Homework 4: 1st Project Check-In

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

10/6/2020

1. What is the general domain/subject area of this project?

Women's health and disparities in cervical cancer screening

2. What data will you use, and what is the source?

We are using the Medical Expenditure Panel Survey (MEPS), collected by the U.S. Department of Health and Human Services. We will be using the 2018 Full Year Consolidated Data File from the publicly available Household Component of MEPS.

3. What primary questions will you seek to answer?

How well do individual-level factors and accessibility to healthcare predict the likelihood of getting a pap smear in the last 5 years?

4. What secondary questions will you seek to answer?

1. What is the effect of access to healthcare variables on the probability of obtaining a pap smear in the last 5 years?
2. Do demographic characteristics of age, race/ethnicity and income confound the effects these access variables have on patients obtaining a pap smear?
3. Are age, race/ethnicity and income effect modifiers of the association between access to healthcare variables and obtaining a pap smear?

5. What outcome(s)/endpoint(s) will you use? (could be continuous, binary, polytomous, Poisson, survival,. . . and you may be considering more than one–and this may be updated/added to, as the semester progresses)

We will use cross validation to predict a binary outcome of success (success = getting a pap smear within the past 5 years, failure = not getting a pap smear within the past 5 years).

6. What is your draft Statistical Analysis Plan? (should be an outline, but more detailed than that for Discussion Board). Note that we will be discussing all forms of outcome/endpoint data in this course, and at present have not yet covered each of these. . . so this plan may be updated/added to as the semester progresses. If your outcome data is other than continuous, you do not need to go into statistical detail other than to state 'Regression modeling involving 'X' outcome data of interest, involving these variables (list them). . . ' and any other concerns or methods of interest (listing potential confounders, effect modifiers, potential use of splines or additive modeling, potential missing data considerations, data reduction methods, regularization methods, etc,. . . or none of these–you will want to consider what is most appropriate for your data and questions at hand).

1. Clean the data
   - Restrict data to those who were asked PAP smear question (8728 individuals)
   - Recode variables as needed
   - Address missing values
2. Explore associations between access to health care variables and obtaining a pap smear by fitting logistic regression models involving binary outcome of obtaining a pap smear test in the last 5 years (`ADPAP42`)

   - Test for confounding from demographic characteristics of age, race/ethnicity and income
   - Test for effect modification of age, race/ethnicity and income

3. Create a tool to predict someone's likelihood of obtaining a pap smear by fitting a logistic regression model involving binary outcome of obtaining a pap smear test in the last 5 years (ADPAP42) and considering these potential predictors:
   - Expenditure variables
     - Total health care expenditure (TOTEXP18)
     - Total out of pocket expenses from patient (TOTSLF18)
     - Patient's ability to afford medical care (AFRDCA42)
   - Access to care/provider variables
     - Patient has usual source of care (USC) provider (HAVEUS42)
     - Combined score for provider availability based on
       - How long it takes to get to USC (TMTKUS42)
       - How difficult to contact USC by phone (PHNREG42)
       - USC has office hours during nights/weekends (OFFHOU42)
     - Combined score reflecting patient satisfaction with provider based on
       - Provider usually ask about prescription medications and treatments other doctors may give them (TREATM42)
       - Provider ask the person to help make decisions between a choice of treatments (DECIDE42)
       - Provider presents and explains all options to the person (EXPLOP42)
       - Provider speaks the person's language or provides translation services (PRVSPK42)
     - Gender of provider (GENDRP42)
     - Health insurance coverage indicator (INSCOV18) and full year insurance coverage status in 2018 (INSURC18) (note: numerous variables address monthly private insurance coverage, we may also use these but we will not list all of the variable names here)
   - Demographic/individual-level variables
     - Region (REGION18)
     - Race/ethnicity (RACETHX)
     - Age (AGE18X)
     - Income
       - Person's total income (TTLP18X)
       - Family's total income (FAMINC18)
       - Family income as % of poverty line represented (POVLEV18)
     - Hours worked per week (combine HOUR31H, HOUR42H and HOUR53H)
     - Perceived mental health status (combine MNHLTH31, MNHLTH42 and MNHLTH53)
     - Perceived general health (combine RTHLTH31, RTHLTH42 and RTHLTH53).
4. Use cross validation to train our model on 90% of the data and test for predictive value on the other 10%.

7. What are the biggest challenges you foresee in answering your proposed questions and completing this project? (logistical, statistical, etc, if there are any)

The data set has over 1500 variables that we need to choose from and then recode as needed, which will take some time. If we include too many variables in our prediction model, we risk overfitting the model, which would make the model inappropriate to use for future samples. Since the MEPS survey used a skip pattern and did not ask every question to each participant, there are missing values for certain variables, so we also risk drastically reducing our data set if we include too many variables with missing values. However, if we do not have enough variables in our model, we may miss important predictors and underfit. Therefore, finding the optimal number of variables and determining the most predictive may present a challenge. On the note of not missing any predictors, we may also be limited by the data set itself — if it does not include certain relevant variables then we are inherently limited from using it. In addition, we have more typical challenges when dealing with data such as: missing values due to survey participants refusing to answer and lack of specific domain knowledge.

8. Will you seek domain expertise? Why or why not? If so, from whom?

We will reach out to a professor at HSPH who does research in general cancer disparities or cervical cancer screening (e.g. Dr. Jane Kim). We will also reference literature on cervical cancer.

9. What software package(s) will you use to complete this project? (It is absolutely fine for different group members to use different packages; in fact some tasks are easier in some packages over others and vice versa.)

We are going to use R.

10. Complete an initial round of exploratory analyses on your data that would be relevant to your plan and responses above, and include any plots, summaries, code and output. Please include exploratory analysis for outcome(s) of continuous form however/wherever possible even if your ultimate goals/questions involve a different form of outcome data such as binary, polytomous, etc. (You may consider this initial analysis as a potential sub-analysis later on.)

