were identified as speaking a language
# other than English at home (OTHLANG = '1') and speaking English either "Not Well" or "Not at All"
# (HWELLSPK = '3' or '4'). PRVSPK42 is set to '-1' (Inapplicable) for all persons not meeting these
# criteria or who were deceased, institutionalized, or younger than 5 years of age.
# I think it may be difficult to deal with since it seems to be dependent on a few other variables and
# those extra variables in to be able to create a usable variable  with levels for
# "inapplicable", "native english speaker", "answers yes", "answers no"


# gender of provider
table(h209red$usc_gender_f)
```

```
##
##           male         female         no usc did not answer   inapplicable
##           1541           1683           1823              6           1583
```

```r
# seems to be too many people skipped for this variable to be used


# insurance coverage variable
table(h209red$inscov_gen_2018_f) # looks good!
```

```
##
## any private public only   uninsured
##        4437        1498         701
```

```r
# inserting NAs into variables

# pap status
h209red <- h209red %>%
  mutate(pap_f = fct_recode(pap_f,
                            NULL = "cannot be computed"))
```

```r
# region
# no NAs

# race
# no NAs

# age
# no NAs

# marital status
# no NAs

# education
h209red <- h209red %>%
  mutate(educ_f = fct_recode(educ_f,
                             NULL = "cannot be computed"))

# smoking frequency
h209red <- h209red %>%
  mutate(smoke_freq_f = fct_recode(smoke_freq_f,
                                   NULL = "did not answer",
                                   NULL = "inapplicable"))


# income variables
# in all 3 of these we have some negative incomes but no missing ones
# should we make the negative ones missing values???

# hours worked
h209red %>% filter(hrsworked_rd1 < 0) %>% summarise(n = n())
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

```r
h209red %>% filter(hrsworked_rd2 < 0) %>% summarise(n = n())
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

```r
h209red %>% filter(hrsworked_rd3 < 0) %>% summarise(n = n())
```

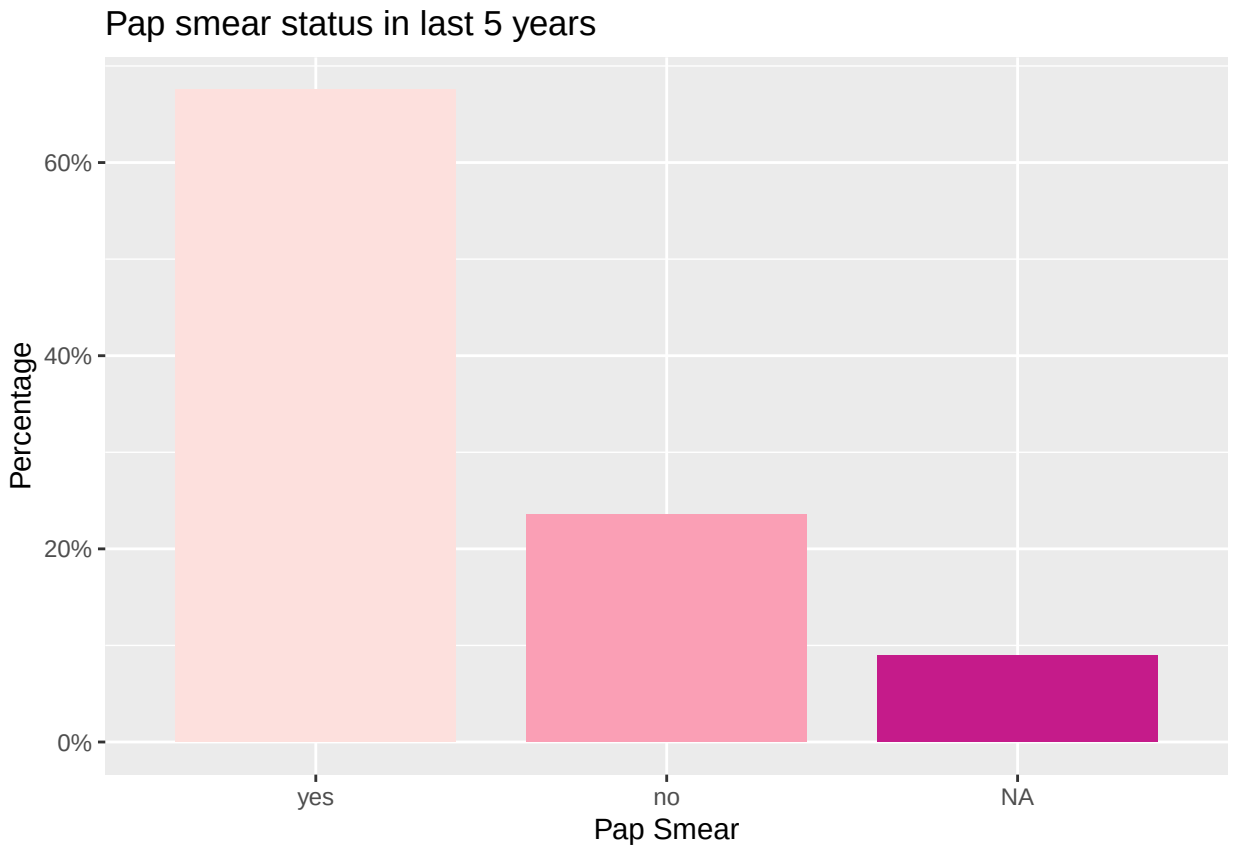
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

