Milestone #4

Rachael Baartmans, Lara Petalio, Christine Truong

11-14-22

```
library(tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#Insert previous code with updated datasets
race_data <-read_csv("ca_csc_outcome_race_data.csv",
            col_select = c(ID, NERVOUS, WORRYING, PROBINTR,
                           PROBDOWN, ASTHMA, HEARTDIS,
                           DIABETES, OTHMENILL, race01, race02, race03,
                           race04, race05, race06, race07, race08,
                           race09, race10, race11, race12, race13,
                           race14, race15),
            na = c("", "NA", "NA/Not Applicable", "N/A", "n/a",
                   "(DO NOT READ) NA/Not Applicable",
                   "(DO NOT READ) Refused",
                   "(DO NOT READ) Don't know"))
```

```
## Rows: 1000 Columns: 24
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (24): ID, NERVOUS, WORRYING, PROBINTR, PROBDOWN, ASTHMA, HEARTDIS, DIABE...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
smoker_data <- read_csv("ca_csc_smoker_data.csv",
            col_select = c(psraid, smokstat, WHEREBUY, BUYCALIF,
                           HOWMANY, SMOK6NUM, SMOK6UNI),
            na = c("", "NA", "NA/Not Applicable", "N/A", "n/a",
                   "(DO NOT READ) NA/Not Applicable",
                   "(DO NOT READ) Refused",
                   "(DO NOT READ) Don't know"))
```

```
## Rows: 1000 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): smokstat, HOWMANY, SMOK6UNI, BUYCALIF, WHEREBUY
## dbl (2): psraid, SMOK6NUM
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Changed casing for variables from capitals to lowercase in both dataframes
#of race_data and smoker_data
names(race_data) <- tolower(names(race_data))
names(smoker_data) <- tolower(names(smoker_data))

#Re-coded "100 or more cigarettes" to "100" for future pack-year calculations
#once the variable `howmany` is converted from character to numeric data type
smoker_data$howmany <- recode(smoker_data$howmany,
                              "100 or more cigarettes" = "100")

#Changed the data type of `howmany` from character to numeric in order to
#perform calculations for pack-years later
smoker_data$howmany <- as.numeric(smoker_data$howmany)

#Re-coded "In military commissaries, or" to "In military commissaries", as well
#as "Somewhere else (SPECIFY)?" to "Somewhere else" to make
#response option more understandable when displayed for the variable `wherebuy`.
smoker_data$wherebuy <- recode(smoker_data$wherebuy,
                  "In military commissaries, or" = "In military commissaries",
                  "Somewhere else (SPECIFY)?" = "Somewhere else")

#Filtered the value of "Years" from the variable `smok6uni` so that "Years"
#would be the only unique value assigned to `smok6uni`. This is because we only
#need the time unit of "Years" for calculating "pack-years" later to describe
#tobacco consumption. This filtered subset was assigned to a new data frame
#called smoker_data_2.
smoker_data_2 <- smoker_data %>% filter(smok6uni == "Years")



#Created new variable `race` to combine variables race01:race15 into one column.
race_data_2 <- race_data %>%
  mutate(race = case_when(race01 == "Yes" ~ "White",
         race02 == "Yes" ~ "Black",
         race03 == "Yes" ~ "Japanese",
         race04 == "Yes" ~ "Chinese",
```

```
        race05 == "Yes" ~ "Filipino",
        race06 == "Yes" ~ "Korean",
        race07 == "Yes" ~ "Other Asian or Pacific Islander",
        race08 == "Yes" ~ "American Indian or Alaskan Native",
        race09 == "Yes" ~ "Mexican",
        race10 == "Yes" ~ "Hispanic/Latino",
        race11 == "Yes" ~ "Other",
        race12 == "Yes" ~ "Vietnamese",
        race13 == "Yes" ~ "Asian Indian",
        race14 == "Yes" ~ "Refused",
        race15 == "Yes" ~ "Don't know")) %>%
  select(-(race01:race15))

#Created new variable "packs_per_day" for future calculations for pack-years
smoker_data_3 <- smoker_data_2 %>% mutate(packs_per_day = howmany/20)


smoker_data_3_no_na <- smoker_data_3 %>%
  drop_na(wherebuy)
```