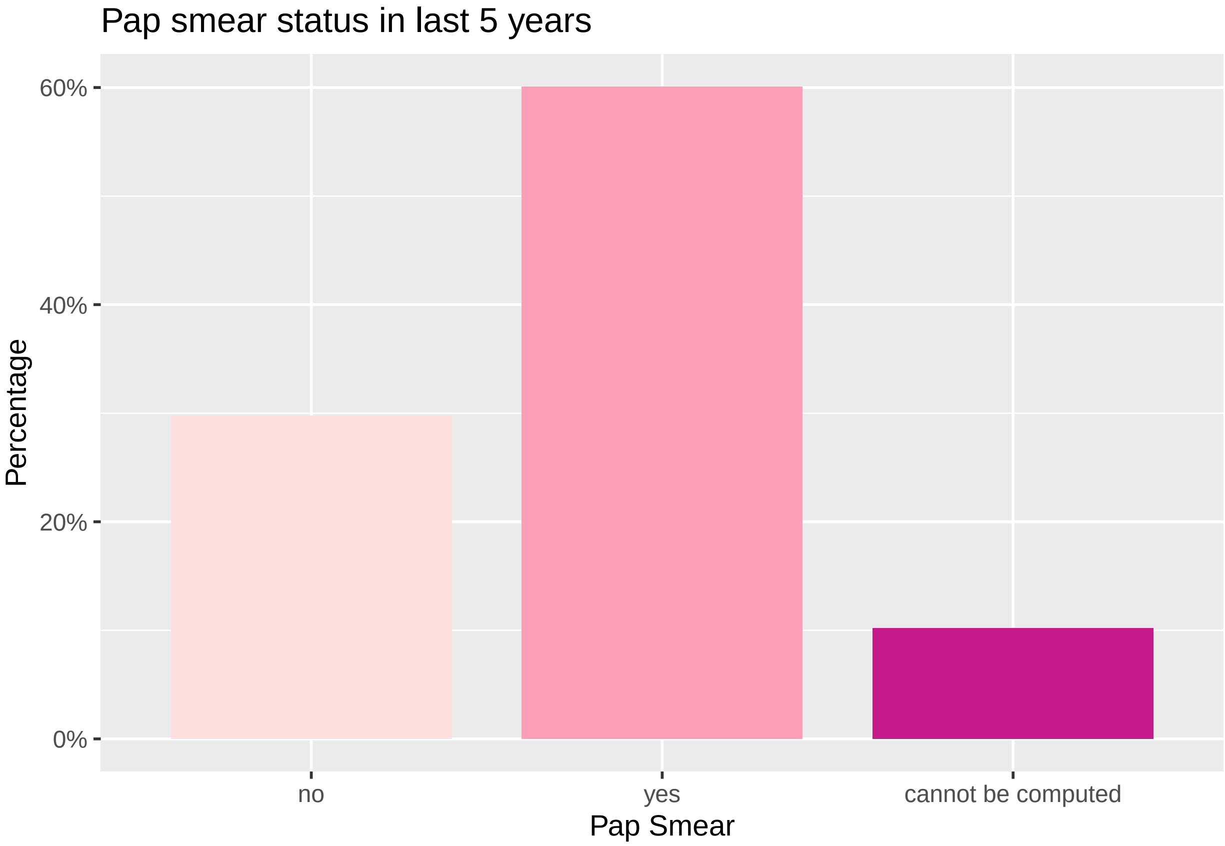
```
url <- "https://meps.ahrq.gov/mepsweb/data_files/pufs/h209dat.zip"
download.file(url, temp <- tempfile())
meps_path <- unzip(temp, exdir = tempdir())
source("https://meps.ahrq.gov/mepsweb/data_stats/download_data/pufs/h209/h209ru.txt")
unlink(temp)
```

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

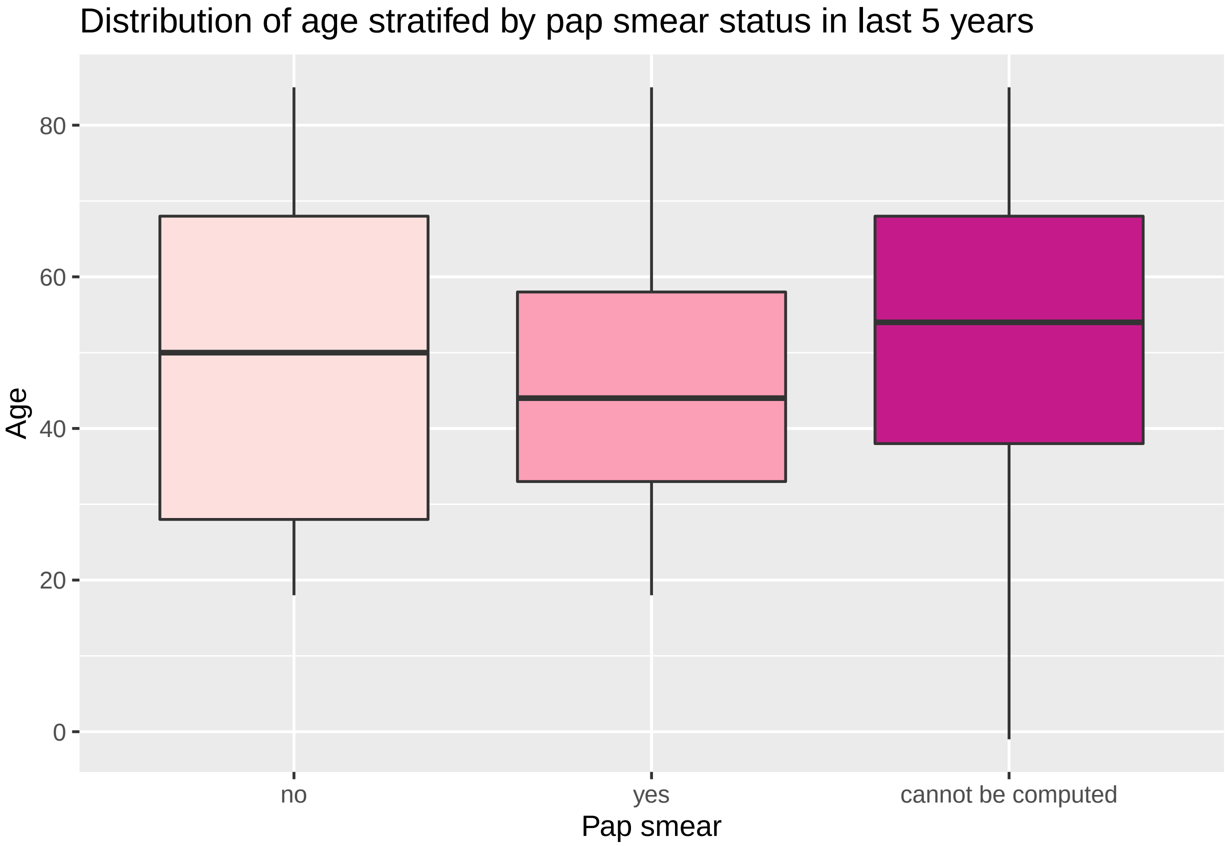
```
h209red <- data.frame("pap" = h209$ADPAP42,
                       "region" = h209$REGION18,
                       "race" = h209$RACETHX,
                       "age" = h209$AGE18X,
                       "income_indiv" = h209$TTLP18X,
                       "income_fam" = h209$FAMINC18,
                       "income_percpov" = h209$POVLEV18,
                       "hrsworked_rd1" = h209$HOUR31H,
                       "hrsworked_rd2" = h209$HOUR42H,
                       "hrsworked_rd3" = h209$HOUR53H,
                       "menhlth_rd1" = h209$MNHLTH31,
                       "menhlth_rd2" = h209$MNHLTH42,
                       "menthlth_rd3" = h209$MNHLTH53,
                       "genhlth_rd1" = h209$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$INSCOV18,
                       "inscov_fullyr_2018" = h209$INSURC18)

h209red <- h209red %>%
  as_tibble() %>%
  filter(pap != -1) # filtering out the people who were not asked pap smear question
```

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

h209red <- h209red %>%
  mutate(region_f = factor(region,
                           levels = c("1", "2", "3", "4", "-1"))) %>%
  mutate(region_f = fct_recode(region_f,
                               "northeast" = "1",
                               "midwest" = "2",
                               "south" = "3",
                               "west" = "4",
                               "not asked" = "-1"))

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"))

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"))

