Milestone #4

Rachael Baartmans, Lara Petalio, Christine Truong

11-14-22

##Relevant Previous Importing and Data-cleaning Steps

```
## 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()

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows

## 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.
## 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.
```

Calculating Pack-years

```
#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`.
#`pack_years` was created conditionally based on the three different time units
#as determined by the existing variable `smok6uni`, which are: "Days", "Months",
#and "Years". To change the unit of "Days" to years, we set up a conditional
#statement in the code to divide `smok6num` by 365 before multiplying the result
#by `packs_per_day` to get pack-years. Similarly, to change the unit of "Months"
#to years, we set up a conditional statement in the code to divide `smok6num`
#by 12 before multiplying the result by `packs_per_day` to get pack-years. For
#`smok6uni` observations that have the value "Years", we just multiplied
#`smok6num` by `packs_per_day` to get pack-years directly. We assigned
#this overall change in the data frame smoker_data_2 to a new data frame
#called smoker_data_3, which includes the new variable `pack_years`.

smoker_data_3 <- smoker_data_2 %>%
  mutate(pack_years =
           case_when(smok6uni == "Days" ~ packs_per_day*(smok6num/365),
                     smok6uni == "Months" ~ packs_per_day*(smok6num/12),
                     smok6uni == "Years" ~ packs_per_day*(smok6num)))

#We then rounded `pack_years` to the nearest whole number for all observations.

smoker_data_3$pack_years <- round(smoker_data_3$pack_years, 0)

#Afterwards, we viewed our new data frame, which includes our new variable
#`pack_years`.

smoker_data_3
```

```
## # A tibble: 1,000 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 100121 Current non~ At conv~ In Cali~       1       NA <NA>               0.05
##  4 100191 Current dai~ At conv~ In Cali~      15       20 Years              0.75
##  5 100206 Current dai~ At conv~ In Cali~      15        7 Years              0.75
##  6 100232 Current dai~ At liqu~ In Cali~      20       45 Years              1
##  7 100256 Current non~ At conv~ Somewhe~       3       NA <NA>               0.15
##  8 100262 Current dai~ At othe~ In Cali~      15       19 Years              0.75
##  9 100317 Current dai~ At conv~ In Cali~       7        2 Years              0.35
## 10 100319 Current dai~ At toba~ In Cali~      20       15 Years              1
## # ... with 990 more rows, and 1 more variable: pack_years <dbl>
```

Joining Cleaned Data Sets Together

```
#In order to join our two cleaned 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_3 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_3$psraid <- as.character(smoker_data_3$psraid)

#Afterward, we performed an inner join between race_data_2 and
#smoker_data_3 by each study participant's unique ID number, which is
#represented by `id` in race_data_2 and `psraid` in smoker_data_3. 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_3,
                                by=c("id" = "psraid"))

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

```
## # A tibble: 1,000 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 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 9 more variables: race <chr>, smokstat <chr>,
## #   wherebuy <chr>, buycalif <chr>, howmany <dbl>, smok6num <dbl>,
## #   smok6uni <chr>, packs_per_day <dbl>, pack_years <dbl>
```

Determining Pack-years for Each Disease by Race

```r
#We created subsets of the data frame joined_smoking_df for each disease of
#interest, with each subset including only the variable of `race` and average
#values of the variable `pack_years` according to disease. This is to simplify
#the process of creating some graphs later by including only the information
#we need in a data frame.

