Milestone #3

Rachael Baartmans, Lara Petalio, Christine Truong

11-07-22

```
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()
```

```
race_data <-read_csv("ca_csc_outcome_race_data.csv",
            col_select = c(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: 23
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (23): NERVOUS, WORRYING, PROBINTR, PROBDOWN, ASTHMA, HEARTDIS, DIABETES,...
##
## 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.
```

```r
smoker_data <- read_csv("ca_csc_smoker_data.csv",
            col_select = c(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: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): smokstat, HOWMANY, SMOK6UNI, BUYCALIF, WHEREBUY
## dbl (1): 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.
```

Subset rows and columns as needed

We have decided that we do not need to subset any columns since we already did this during the importing process of our data (specified in the col_select argument of the read_csv function). However, we have subsetted the smoker_data data frame for only rows that have "Years" as the value for the variable `smok6uni` for pack-year calculations that we will be performing later on our analysis. This subsetting is done while we cleaned our variables, as shown in the next step of this milestone.

Clean variables for analysis

Minimum of 2
**Examples: Recode invalid values/handle missing fields/recode categories **

```
#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")

#Viewed unique values assigned to the cleaned variables `howmany`, `wherebuy`,
# and `smok6uni` to see the changes in the variables' values
unique(smoker_data_2$howmany)
```

```
##  [1]  30  20  15   7  10   5   6  60   8  25  40 100   4  18  24   2   9  12  NA
## [20]  35  11  48  50   3  13   1  21  17  14  29  16
```

```
unique(smoker_data_2$wherebuy)
```

```
## [1] "At other discount or warehouse stores such as Wal-Mart or Costco"
## [2] "At tobacco discount stores"
## [3] "At convenience stores or gas stations"
## [4] "At liquor stores or drug stores"
## [5] "In military commissaries"
## [6] NA
## [7] "At supermarkets"
## [8] "Somewhere else"
## [9] "On Indian reservations"
```

```
unique(smoker_data_2$smok6uni)
```

```
## [1] "Years"
```

Create New Variables needed for analysis

Minimum of 2 created from existing columns
Examples: calculating the rate or combining character strings

```
#Created new variable `race` to combine variables race01:race15 into a single
#column. We also renamed race01 through race15 as the race categories they stand
#for, such as, White, Black, Japanese, and so on to facilitate comprehension of
#the race of each individual in our data set at first glance
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))
#Used select() function to remove original race01:race15 variables following
#combining all race variables into one column of `race`

#Viewed the updated data set, race_data_2
race_data_2
```

```
## # A tibble: 1,000 x 9
##    nervous   worrying probintr probdown asthma heartdis diabetes othmenill race
##    <chr>     <chr>    <chr>    <chr>    <chr>  <chr>    <chr>    <chr>     <chr>
##  1 Nearly e~ Not at ~ Nearly ~ Not at ~ No     Yes      No       No        White
##  2 Several ~ Several~ Several~ Not at ~ No     No       No       No        White
##  3 Not at a~ Not at ~ Not at ~ Not at ~ No     No       No       No        White
##  4 Several ~ Not at ~ Not at ~ Not at ~ Yes    No       No       No        White
##  5 Not at a~ Several~ Not at ~ Not at ~ No     No       No       No        White
##  6 Not at a~ Not at ~ Not at ~ Not at ~ No     Yes      No       No        White
##  7 Not at a~ Not at ~ Not at ~ Several~ Yes    Yes      No       No        White
##  8 Several ~ Nearly ~ Several~ Several~ No     No       No       No        White
##  9 Several ~ Several~ Several~ <NA>     No     No       No       Yes       White
## 10 More tha~ Several~ Not at ~ Not at ~ No     No       No       No        White
## # ... with 990 more rows
```