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"))

h209red <- h209red %>%
  mutate(usc_gender_f = factor(usc_gender,
                               levels = c("1", "2", "-8", "-1"))) %>%
  mutate(usc_gender_f = fct_recode(usc_gender_f,
                                   "male" = "1",
                                   "female" = "2",
                                   "did not answer" = "-8",
                                   "not asked" = "-1"))
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"))
```

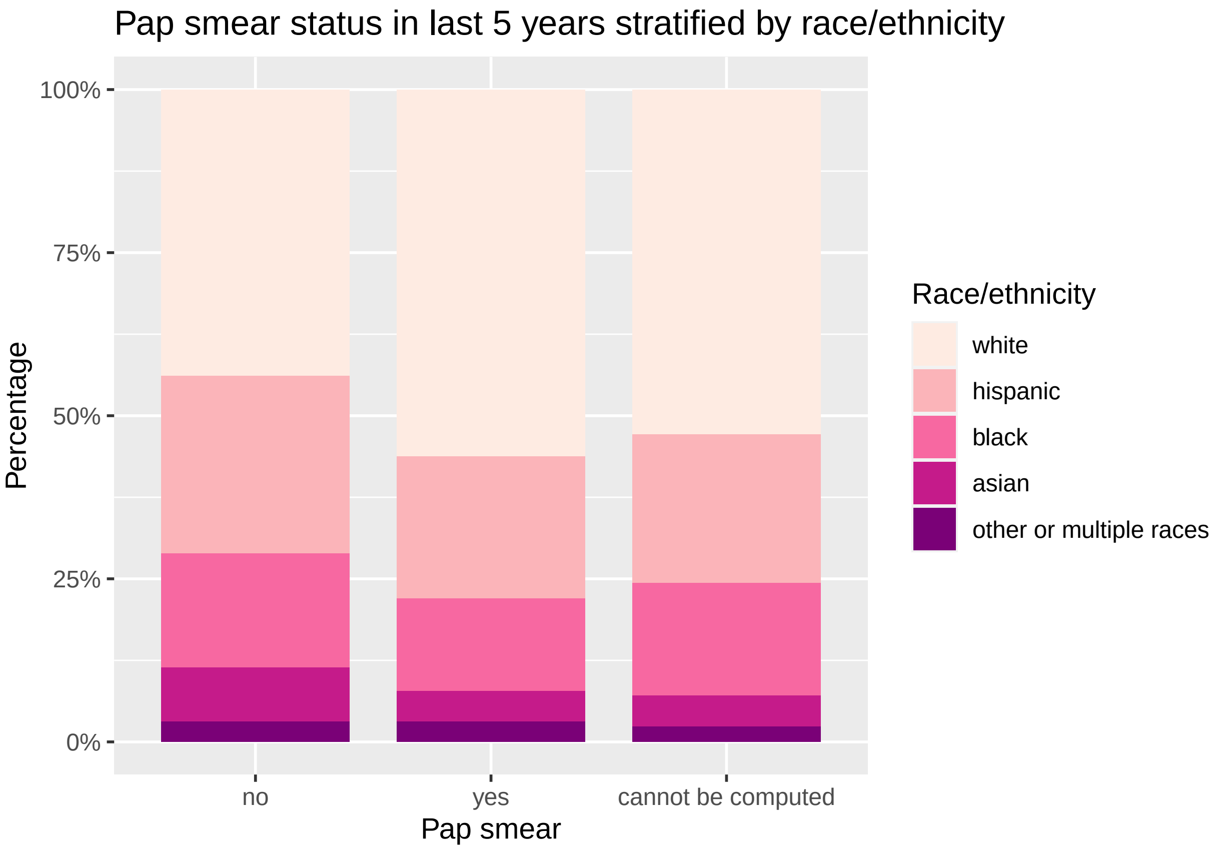
```
h209red %>%
  ggplot(aes(x = pap_f)) +
  geom_bar(aes(y = ..prop.., group = 1),
           stat = "count",
           fill = brewer.pal(n = 3, name = "RdPu"),
           width = 0.8) +
  theme(axis.text = element_text(size = 9),
        axis.title = element_text(size = 11)) +
  scale_y_continuous(labels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



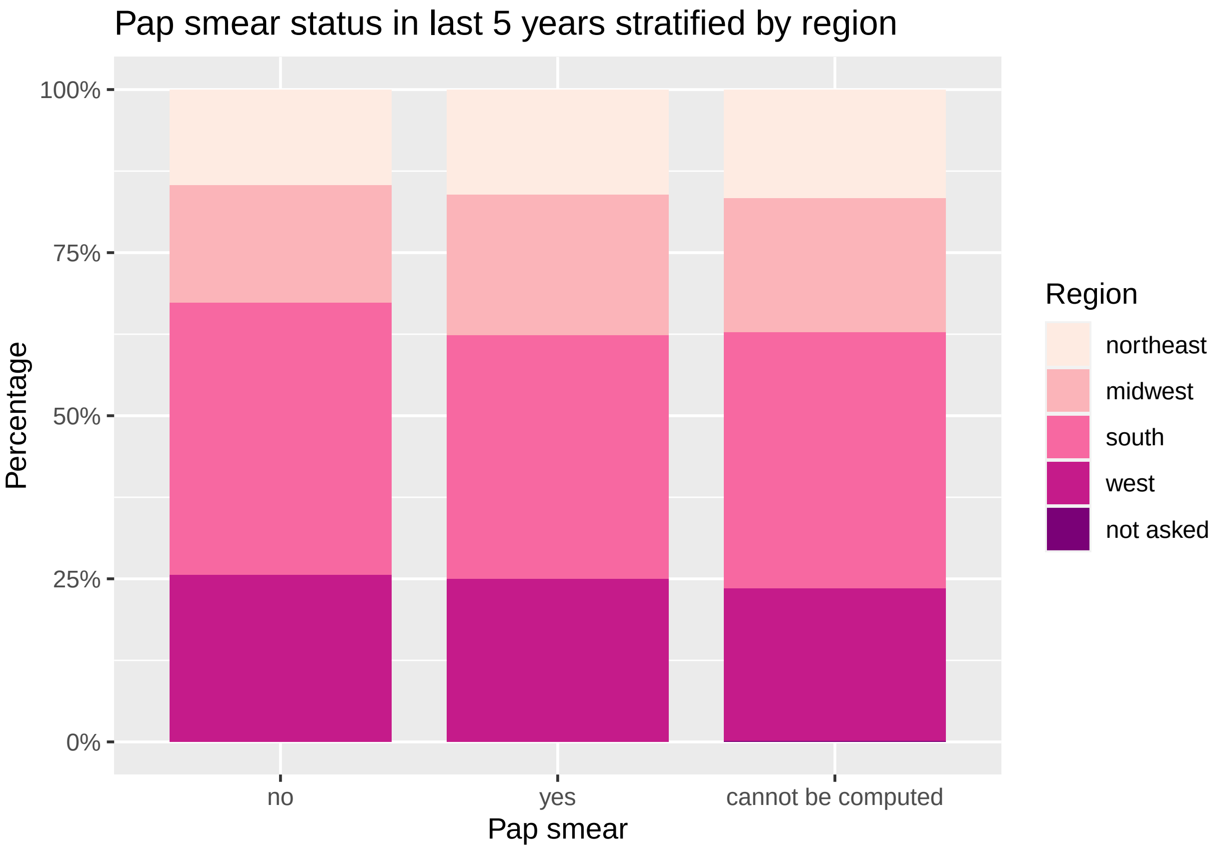
```
h209red %>%
  ggplot(aes(x = pap_f, y = age)) +
  geom_boxplot(fill = brewer.pal(n = 3, name = "RdPu")) +
  labs(x = "Pap smear",
       y = "Age") +
  ggtitle("Distribution of age stratifed by pap smear status in last 5 years")
```