#asthma
avg_pack_years_race_asthma <- joined_smoking_df %>%
  filter(!is.na(pack_years)) %>%
  group_by(race, asthma) %>%
  summarize(avg_pack_years = sum(pack_years)/n())
```

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

```r
#heartdis
avg_pack_years_race_heartdis <- joined_smoking_df %>%
  filter(!is.na(pack_years)) %>%
  group_by(race, heartdis) %>%
  summarize(avg_pack_years = sum(pack_years)/n())
```

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

```r
#diabetes
avg_pack_years_race_diabetes <- joined_smoking_df %>%
  filter(!is.na(pack_years)) %>%
  group_by(race, diabetes) %>%
  summarize(avg_pack_years = sum(pack_years)/n())
```

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

```r
#othmenill
avg_pack_years_race_othmenill <- joined_smoking_df %>%
  filter(!is.na(pack_years)) %>%
  group_by(race, othmenill) %>%
  summarize(avg_pack_years = sum(pack_years)/n())
```

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

Visualizations

Average Pack-years by Disease Status (Two Complementary Tables)

```r
#Table for avg pack-years per disease for smokers who have a disease
t_avg_pack_years_disease <- joined_smoking_df %>%
  mutate(disease = case_when(asthma == "Yes" ~ "Asthma",
                             heartdis == "Yes" ~ "Heart Disease",
                             diabetes == "Yes" ~ "Diabetes",
                             othmenill == "Yes" ~ "Mental Illness")) %>%
  select(disease, pack_years) %>%
  filter(!is.na(pack_years), !is.na(disease)) %>%
  group_by(disease) %>%
  summarize(avg_pack_years = round(sum(pack_years)/n(), 0))
t_avg_pack_years_disease
```

```
## # A tibble: 4 x 2
##   disease        avg_pack_years
##   <chr>                   <dbl>
## 1 Asthma                     25
## 2 Diabetes                   25
## 3 Heart Disease              28
## 4 Mental Illness             17
```

```r
#Kable table for avg pack-years per disease for smokers who have a disease
kable(t_avg_pack_years_disease,
      booktabs=T,
      col.names=c("Disease", "Average Pack-years"),
      align='lcccc',
      caption= 'Average Pack-years for Smokers Who Have a Disease') %>%
      kable_styling(full_width = T) %>%
      kable_styling(latex_options = "hold_position")
```

Table 1: Average Pack-years for Smokers Who Have a Disease

Disease	Average Pack-years
Asthma	25
Diabetes	25
Heart Disease	28
Mental Illness	17

```r
#Table for avg pack-years per disease for those who do not have a disease
t_avg_pack_years_no_disease <- joined_smoking_df %>%
  mutate(disease = case_when(asthma == "No" ~ "Does Not Have Asthma",
                             heartdis == "No" ~ "Does Not Have Heart Disease",
                             diabetes == "No" ~ "Does Not Have Diabetes",
                             othmenill == "No" ~ "Does Not Have Mental Illness")) %>%
  select(disease, pack_years) %>%
  filter(!is.na(pack_years), !is.na(disease)) %>%
  group_by(disease) %>%
  summarize(avg_pack_years = round(sum(pack_years)/n(), 0))
t_avg_pack_years_no_disease
```

```
## # A tibble: 4 x 2
##   disease                 avg_pack_years
##   <chr>                            <dbl>
## 1 Does Not Have Asthma                21
## 2 Does Not Have Diabetes              28
## 3 Does Not Have Heart Disease         23
## 4 Does Not Have Mental Illness       102
```