```r
# limitation
h209red <- h209red %>%
  mutate(limitation_f = fct_recode(limitation_f,
                                   NULL = "did not answer",
                                   NULL = "inapplicable"))
```

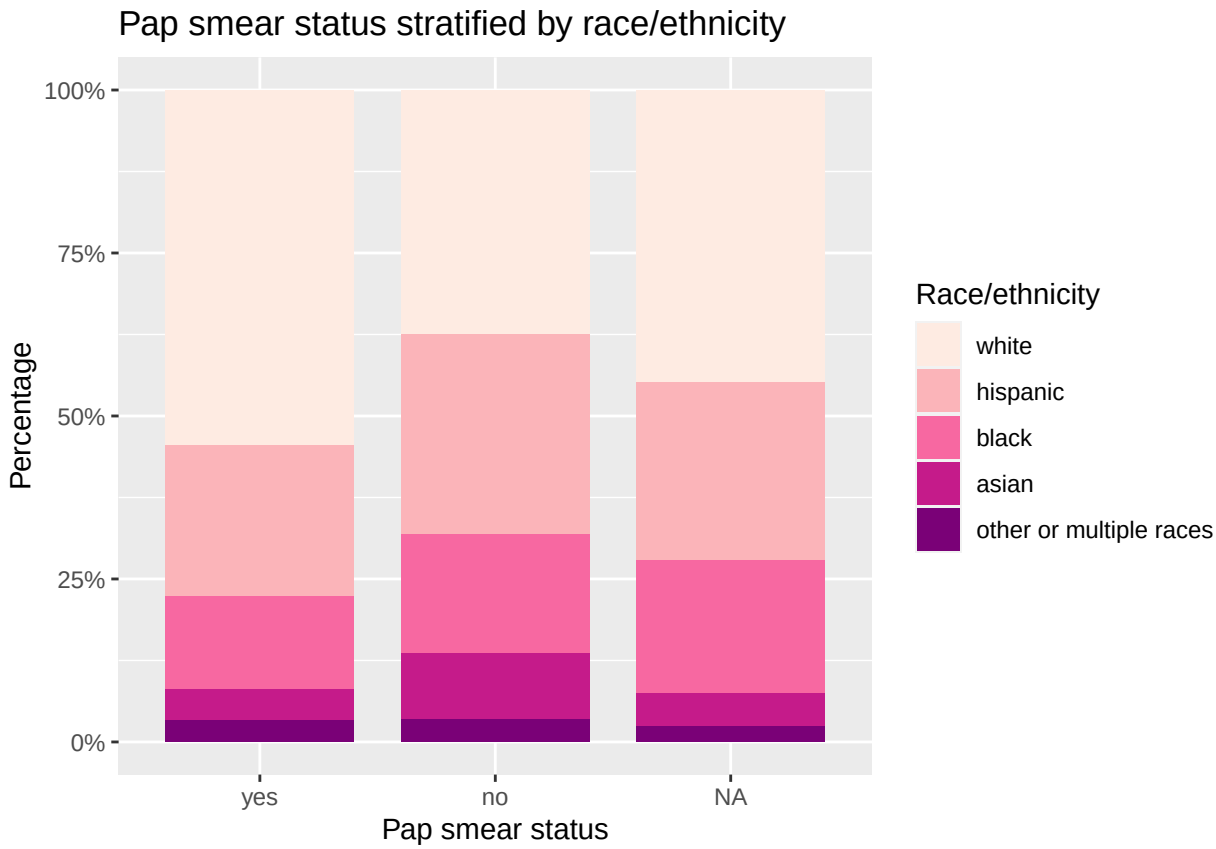
```r
# perceived mental health (2nd round)
h209red <- h209red %>%
  mutate(menhlth_rd2_f = fct_recode(menhlth_rd2_f,
                                    NULL = "did not answer"))