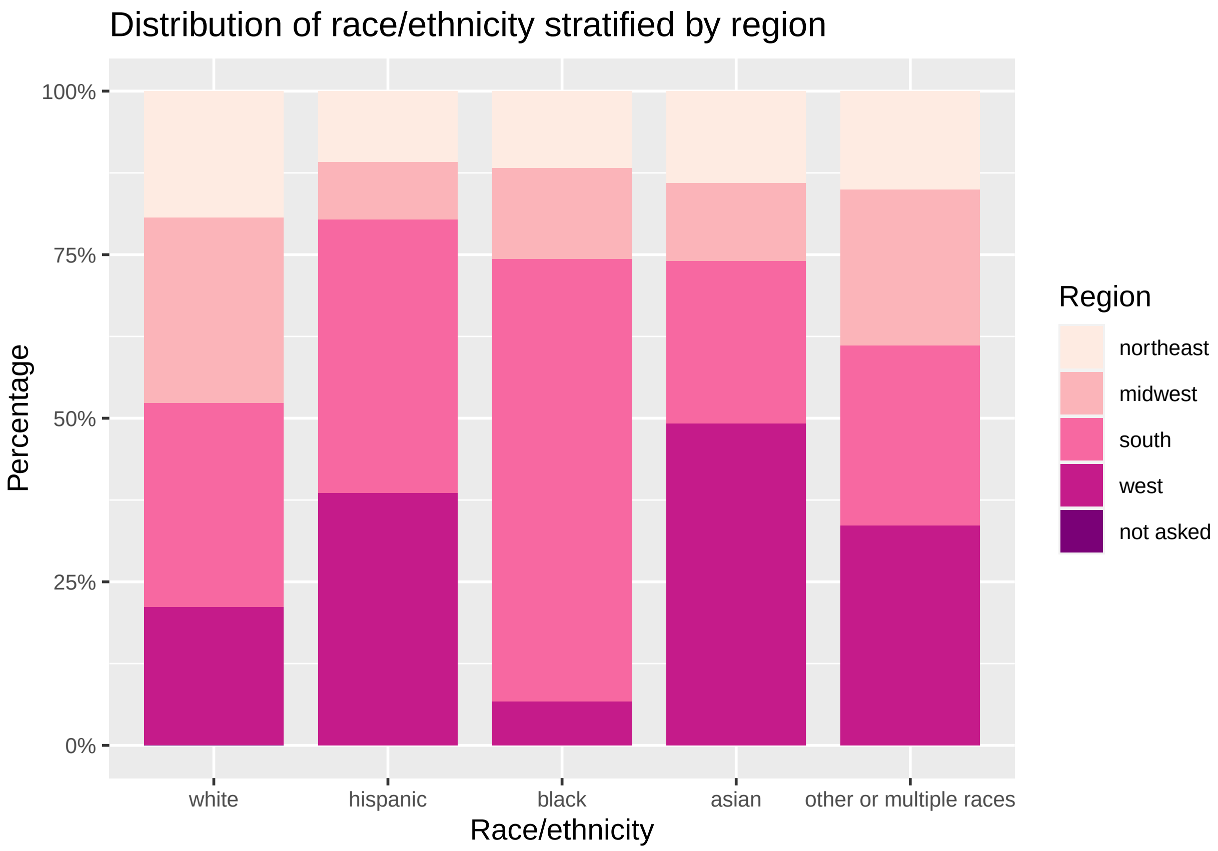
```
h209red %>%
  ggplot(aes(x = pap_f, fill = race_f)) +
  geom_bar(position = "fill", width = 0.8) +
  theme(axis.text = element_text(size = 9),
        axis.title = element_text(size = 11),
        legend.title = element_text(size = 11),
        legend.text = element_text(size = 9)) +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_brewer(palette = "RdPu") +
  labs(x = "Pap smear",
       y = "Percentage",
       fill = "Race/ethnicity") +
  ggtitle("Pap smear status in last 5 years stratified by race/ethnicity")
```



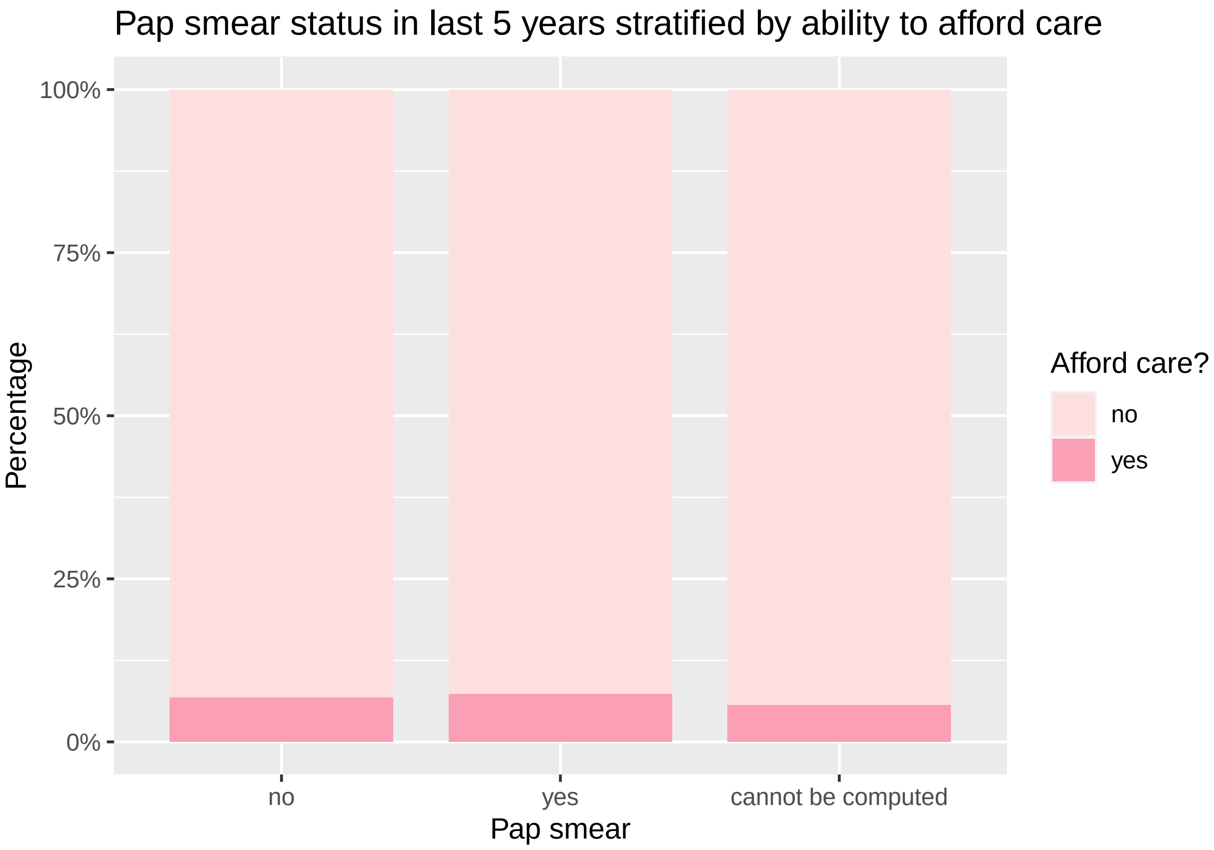
```
h209red %>%
  ggplot(aes(x = pap_f, fill = region_f)) +
  geom_bar(position = "fill", width = 0.8) +
  theme(axis.text = element_text(size = 9),
        axis.title = element_text(size = 11),
        legend.title = element_text(size = 11),
        legend.text = element_text(size = 9)) +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_brewer(palette = "RdPu") +
  labs(x = "Pap smear",
       y = "Percentage",
       fill = "Region") +
  ggtitle("Pap smear status in last 5 years stratified by region")
```