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

#Viewed the final cleaned data set, smoker_data_4
smoker_data_3
```

```
## # A tibble: 825 x 7
```

```
##    smokstat            wherebuy buycalif howmany smok6num smok6uni packs_per_day
##    <chr>               <chr>    <chr>      <dbl>    <dbl> <chr>             <dbl>
##  1 Current daily smok~ At othe~ In Cali~      30       36 Years              1.5
##  2 Current daily smok~ At toba~ In Cali~      20       25 Years              1
##  3 Current daily smok~ At conv~ In Cali~      15       20 Years              0.75
##  4 Current daily smok~ At conv~ In Cali~      15        7 Years              0.75
##  5 Current daily smok~ At liqu~ In Cali~      20       45 Years              1
##  6 Current daily smok~ At othe~ In Cali~      15       19 Years              0.75
##  7 Current daily smok~ At conv~ In Cali~       7        2 Years              0.35
##  8 Current daily smok~ At toba~ In Cali~      20       15 Years              1
##  9 Current daily smok~ In mili~ In Cali~      10       40 Years              0.5
## 10 Current daily smok~ <NA>     <NA>          20       27 Years              1
## # ... with 815 more rows
```

Data dictionary based on clean data set

must include: variable name, data type, and description Our group decided to pick 6 data elements from our cleaned data sets of race_data_2 and smoker_data_3 to include in our data dictionary. These variables are `wherebuy`, `howmany`, `nervous`, `asthma`, `race,`, and `packs_per_day`. The variables of `wherebuy`, `howmany`, and `packs_per_day` come from smoker_data_3, and the variables of `nervous`, `asthma`, and `race` come from race_data_2.

For each of the 6 data elements we picked, we must use the typeof() function to define its data type, as well as describe what the variable itself stands for in the context of this study question using the research documents published by the study researchers.

```
typeof(smoker_data_3$wherebuy)
```

```
## [1] "character"
```

```
typeof(smoker_data_3$howmany)
```

```
## [1] "double"
```

```
typeof(smoker_data_3$packs_per_day)
```

```
## [1] "double"
```

```
typeof(race_data_2$nervous)
```

```
## [1] "character"
```

```
typeof(race_data_2$asthma)
```

```
## [1] "character"
```

```
typeof(race_data_2$race)
```

```
## [1] "character"
```

Variable 1 name: wherebuy

- Data Type: character
- Description: The `wherebuy` variable contains the responses to the survey question of "Where do/did you usually buy your cigarettes?". Response options for participants include general locations (e.g., convenience stores, gas stations, super markets, liquor/drug stores, discount/warehouse stores) and more niche locations, such as Indian reservations and military commissaries. Other response options for this survey question include "Somewhere else" for locations not mentioned and for the participant to specify, as well as "Refused", or "Don't Know". Values of "Refused" and "Don't Know" were re-coded as NA during the initial import process of our two original data frames of race_data and smoker_data.

Variable 2 name: howmany

- Data Type: double
- Description: The variable `howmany` contains the numeric data related to how many cigarettes were smoked per day in the last 30 days if participants did smoke during that time frame. The values given for this question were 1 to 100, as well as "Refused" and "Don't Know". Responses of "Refused" and "Don't Know" were turned into NA values in our data during the initial import process of the original data.

Variable 3 name: packs_per_day

- Data Type: double
- Description: The variable `packs_per_day` contains the number of packs of cigarettes smoked per day for each observation in our data set. Our group created this variable in order to facilitate our calculations for tobacco consumption in "pack-years", which is defined by multiplying the number of packs of cigarettes smoked per day by the number of years a person has smoked.

Variable 4 name: nervous

- Data Type: character
- Description: The variable `nervous` is a character variable that contains responses to the question of whether individuals felt nervous, anxious, or on edge in the last two weeks at the time of interviews for the study. The responses for this question were "Not at all", "Several days", "More than half the days", "Nearly every day", "Refused", and "Don't Know". The responses of "Refused" and "Don't Know" were re-coded as NA during the initial importing process of the original data frame, race_data; all other responses remain untouched in our cleaned data frame, race_data_2.

Variable 5 name: asthma

- Data Type: character
- Description: The variable `asthma` pertains to medical history of individuals, as diagnosed by a doctor in the past. More specifically, the question asked to participants was, "Has a physician ever told you that you have any of the following conditions?". Possible responses were "Yes", "No", "Refused", and "Don't Know"; however, the original data set contained only "Yes" and "No" answers, which are shown in our cleaned data frame, race_data_2.

Variable 6 name: race

- Data Type: character
- Description: The variable `race` describes the racial background of individuals that partook in the study. The racial categories for this variable include White, Black, Japanese, Chinese, Filipino, Korean, Vietnamese, other Asian or Pacific Islander, American Indian or Alaskan Native, Mexican, Hispanic/Latino, Asian Indian, and other, as well as participant responses of "Refused" or "Don't Know".

Tables with descriptive statistics for 4 data elements