Milestone #4 assignments

```
#We calculated pack-years, which is given by the formula of
#pack-years = # of packs of cigarettes smoked per day * years a person has smoked.
#This calculation led to the creation of a new variable called `pack_years`.
#We assigned this change in the smoker_data_3_no_na data frame to a new data
#frame called smoker_data_4, which includes the new variable `pack_years`.

smoker_data_4 <- smoker_data_3_no_na %>%
  mutate(pack_years = packs_per_day*smok6num)

#Afterwards, we viewed our new data frame with the new variable `pack_years`.
smoker_data_4
```

```
## # A tibble: 753 x 9
##    psraid smokstat     wherebuy buycalif howmany smok6num smok6uni packs_per_day
##     <dbl> <chr>        <chr>    <chr>      <dbl>    <dbl> <chr>            <dbl>
##  1 100099 Current dai~ At othe~ In Cali~      30       36 Years             1.5
##  2 100109 Current dai~ At toba~ In Cali~      20       25 Years             1
##  3 100191 Current dai~ At conv~ In Cali~      15       20 Years             0.75
##  4 100206 Current dai~ At conv~ In Cali~      15        7 Years             0.75
##  5 100232 Current dai~ At liqu~ In Cali~      20       45 Years             1
##  6 100262 Current dai~ At othe~ In Cali~      15       19 Years             0.75
##  7 100317 Current dai~ At conv~ In Cali~       7        2 Years             0.35
##  8 100319 Current dai~ At toba~ In Cali~      20       15 Years             1
##  9 100351 Current dai~ In mili~ In Cali~      10       40 Years             0.5
## 10 100411 Current dai~ At supe~ In Cali~       5        4 Years             0.25
## # ... with 743 more rows, and 1 more variable: pack_years <dbl>
```

```
#In order to join our two data sets together, we first had to remove the strings
#of 'DIS' and 'STAT' from the `id` column of race_data_2 by using gsub().
#We overwrote these changes in the race_data_2 data frame and viewed these new
#changes to make sure the `id` variable only contains numbers and no
#characters.

race_data_2$id <- gsub('[DISSTAT]', '', race_data_2$id)
race_data_2
```

```
## # A tibble: 1,000 x 10
##    id     nervous  worrying probintr probdown asthma heartdis diabetes othmenill
##    <chr>  <chr>    <chr>    <chr>    <chr>    <chr>  <chr>    <chr>    <chr>
##  1 100099 Nearly ~ Not at ~ Nearly ~ Not at ~ No     Yes      No       No
##  2 100109 Several~ Several~ Several~ Not at ~ No     No       No       No
##  3 100121 Not at ~ Not at ~ Not at ~ Not at ~ No     No       No       No
##  4 100191 Several~ Not at ~ Not at ~ Not at ~ Yes    No       No       No
##  5 100206 Not at ~ Several~ Not at ~ Not at ~ No     No       No       No
##  6 100232 Not at ~ Not at ~ Not at ~ Not at ~ No     Yes      No       No
##  7 100256 Not at ~ Not at ~ Not at ~ Several~ Yes    Yes      No       No
##  8 100262 Several~ Nearly ~ Several~ Several~ No     No       No       No
##  9 100317 Several~ Several~ Several~ <NA>     No     No       No       Yes
## 10 100319 More th~ Several~ Not at ~ Not at ~ No     No       No       No
## # ... with 990 more rows, and 1 more variable: race <chr>
```

```
#Next, looking at the smoker_data_4 data frame, we see that the `psraid`
#variable contains each study participant's unique ID number, but the variable
#is a numeric data type. On the other hand, `id` from the race_data_2
#data frame is a character data type. We needed to convert `psraid` then from
#character to numeric data type because 1) `psraid` is an identifier rather than
#a numeric value to mathematically manipulate even if it does contain numbers
#and 2) in order to perform a join, the two variables must be the same data
#type.

smoker_data_4$psraid <- as.character(smoker_data_4$psraid)

#Afterward, we performed an inner join between race_data_2 and
#smoker_data_4 by each study participant's unique ID number, which is
#represented by `id` in race_data_2 and `psraid` in smoker_data_4. We
#chose to do an inner join because we wanted to select participants that
#exist in each of our two data sets for our final data frame.
#We assigned this join to a new data frame called joined_smoking_df.

joined_smoking_df <- inner_join(x = race_data_2, y = smoker_data_4,
                                by=c("id" = "psraid"))

#Then, we viewed the new data frame we created
joined_smoking_df
```

```
## # A tibble: 753 x 18
##    id     nervous  worrying probintr probdown asthma heartdis diabetes othmenill
##    <chr>  <chr>    <chr>    <chr>    <chr>    <chr>  <chr>    <chr>    <chr>
##  1 100099 Nearly ~ Not at ~ Nearly ~ Not at ~ No     Yes      No       No
##  2 100109 Several~ Several~ Several~ Not at ~ No     No       No       No
```

```
##  3 100191 Several~ Not at ~ Not at ~ Not at ~ Yes    No       No       No
##  4 100206 Not at ~ Several~ Not at ~ Not at ~ No     No       No       No
##  5 100232 Not at ~ Not at ~ Not at ~ Not at ~ No     Yes      No       No
##  6 100262 Several~ Nearly ~ Several~ Several~ No     No       No       No
##  7 100317 Several~ Several~ Several~ <NA>     No     No       No       Yes
##  8 100319 More th~ Several~ Not at ~ Not at ~ No     No       No       No
##  9 100351 Several~ Several~ Several~ Several~ No     No       No       Yes
## 10 100411 Nearly ~ Several~ More th~ Not at ~ No     No       No       No
## # ... with 743 more rows, and 9 more variables: race <chr>, smokstat <chr>,
## #   wherebuy <chr>, buycalif <chr>, howmany <dbl>, smok6num <dbl>,
## #   smok6uni <chr>, packs_per_day <dbl>, pack_years <dbl>
```