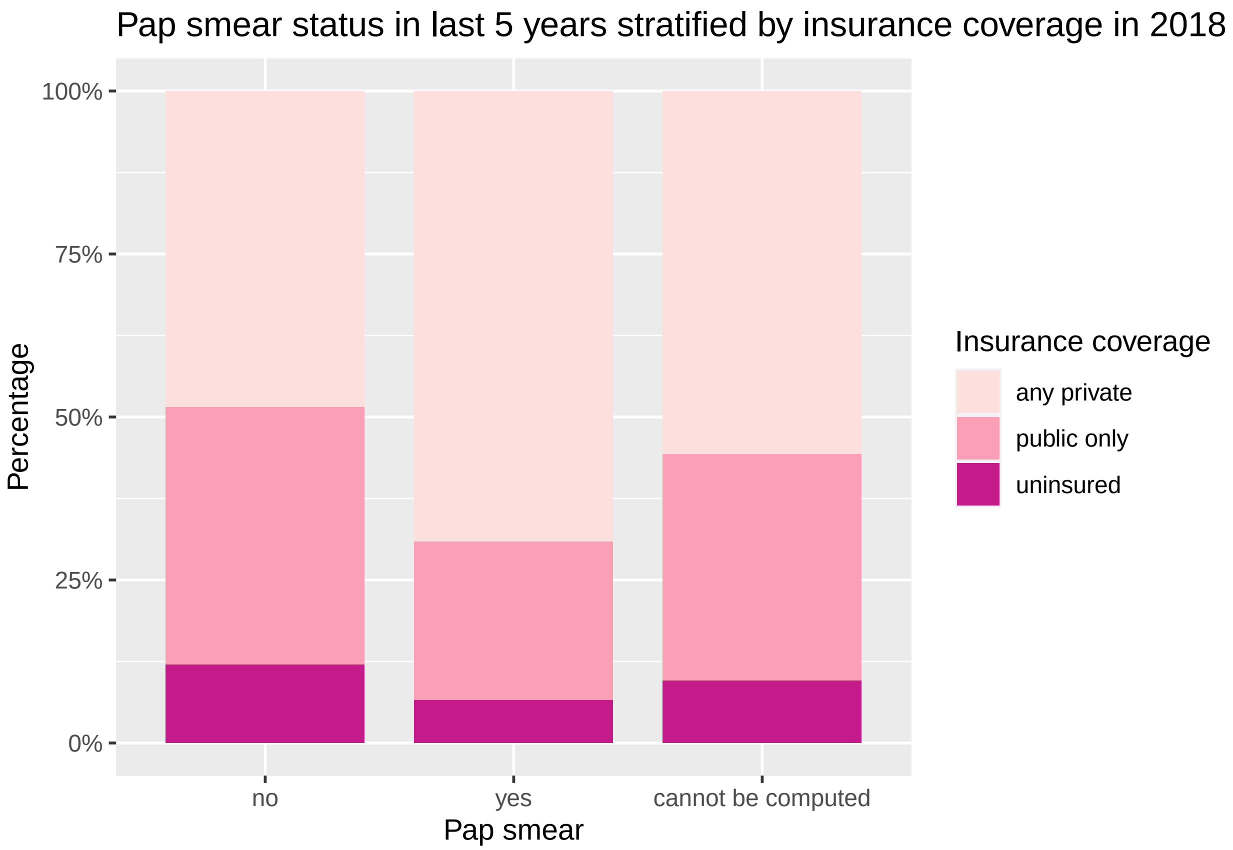
```
h209red %>%
  ggplot(aes(x = race_f, fill = region_f)) +
  geom_bar(position = "fill", width = 0.8) +
  theme(axis.text = element_text(size = 7.5),
        axis.title = element_text(size = 11),
        legend.title = element_text(size = 11),
        legend.text = element_text(size = 7.5)) +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_brewer(palette = "RdPu") +
  labs(x = "Race/ethnicity",
       y = "Percentage",
       fill = "Region") +
  ggtitle("Distribution of race/ethnicity stratified by region")
```



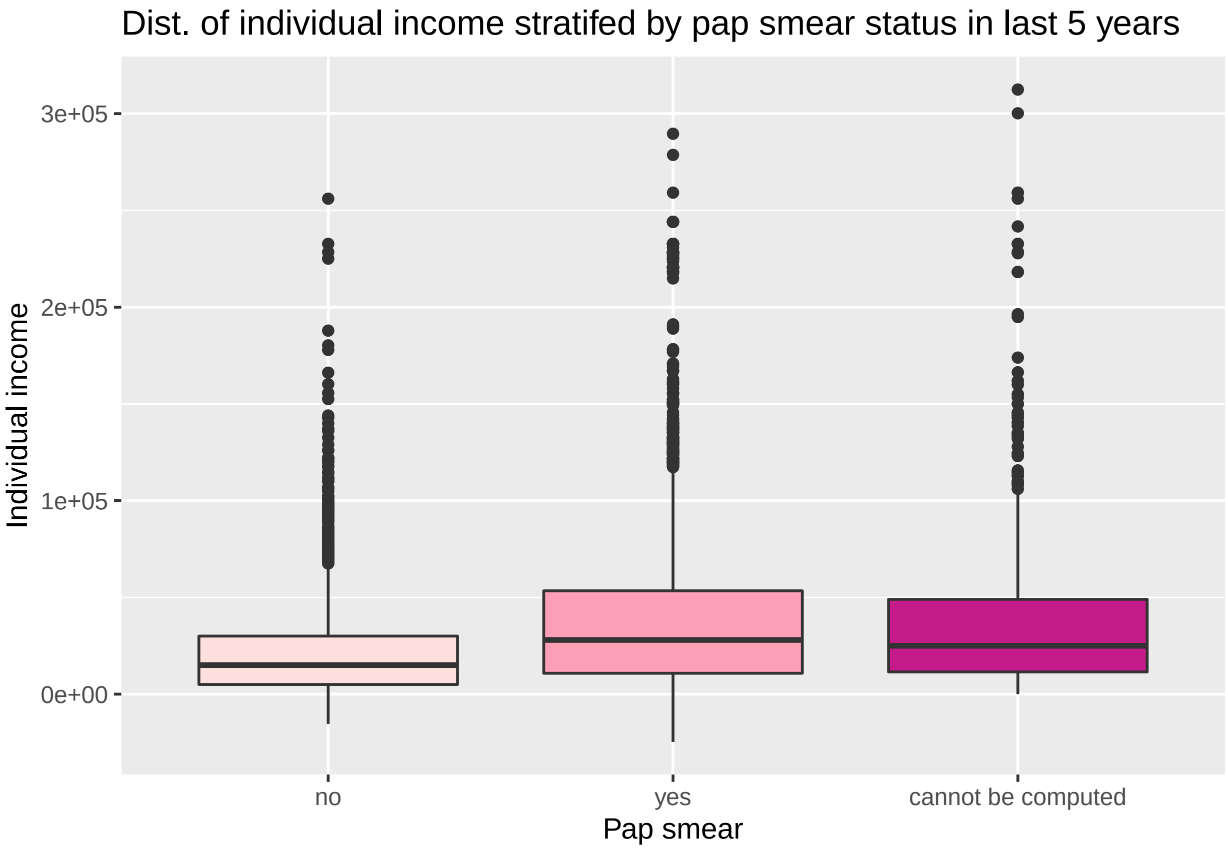
```
h209red %>%
  filter(afford_care_f != "did not answer") %>%
  ggplot(aes(x = pap_f, fill = afford_care_f)) +
  geom_bar(position = "fill", width = 0.8) +
  theme(axis.text = element_text(size = 9),
        axis.title = element_text(size = 11),
        legend.title = element_text(size = 11),
        legend.text = element_text(size = 9)) +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_brewer(palette = "RdPu") +
  labs(x = "Pap smear",
       y = "Percentage",
       fill = "Afford care?") +
  ggtitle("Pap smear status in last 5 years stratified by ability to afford care")
```