```r
#Kable table for avg pack-years per disease for those who do not have a disease
kable(t_avg_pack_years_no_disease,
booktabs=T,
col.names=c("Disease", "Average Pack-years"),
align='lcccc',
caption= 'Average Pack-years Among Smokers Responding "No" to Having a Disease') %>%
      kable_styling(full_width = T) %>%
      kable_styling(latex_options = "hold_position")
```

Table 2: Average Pack-years Among Smokers Responding "No" to Having a Disease

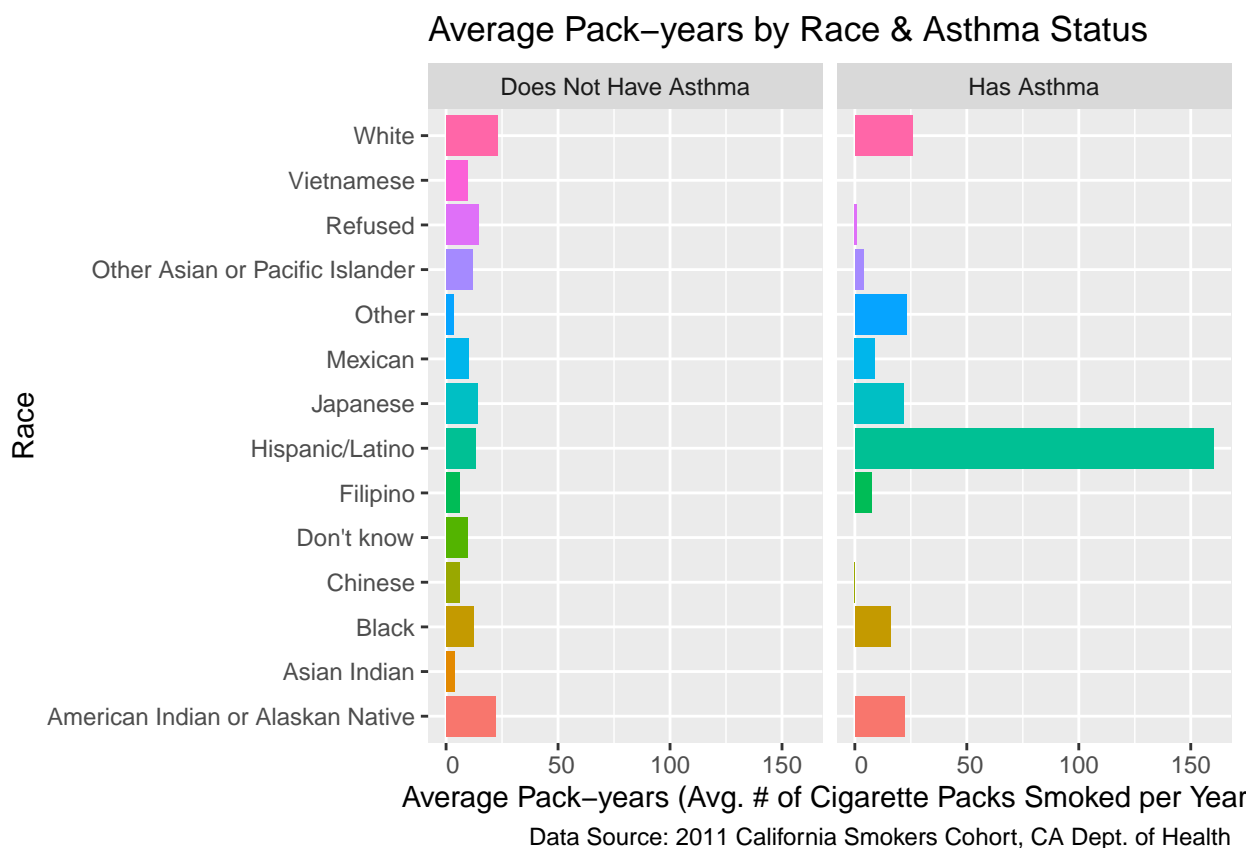| Disease	Average Pack-years
Does Not Have Asthma	21
Does Not Have Diabetes	28
Does Not Have Heart Disease	23
Does Not Have Mental Illness	102

#We created another table with all diseases present with yes/no statuses, #showing unique combinations of diseases and the avg pack-years per combination #table_option_2 <- joined_smoking_df %>% #filter(!is.na(pack_years)) %>% #group_by(asthma, heartdis, diabetes, othmenill) %>% #summarize(avg_pack_years = sum(pack_years)/n()) #table_option_2

Interpretation of the Three Tables of Average Pack-years by Disease Status:
Among smokers that have reported the presence of the diseases of interest (i.e., asthma, heart disease, diabetes, and mental illness), those with heart disease have the highest number of average pack-years (28), while those with mental illness have the lowest number of pack years (17). Among smokers that reported "no" to having a disease, those with no mental illness have the highest average pack years (102), while those with no asthma had the lowest average pack years (21).

Average Pack-years by Race and Disease (Series of Four Graphs)
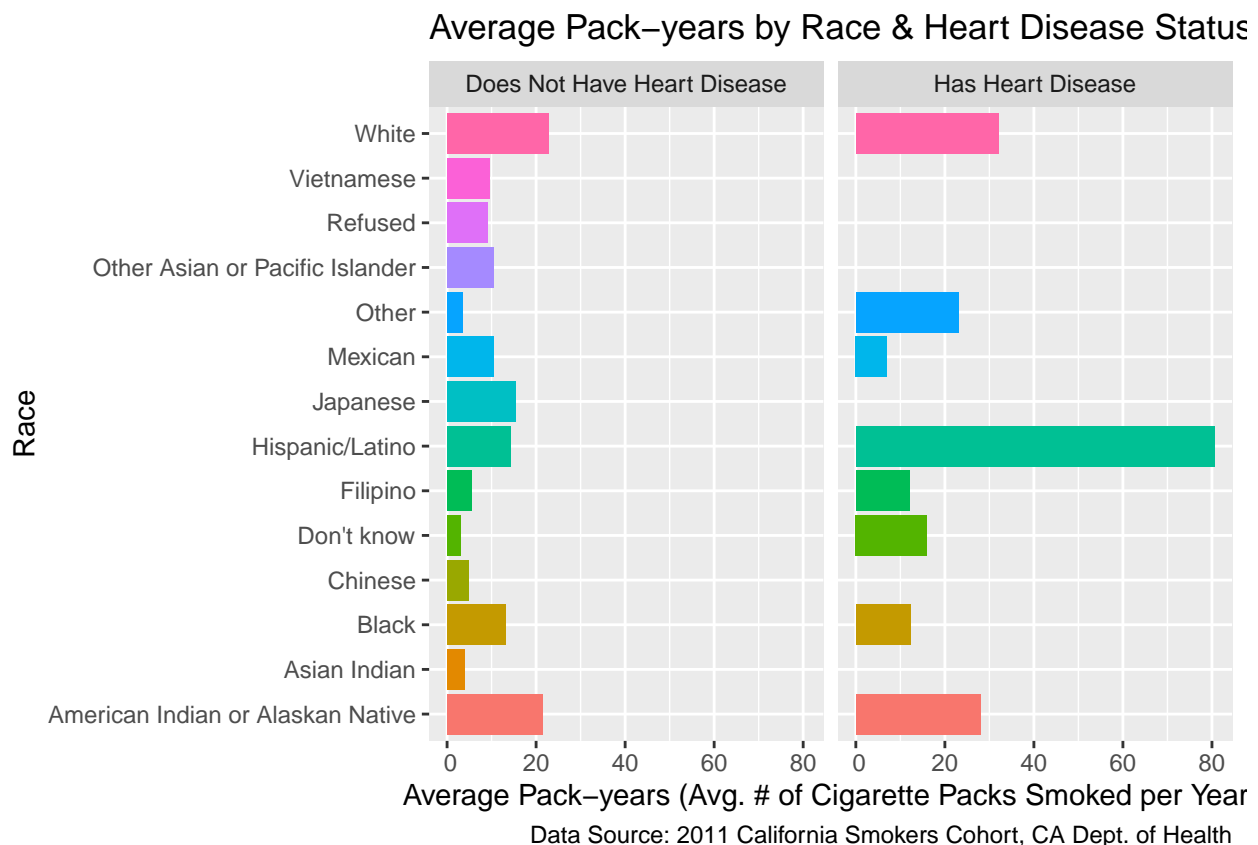
```
avg_pack_years_race_asthma %>%
  drop_na(c(asthma, avg_pack_years)) %>%
  ggplot(aes(x = race, y = avg_pack_years)) +
  geom_bar(aes(fill=race), stat="identity", position = "dodge") +
  coord_flip() +
  guides(fill = "none") +
  labs(x = "Race",
       y = "Average Pack-years (Avg. # of Cigarette Packs Smoked per Year)",
  title = "Average Pack-years by Race & Asthma Status",
  caption = "Data Source: 2011 California Smokers Cohort, CA Dept. of Health") +
  scale_y_continuous(labels = function(x) format(x,big.mark=",",
                                                 scientific=FALSE)) +
  facet_wrap(~ asthma, labeller = labeller(asthma =
                                    c( "Yes" = "Has Asthma","No" = "Does Not Have Asthma"
                                    )))
```