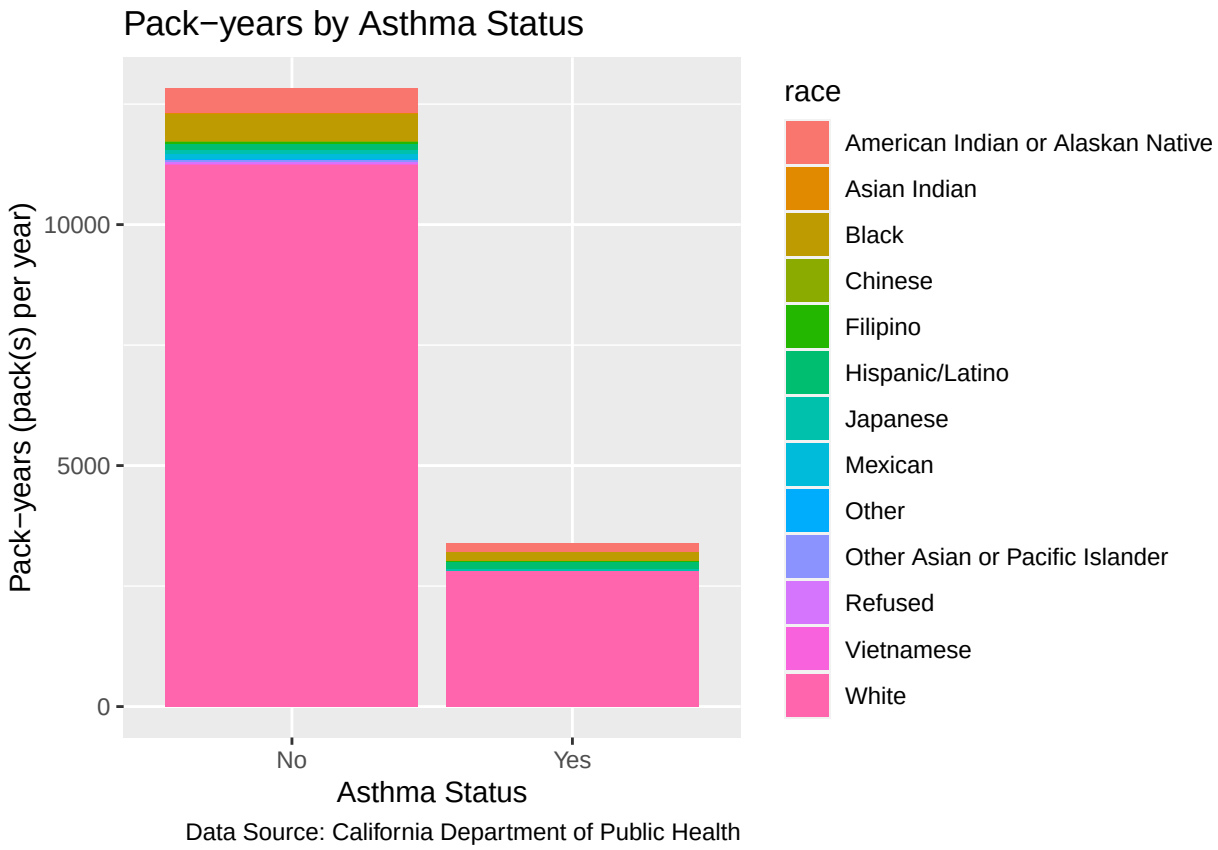
##OUR GROUP'S RESEARCH QUESTION: "For this project, we aim to investigate how tobacco use primarily impacts mental illness among smokers in California in 2011, as well as explore how race and location of cigarette purchase can impact disease status."