```
#We wanted to create a print-quality table to exhibit how many smokers bought their
#cigarettes from each unique location option listed in the study. To do this, we
#first created a new data frame called smoker_data_3_no_na, which contains
#only non-NA values for the variable `wherebuy` while keeping everything else
#the same from the smoker_data_3 data frame. This is because we didn't want any
#NA values shown in our frequency table.

smoker_data_3_no_na <- smoker_data_3 %>%
  drop_na(wherebuy)

#We then made the `wherebuy` variable in smoker_data_3_no_na into a table
#using the table() function and assigned it to the object table_smoker_wherebuy.

table_smoker_wherebuy <- table(smoker_data_3_no_na$wherebuy)

#Finally, we used the kable() function from the kableExtra package to create our
#print-quality table from table_smoker_wherebuy that shows the frequencies per
#unique cigarette-buying location mentioned in the study.

library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'

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

```
kable(table_smoker_wherebuy,
      booktabs=T,
      col.names=c("Buying Location", "Frequency"),
      align='lcccc',
      caption='\\textbf{Frequencies Per Cigarette-buying Location}',
      format = 'latex',
      format.args=list(big.mark=","))%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 1: **Frequencies Per Cigarette-buying Location**

Buying Location	Frequency
At convenience stores or gas stations	328
At liquor stores or drug stores	103
At other discount or warehouse stores such as Wal-Mart or Costco	46
At supermarkets	30
At tobacco discount stores	201
In military commissaries	7
On Indian reservations	20
Somewhere else	18

```r
#We created a print-quality table to show the average number of cigarettes smoked
#in the past 30 days before the study interview based on cigarette-buying location.
#To do this, we first subsetted data from the smoker_data_3 data frame, where we
#only kept the variables `wherebuy` and `howmany`. We then dropped all NA values
#from both of these variables, then grouped by `wherebuy` and calculated the mean
#of `howmany` for all grouped observations (i.e., mean number of cigarettes
#smoked per purchasing location). This subsetted data was assigned to the object
#table_smoker_howmany.

table_smoker_howmany <- smoker_data_3 %>%
  select(wherebuy, howmany) %>%
  drop_na(wherebuy, howmany) %>%
  group_by(wherebuy) %>%
  summarize(mean_number_of_cigarettes_smoked = mean(howmany))

#Again, we used the kable() function from the kableExtra package to create our
#print-quality table from the new subset data frame of table_smoker_howmany that
#shows the average number of cigarettes smoked per unique cigarette-buying location.

kable(table_smoker_howmany,
      booktabs=T,
      col.names=c("Buying Location", "Average Number of Cigarettes Smoked"),
      align='lcccc',
      caption='\\textbf{Mean No. of Cigarettes Smoked In
      the Past Month Based on Buying Location}',
      format = 'latex',
      format.args=list(big.mark=","), digits=2)%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 2: **Mean No. of Cigarettes Smoked In the Past Month Based on Buying Location**