# perceived general heath (2nd round)
h209red <- h209red %>%
  mutate(genhlth_rd2_f = fct_recode(genhlth_rd2_f,
                                    NULL = "did not answer"))

# total exp
# no NAs

# out of pocket
# no NAs

# ability to afford care
h209red <- h209red %>%
  mutate(afford_care_f = fct_recode(afford_care_f,
                                    NULL = "did not answer"))

# usual source of care status
h209red <- h209red %>%
  mutate(have_usc_f = fct_recode(have_usc_f,
                                 NULL = "did not answer"))

# distance from provider
h209red <- h209red %>%
  mutate(dist_from_usc_f = fct_recode(dist_from_usc_f,
                                      NULL = "did not answer",
                                      NULL ="inapplicable"))

# ability to reach provider by phone
h209red <- h209red %>%
  mutate(rch_usc_byphn_f = fct_recode(rch_usc_byphn_f,
                                      NULL = "did not answer",
                                      NULL = "inapplicable"))


# provider offers office hours during nights/weekends
h209red <- h209red %>%
  mutate(usc_offhrs_nw_f = fct_recode(usc_offhrs_nw_f,
                                      NULL = "did not answer",
                                      NULL = "inapplicable"))


# provider asks about treatments
h209red <- h209red %>%
  mutate(usc_asks_abt_trts_f = fct_recode(usc_asks_abt_trts_f,
                                          NULL = "did not answer",
                                          NULL = "inapplicable"))

# provider asks person to help make decisions
```

```r
h209red <- h209red %>%
  mutate(usc_asks_hlp_dec_f = fct_recode(usc_asks_hlp_dec_f,
                                         NULL = "did not answer",
                                         NULL = "inapplicable"))

# provider presents and explains all options
h209red <- h209red %>%
  mutate(usc_expln_options_f = fct_recode(usc_expln_options_f,
                                          NULL = "did not answer",
                                          NULL = "inapplicable"))

# gender of provider
h209red <- h209red %>%
  mutate(usc_gender_f = fct_recode(usc_gender_f,
                                   NULL = "did not answer",
                                   NULL = "inapplicable"))