Visualizations (3 total) **one print quality tables per scenario** With Kable: pack-years vs. asthma/heart disease/diabetes/mental illness (1 table)

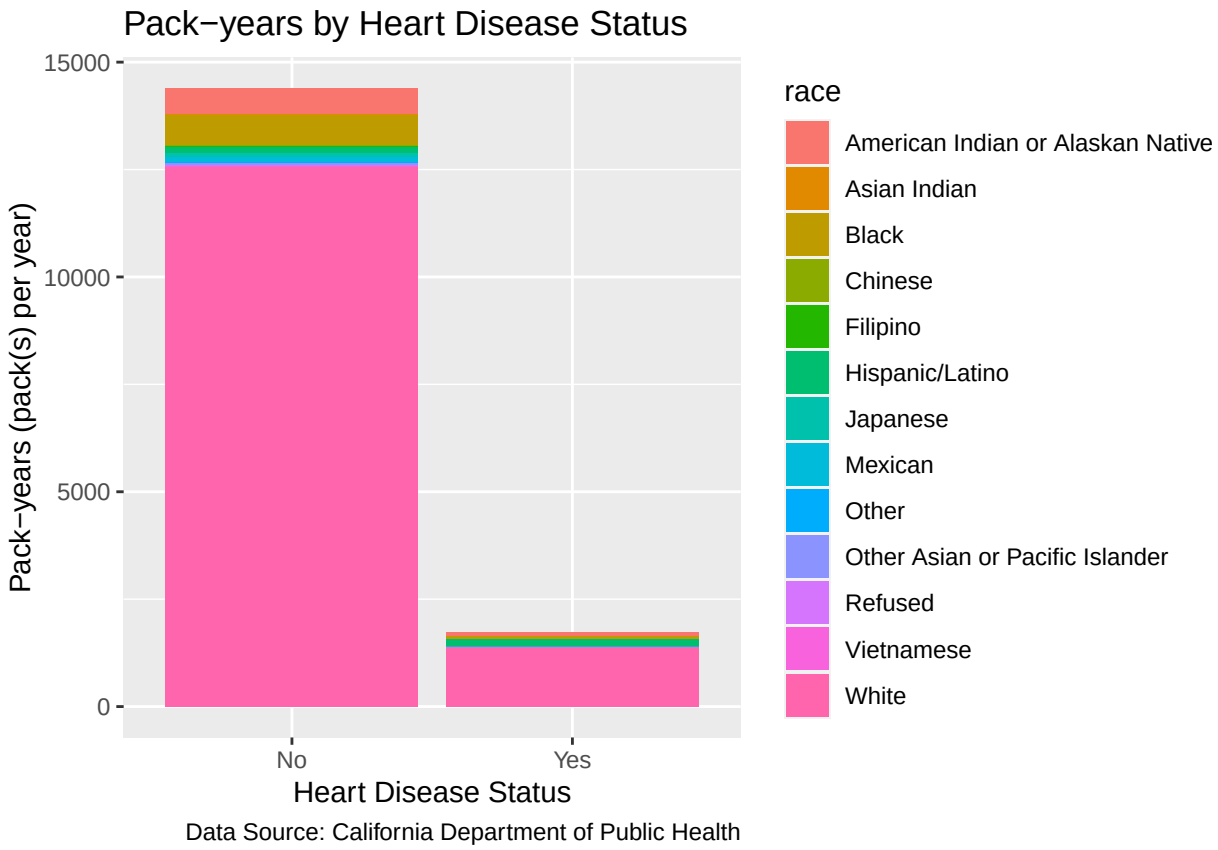
"Compare the average number of pack-years by at least four disease outcomes (e.g. asthma, heart disease, diabetes, physical illness, and/or mental illness). Provide a print-quality table that shows the average number of pack-years and the disease outcomes."

one print quality plot or chart per scenario With ggplot: 1 bar graph (x = disease, y = pack-years)