```
avg_pack_years_race_heartdis %>%
  drop_na(c(heartdis, avg_pack_years)) %>%
  ggplot(aes(x = race, y = avg_pack_years)) +
  geom_bar(aes(fill=race), stat="identity", position = "dodge") +
  coord_flip() +
  guides(fill = "none") +
  labs(x = "Race",
       y = "Average Pack-years (Avg. # of Cigarette Packs Smoked per Year)",
  title = "Average Pack-years by Race & Heart Disease Status",
  caption = "Data Source: 2011 California Smokers Cohort, CA Dept. of Health") +
  scale_y_continuous(labels = function(x) format(x,big.mark=",",
                                           scientific=FALSE)) +
  facet_wrap(~ heartdis, labeller = labeller(heartdis =
                                 c("No" = "Does Not Have Heart Disease",
                                   "Yes" = "Has Heart Disease")))
```
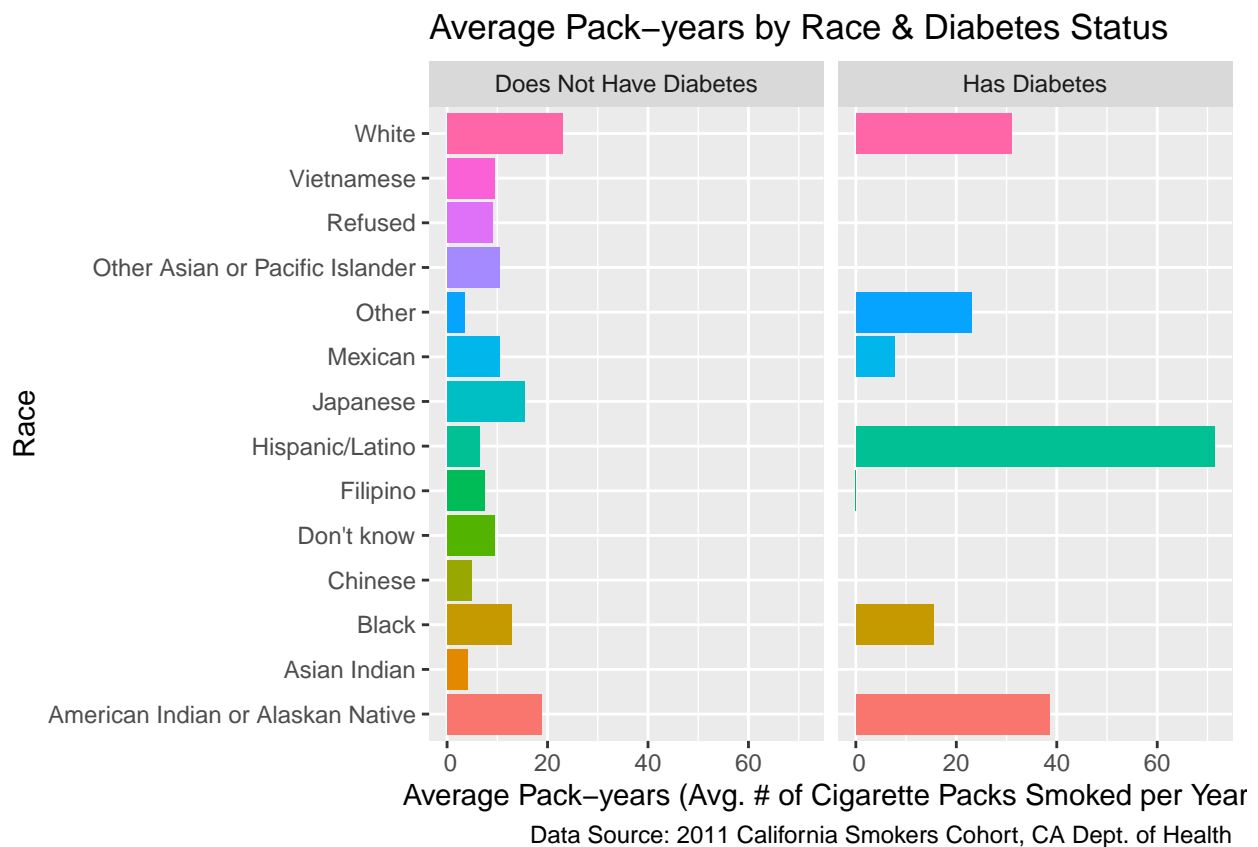

Average Pack–years by Race & Heart Disease Status

Data Source: 2011 California Smokers Cohort, CA Dept. of Health