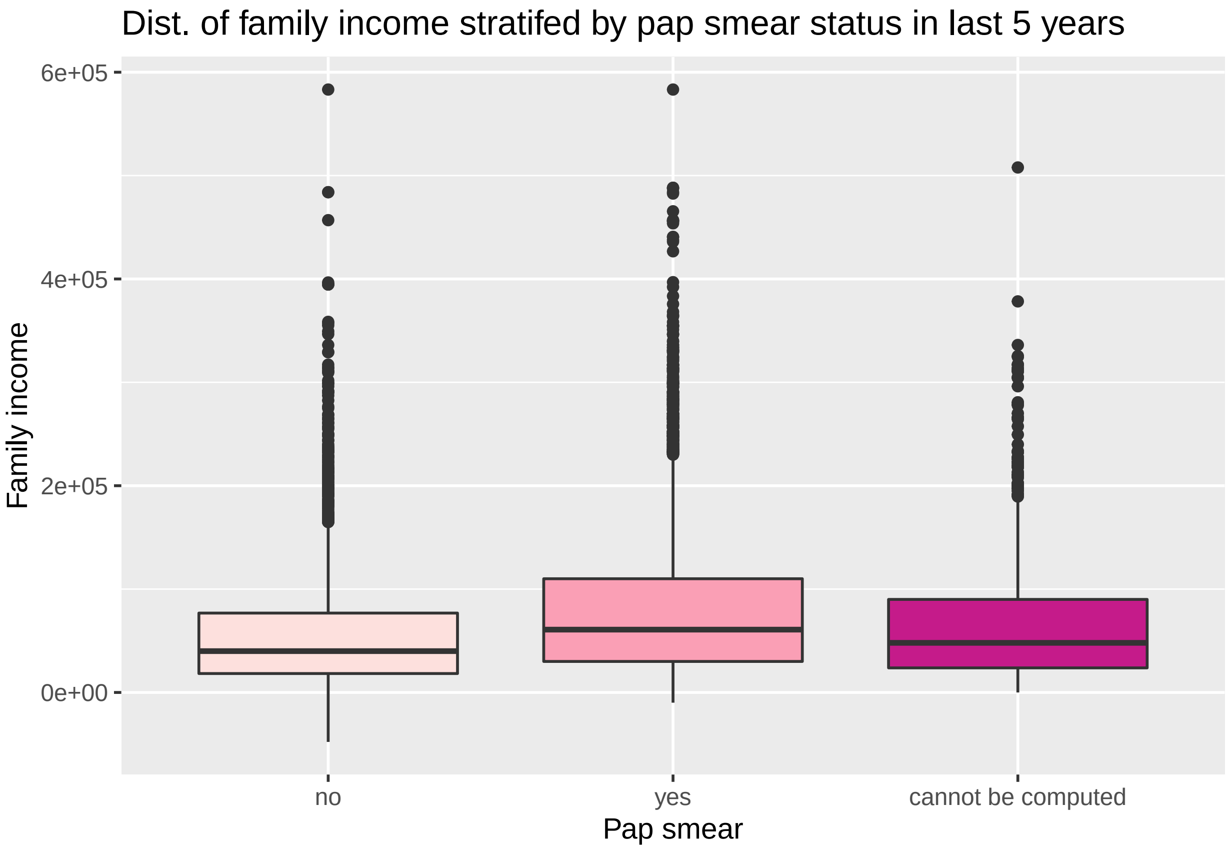
```
h209red %>%
  ggplot(aes(x = pap_f, fill = inscov_gen_2018_f)) +
  geom_bar(position = "fill", width = 0.8) +
  theme(axis.text = element_text(size = 9),
        axis.title = element_text(size = 11),
        legend.title = element_text(size = 11),
        legend.text = element_text(size = 9)) +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_brewer(palette = "RdPu") +
  labs(x = "Pap smear",
       y = "Percentage",
       fill = "Insurance coverage") +
  ggtitle("Pap smear status in last 5 years stratified by insurance coverage in 2018")
```