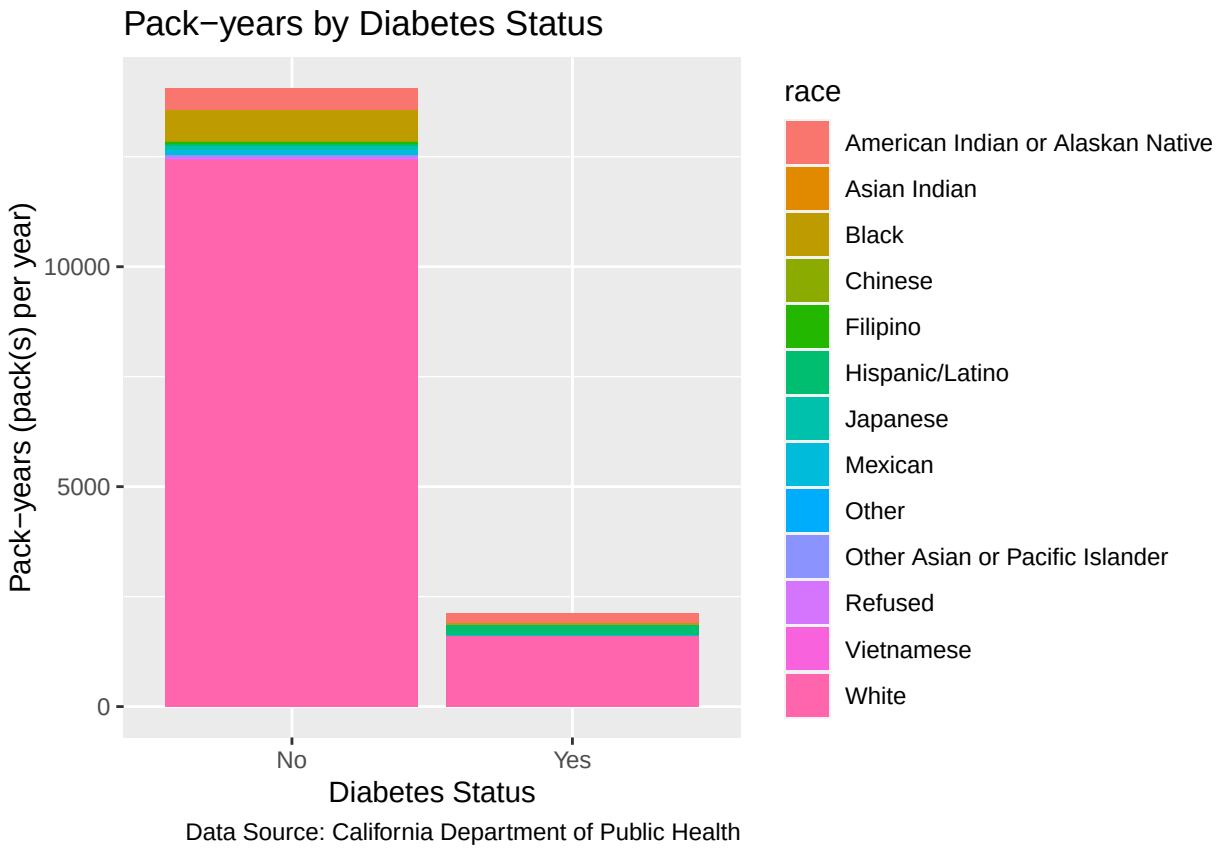
```
joined_smoking_df %>%
  drop_na(c(asthma, pack_years)) %>%
  ggplot(aes(x = asthma, pack_years)) +
  geom_bar(aes(fill=race), stat="identity") +
  labs(x = "Asthma Status", y = "Pack-years (pack(s) per year)",
       title = "Pack-years by Asthma Status",
       caption = "Data Source: California Department of Public Health")
```