```
avg_pack_years_race_diabetes %>%
  drop_na(c(diabetes, avg_pack_years)) %>%
  ggplot(aes(x = race, y = avg_pack_years)) +
  geom_bar(aes(fill=race), stat="identity", position = "dodge") +
  coord_flip() +
  guides(fill = "none") +
  labs(x = "Race",
       y = "Average Pack-years (Avg. # of Cigarette Packs Smoked per Year)",
  title = "Average Pack-years by Race & Diabetes Status",
  caption = "Data Source: 2011 California Smokers Cohort, CA Dept. of Health") +
  scale_y_continuous(labels = function(x) format(x,big.mark=",",
                                                 scientific=FALSE)) +
  facet_wrap(~ diabetes,
             labeller = labeller(diabetes = c("No" = "Does Not Have Diabetes",
                                               "Yes" = "Has Diabetes")))
```
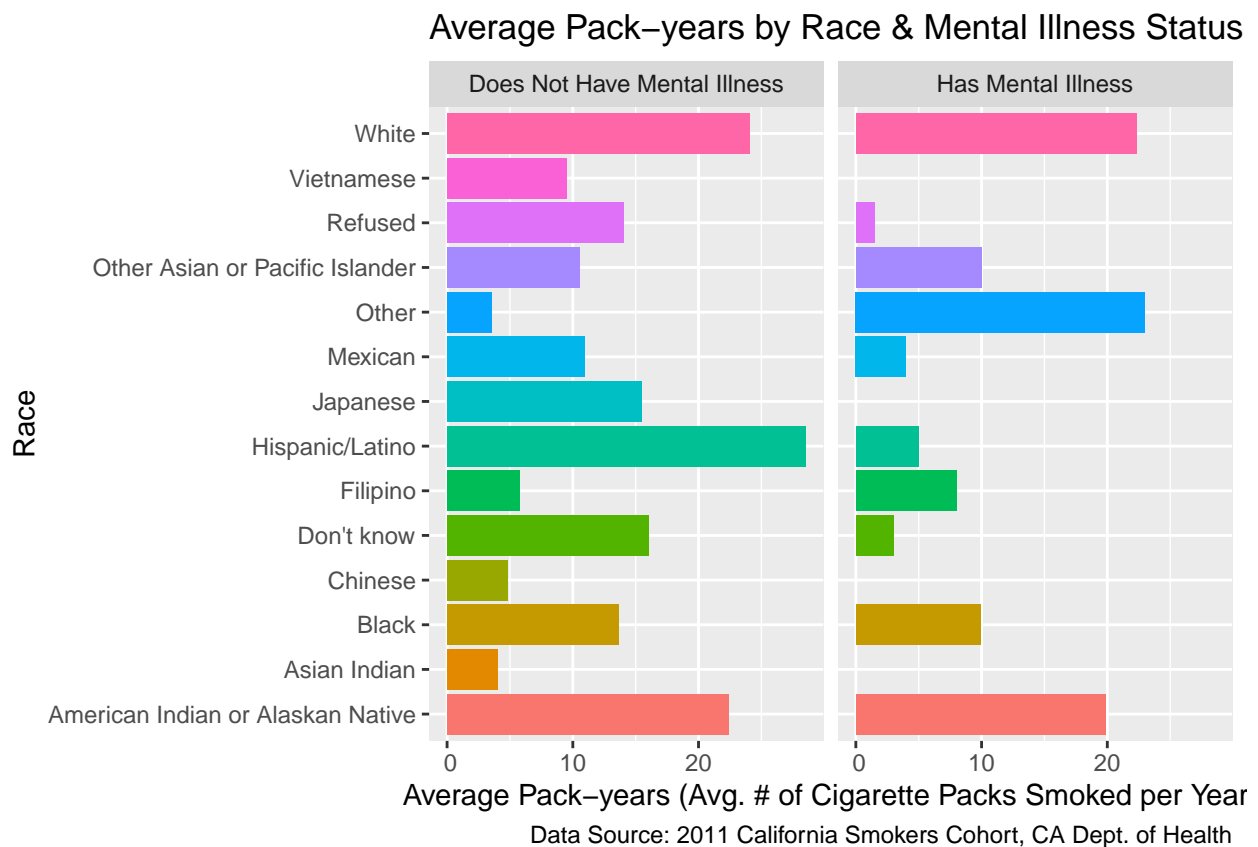

Average Pack-years by Race & Diabetes Status

Data Source: 2011 California Smokers Cohort, CA Dept. of Health