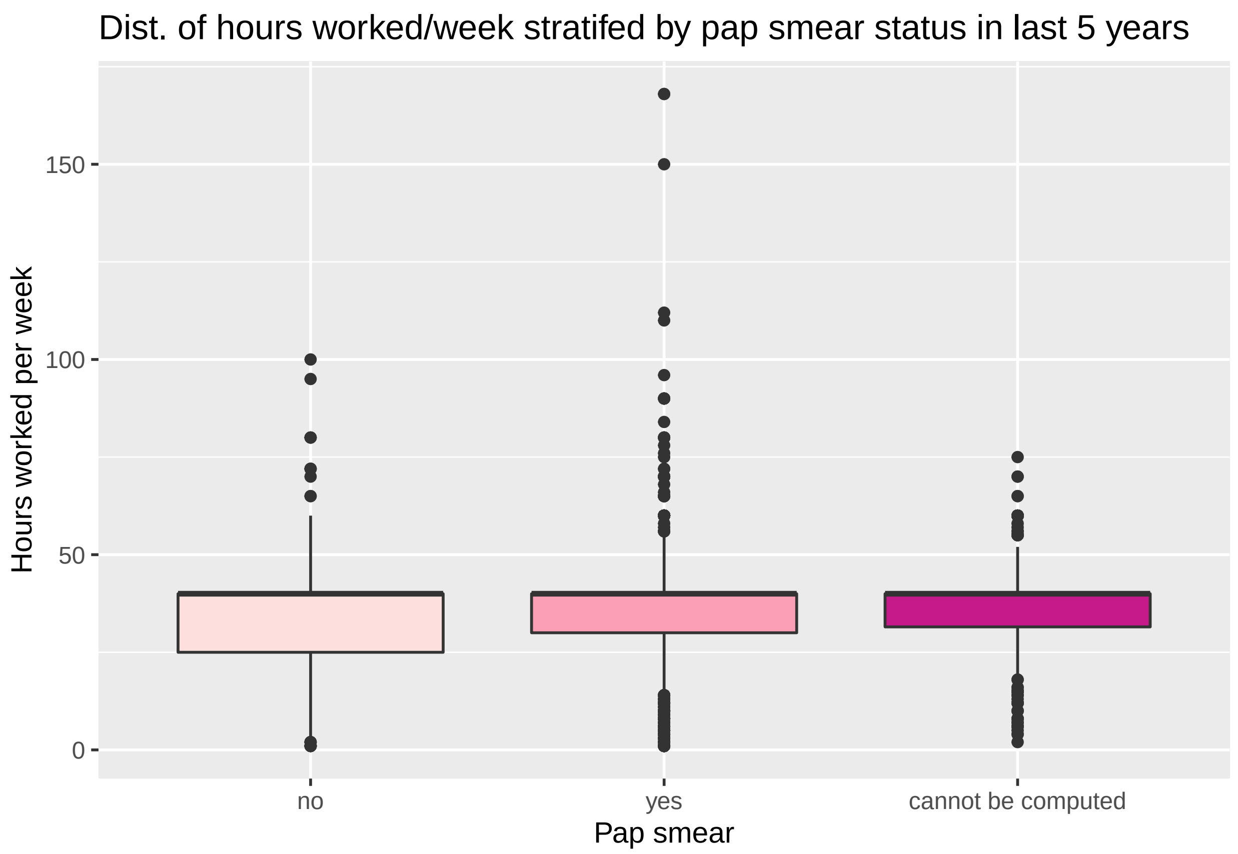
```r
h209red %>%
  ggplot(aes(x = pap_f, y = income_indiv)) +
  geom_boxplot(fill = brewer.pal(n = 3, name = "RdPu")) +
  labs(x = "Pap smear",
       y = "Individual income") +
  ggtitle("Dist. of individual income stratifed by pap smear status in last 5 years")
```



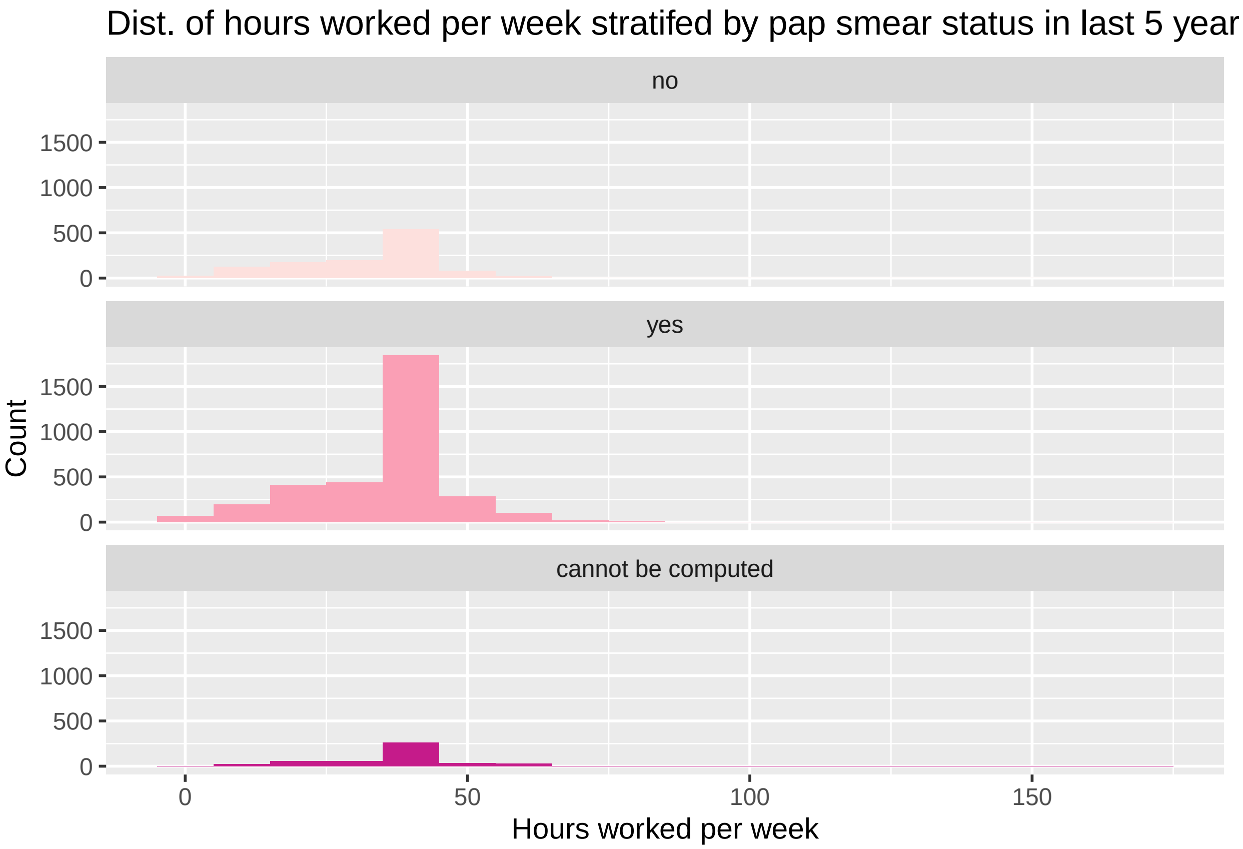
```r
h209red %>%
  ggplot(aes(x = pap_f, y = income_fam)) +
  geom_boxplot(fill = brewer.pal(n = 3, name = "RdPu")) +
  labs(x = "Pap smear",
       y = "Family income") +
  ggtitle("Dist. of family income stratifed by pap smear status in last 5 years")
```