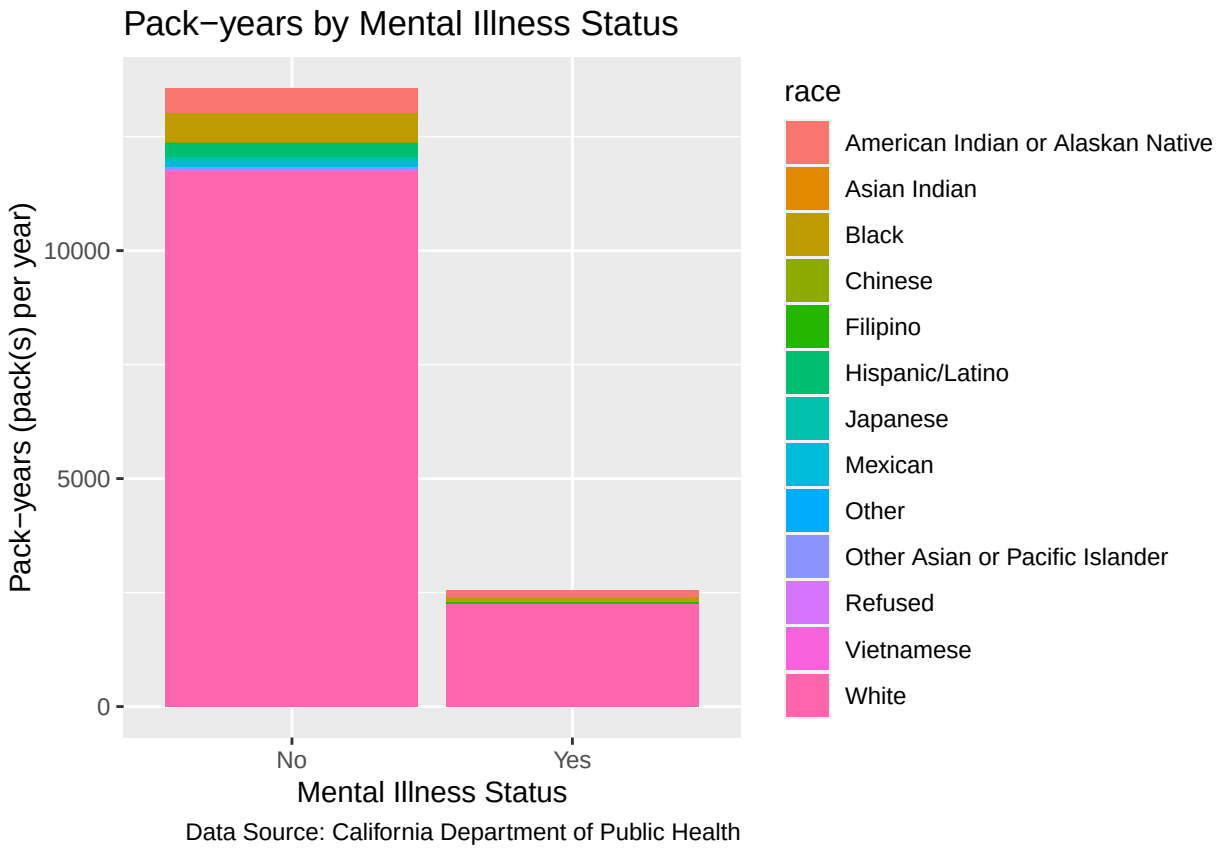

```
joined_smoking_df %>%
  drop_na(c(heartdis, pack_years)) %>%
  ggplot(aes(x = heartdis, pack_years)) +
  geom_bar(aes(fill=race), stat="identity") +
  labs(x = "Heart Disease Status", y = "Pack-years (pack(s) per year)",
       title = "Pack-years by Heart Disease Status",
       caption = "Data Source: California Department of Public Health")
```




```
joined_smoking_df %>%
  drop_na(c(diabetes, pack_years)) %>%
  ggplot(aes(x = diabetes, pack_years)) +
  geom_bar(aes(fill=race), stat="identity") +
  labs(x = "Diabetes Status", y = "Pack-years (pack(s) per year)",
       title = "Pack-years by Diabetes Status",
       caption = "Data Source: California Department of Public Health")
```




```
joined_smoking_df %>%
  drop_na(c(othmenill, pack_years)) %>%
  ggplot(aes(x = othmenill, pack_years)) +
  geom_bar(aes(fill=race), stat="identity") +
  labs(x = "Mental Illness Status", y = "Pack-years (pack(s) per year)",
       title = "Pack-years by Mental Illness Status",
       caption = "Data Source: California Department of Public Health")
```



one additional table or plot With ggplot: dodged bar chart with (x = cigarette purchase location, aes(fill = othmenill) to view cigarette purchase location and mental illness status

Each visual should include:

code legend (if necessary) Unless we decide to input a third variable in a graph.

interpretation (1 to 2 sentences)

##PDF should be prepared for presentation **each part of milestone on new page only necessary info outputted show work with "echo"**