Buying Location	Average Number of Cigarettes Smoked
At convenience stores or gas stations	15.16
At liquor stores or drug stores	14.81
At other discount or warehouse stores such as Wal-Mart or Costco	18.33
At supermarkets	16.90
At tobacco discount stores	16.13
In military commissaries	11.00
On Indian reservations	16.85
Somewhere else	24.56

```
#We wanted to create a print-quality table to exhibit how many smokers who were
#previously diagnosed with mental illness by a doctor felt nervous, anxious, or
#on edge over the past two weeks before the study interviews. To do this, we
#first subsetted data from the race_data_2 data frame, where we only kept the
#variables `nervous` and `othmenill`. Subsequently, we dropped all NA values
#from both of these variables because we did not want any NA variables to show
#up in the table we are creating. We then grouped by both variables and found the
#total number of smokers with or without previously diagnosed mental illness per
#value category for the variable `nervous`. Afterwards, we used pivot_wider() to
#expand our table horizontally with separate columns for `othmenill`'s values for
#better visualization of the table frequencies. Finally, we ordered the string
#values of `nervous` by converting the variable into a factor and arranging them
#from lowest to highest. The values being ordered range from lowest levels of
#nervousness/anxiousness/feeling on edge (i.e., "Not at all") to highest levels
#(i.e., "Nearly every day").

table_nervous_othmenill <- race_data_2 %>%
  select(nervous, othmenill) %>%
  drop_na(nervous, othmenill) %>%
  arrange(nervous) %>%
  group_by(othmenill, nervous) %>%
  summarize(count = n()) %>%
  pivot_wider(names_from = "othmenill", values_from = "count") %>%
  mutate(nervous = factor(nervous,
                          levels = c("Not at all", "Several days",
                                                   "More than half the days",
                                                   "Nearly every day"),
                          ordered = TRUE)) %>%
  arrange(nervous)
```

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

```
#Lastly, we used the kable() function from the kableExtra package to create our
#print-quality table from table_nervous_othmenill that demonstrates how many
#smokers with previously diagnosed mental illness experienced to some level or
#did not experience nervousness, anxiousness, or felt on edge in the past two
#weeks before the study interviews were conducted.

kable(table_nervous_othmenill,
      booktabs=T,
      col.names=c("Level of Nervousness/Anxiousness/Feeling On Edge",
                  "No Diagnosed Mental Illness", "Diagnosed Mental Illness"),
      align='lcccc',
      caption='\\textbf{Number of Smokers Per Level of
      Anxiety Feelings By Mental Illness Status}',
      format = 'latex',
      format.args=list(big.mark=","))%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 3: **Number of Smokers Per Level of Anxiety Feelings By Mental Illness Status**

Level of Nervousness/Anxiousness/Feeling On Edge	No Diagnosed Mental Illness	Diagnosed Mental Illness
Not at all	345	29
Several days	247	52
More than half the days	95	30
Nearly every day	119	58

```r
#We wanted to create a print-quality table to examine how many cases of mental
#illness or no mental illness exist among the smokers in this study by race.
#To do this, we first subsetted data from the race_data_2 data frame by
#selecting only the variables of `race` and `othmenill`, then dropping all NA
#values for both variables since we don't want any NA variables to show up in
#our table. Subsequently, we grouped by both variables and generated the
#frequencies per categorical combination of race and mental illness status.
#Following this step, we used pivot_wider() to expand our table horizontally
#with separate columns for `othmenill`'s values for better visualization of the
#table frequencies.

table_race_othmenill <- race_data_2 %>%
  select(race, othmenill) %>%
  drop_na(race, othmenill) %>%
  group_by(race, othmenill) %>%
  summarize(count = n()) %>%
  pivot_wider(names_from = "othmenill", values_from = "count")
```

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

```r
#Lastly, we used the kable() function from the kableExtra package to create our
#print-quality table from table_nervous_othmenill that exhibits how many smokers
#were or were not clinically diagnosed with mental illness by race category.

kable(table_race_othmenill,
      booktabs=T,
      col.names=c("Race", "No Diagnosed Mental Illness",
                  "Diagnosed Mental Illness"),
      align='lcccc',
      caption='\\textbf{Race and Mental Illness Status}',
      format.args=list(big.mark=","))%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 4: **Race and Mental Illness Status**

Race	No Diagnosed Mental Illness	Diagnosed Mental Illness
American Indian or Alaskan Native	29	10
Asian Indian	1	NA
Black	64	13
Chinese	7	NA
Don't know	1	1
Filipino	6	2
Hispanic/Latino	16	1
Japanese	6	NA
Mexican	17	2
Other	2	1
Other Asian or Pacific Islander	5	1
Refused	4	3
Vietnamese	2	NA
White	660	137