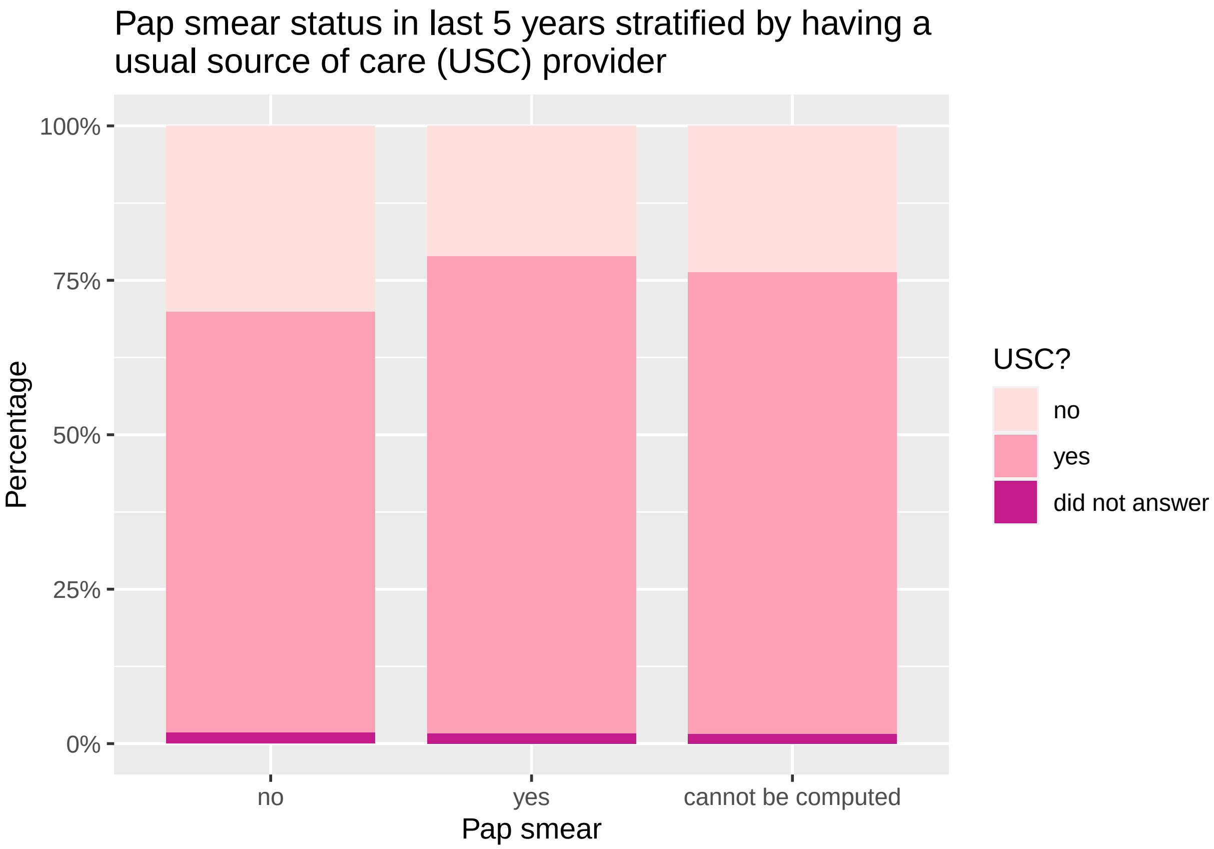
```r
h209red %>%
  filter(hrsworked_rd1 != -1) %>%
  ggplot(aes(x = pap_f, y = hrsworked_rd1)) +
  geom_boxplot(fill = brewer.pal(n = 3, name = "RdPu")) +
  labs(x = "Pap smear",
       y = "Hours worked per week") +
  ggtitle("Dist. of hours worked/week stratifed by pap smear status in last 5 years")
```



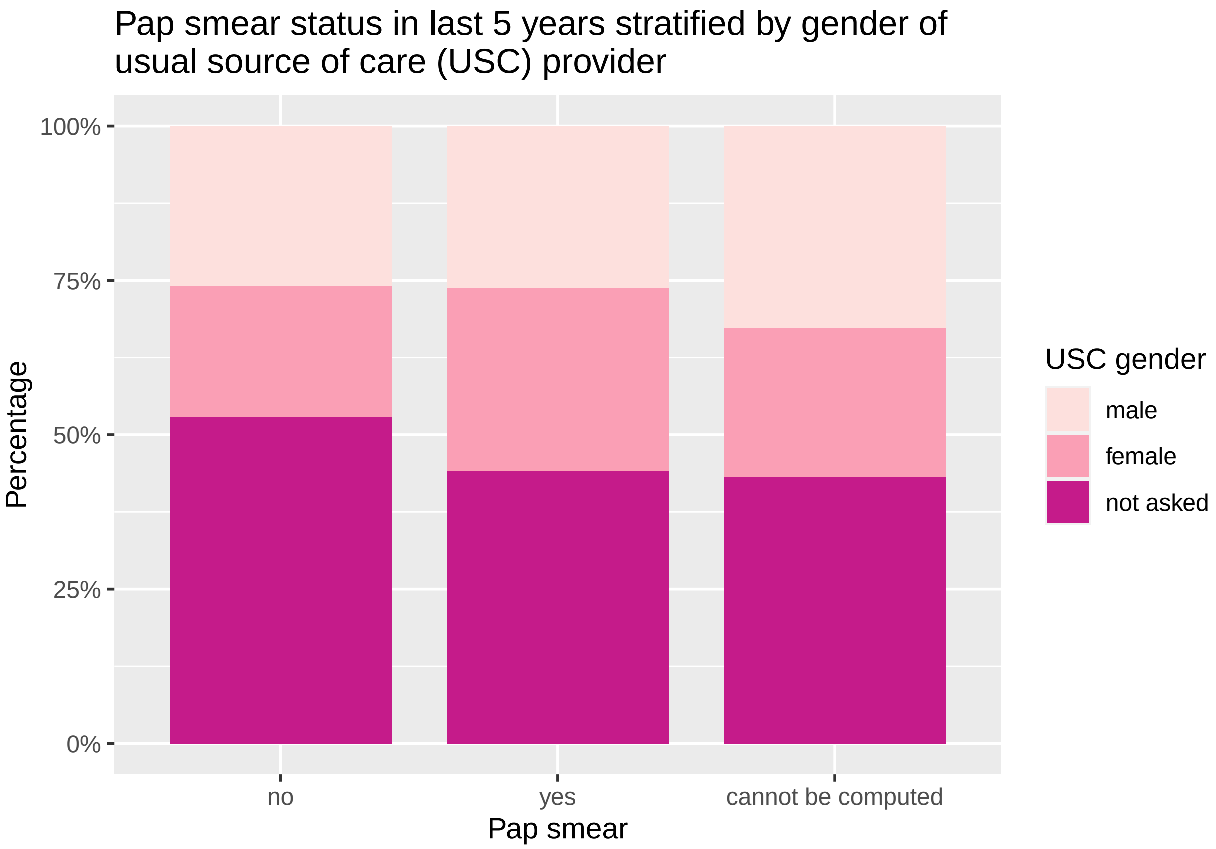
```
h209red %>%
  filter(hrsworked_rd1 != -1) %>%
  ggplot(aes(x = hrsworked_rd1, fill = pap_f)) +
  geom_histogram(binwidth = 10) +
  facet_wrap(~ pap_f, ncol = 1) +
  scale_fill_brewer(palette = "RdPu") +
  theme(legend.position = "none") +
  labs(x = "Hours worked per week",
       y = "Count") +
  ggtitle("Dist. of hours worked per week stratifed by pap smear status in last 5 years")
```



```
h209red %>%
  ggplot(aes(x = pap_f, fill = have_usc_f)) +
  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 = "Pap smear",
       y = "Percentage",
       fill = "USC?") +
  ggtitle("Pap smear status in last 5 years stratified by having a \nusual source of care (USC) provide
```



```
h209red %>%
  filter(usc_gender_f != "did not answer") %>%
  ggplot(aes(x = pap_f, fill = usc_gender_f)) +
  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 = "Pap smear",
       y = "Percentage",
       fill = "USC gender") +
  ggtitle("Pap smear status in last 5 years stratified by gender of \nusual source of care (USC) provid
```



11. Project Attestation: No member of this group is using these data or same/similar questions in any other course or course project, at HSPH.