```
avg_pack_years_race_othmenill %>%
  drop_na(c(othmenill, avg_pack_years)) %>%
  ggplot(aes(x = race, y = avg_pack_years)) +
  geom_bar(aes(fill=race), stat="identity", position = "dodge") +
  coord_flip() +
  guides(fill = "none") +
  labs(x = "Race",
       y = "Average Pack-years (Avg. # of Cigarette Packs Smoked per Year)",
  title = "Average Pack-years by Race & Mental Illness Status",
  caption = "Data Source: 2011 California Smokers Cohort, CA Dept. of Health") +
  scale_y_continuous(labels = function(x) format(x,big.mark=",",
                                          scientific=FALSE)) +
  facet_wrap(~ othmenill, labeller = labeller(othmenill =
                                c("No" = "Does Not Have Mental Illness",
                                  "Yes" = "Has Mental Illness")))
```


Average Pack-years by Race & Mental Illness Status

Data Source: 2011 California Smokers Cohort, CA Dept. of Health

Interpretation of the Four Graphs of Average Pack-years by Race and Disease:
Except for mental illness, among individuals who reported that they do not have asthma, heart disease, and/or diabetes, those who identified as either "White" or "American Indian or Alaskan Native" by race appear to have a greater number of average pack-years (i.e., average number of cigarette packs smoked per year) compared to other races in the 2011 California Smokers Cohort.

Except for mental illness, among individuals who reported that they do have asthma, heart disease, and/or diabetes, those who identified as "Hispanic/Latino" by race appear to have the greatest number of average pack-years (i.e., average number of cigarette packs smoked per year) compared to other races in the 2011 California Smokers Cohort.

Work in progress:

##OUR GROUP'S RESEARCH QUESTION (For reference only to help us think of relevant #graphs - delete later): "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** kable tables for avg pack-years by disease – GOT THEM

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

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"**