# insurance indicator in 2018
# no NAs
```

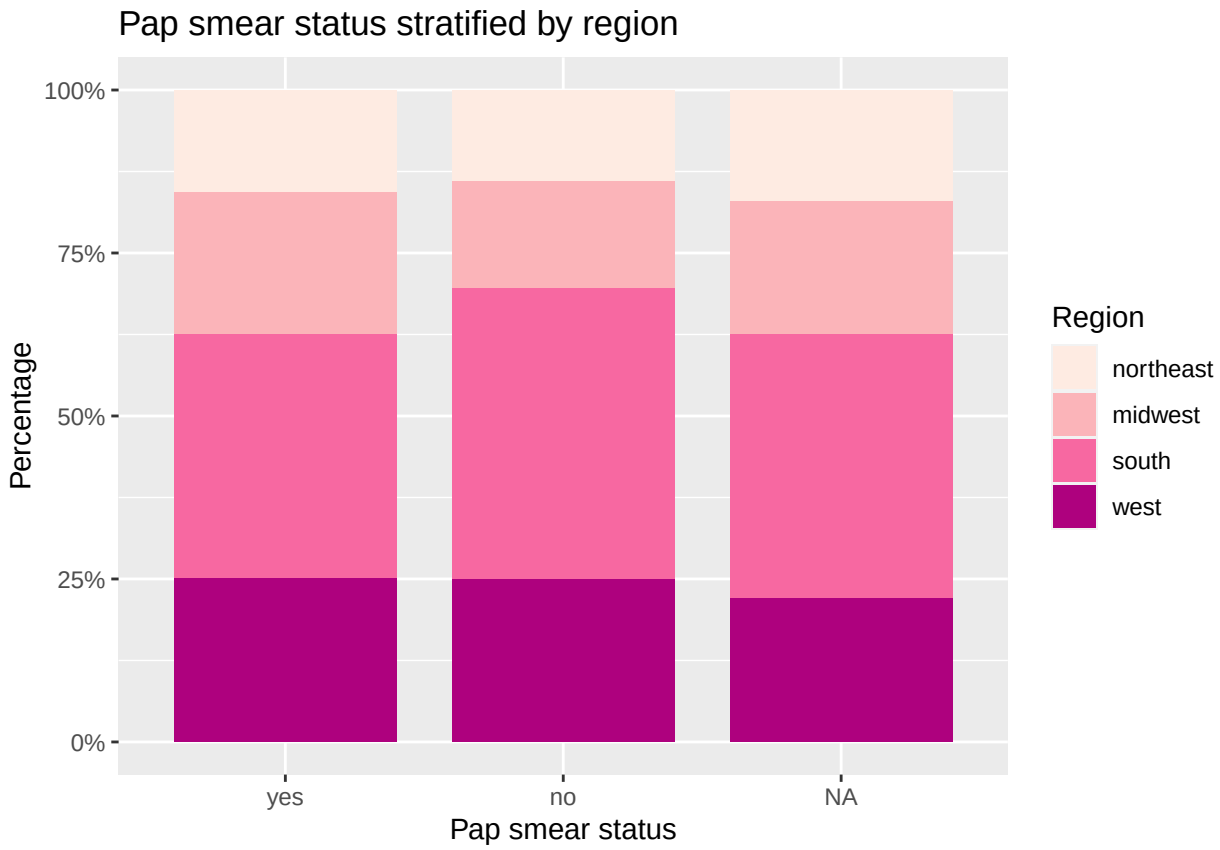
```r
# looking at pap smear variable
h209red %>%
  ggplot(aes(x = pap_f)) +
  geom_bar(aes(y = ..prop.., group = 1),
           stat = "count",
           fill = brewer.pal(n = 3, name = "RdPu"),
           width = 0.8) +
  theme(axis.text = element_text(size = 9),
        axis.title = element_text(size = 11)) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Pap Smear",
       y = "Percentage") +
  ggtitle("Pap smear status in last 5 years")
```




```
# creating function for two (categorical) variable bar plots with percentages
create_two_var_bar <- function(x_var, fill_var, x_label, fill_label) {
  h209red %>%
    ggplot(aes_string(x = x_var, fill = fill_var)) +
    geom_bar(position = "fill", width = 0.8) +
    theme(axis.text = element_text(size = 9),
          axis.title = element_text(size = 11),
          legend.title = element_text(size = 11),
          legend.text = element_text(size = 9)) +
    scale_y_continuous(labels = scales::percent) +
    scale_fill_brewer(palette = "RdPu") +
    labs(x = x_label,
         y = "Percentage",
         fill = fill_label) +
    ggtitle(paste(x_label, "stratified by", tolower(fill_label)))
}

# bar plot of pap smear vs. race/ethnicity
create_two_var_bar("pap_f", "race_f", "Pap smear status", "Race/ethnicity")
```




```
# bar plot of pap smear vs. region
create_two_var_bar("pap_f", "region_f", "Pap smear status", "Region")
```


Pap smear status stratified by region
100%
75%
50%
25%
0%
Percentage
yes
no
NA
Pap smear status
Region
northeast
midwest
south
west