Milestone #2

Rachael Baartmans, Lara Petalio, Christine Truong

2022-10-02

What is the data source? (1-2 sentences on where the data is coming from, dates included, etc.)

Containing two data sets, the data source for this project is the 2011 California Smokers Cohort (CSC), which is a part of the California Tobacco Surveys (CTS) that aimed to investigate statewide tobacco use and behaviors among smokers to assess the effectiveness of smoking cessation strategies in California. It was sponsored by the California Department of Public Health (CDPH), and data was collected among smokers identified through the California Health Interview Survey Longitudinal Smokers Survey (CLSS) between July 8, 2011 and December 8, 2011.

How does the dataset relate to the group problem statement and question?

For this project, we aim to investigate how tobacco use impact mental illness among smokers in California in 2011 as well as explore how race and location of cigarette purchase can impact disease status. Since the data source include two separate data sets, they need to be joined by participant's ID, and present only the information necessary for our study, namely the participants' smoking status, location of cigarette purchase, race, and health outcomes. Tobacco consumption will also be compared to at least four disease outcomes (NEED TO CHOOSE FOUR: asthma, heart disease, diabetes, physical illness, and/or mental illness). We hope to provide some insight and suggestions for CDPH on how to redirect resources that can strengthen smoking cessation strategies at the end of the study.

Use appropriate import function and package based on type of file:

```
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_types = cols_only(NERVOUS = col_factor(),
                                           WORRYING = col_factor(),
                                           PROBINTR = col_factor(),
                                           PROBDOWN = col_factor(),
                                           OPHYSILL = col_factor(),
                                           race01 = col_factor(),
                                           race02 = col_factor(),
                                           race03 = col_factor(),
                                           race04 = col_factor(),
                                           race05 = col_factor(),
                                           race06 = col_factor(),
                                           race07 = col_factor(),
                                           race08 = col_factor(),
                                           race09 = col_factor(),
                                           race10 = col_factor(),
                                           race11 = col_factor()))
smoker_data <- read_csv("ca_csc_smoker_data.csv",
                        col_types = cols_only(smokstat = col_factor(),
                                              WHEREBUY = col_character(),
                                              BUYCALIF = col_character()))
```

Description of import process: We utilized the read_csv() function from the R readr package to import/read in our two csv-formatted data sets (one at a time) that contain information regarding the demographics, behaviors, smoking status, disease outcome, and race of individuals in the 2011 California Smokers Cohort. For the arguments of the read_csv() function, we only used col_types in order to specify the data types we wanted each column in the data sets to be imported as; no other arguments were deemed necessary to include in our code, as the data sets did not initially appear to have any missing values read incorrectly by R (i.e., not "NA") or any mislabeled columns or empty rows. We also included a nested function of cols_only() within the col_types argument so that we could read in only a subset of the columns from the original data sets, which were our variables of interest for the this project. In the data set containing demographic/behavioral/race information about the study participants, our variables of interest are NERVOUS, WORRYING, PROBINTR, PROBDOWN, OPHYSILL, race01, race02, race03, race04, race05, race06, race07, race08, race09, race10. In the data set containing smoking status/location of cigarette purchase/disease outcomes, our variables of interest are smokstat, WHEREBUY, and BUYCALIF. When our group imported each data set into R separately with the read_csv() function, we also assigned each subsetted csv to separate objects, which are called race_data and smoker_data.

Utilize function arguments to control relevant components: Arguments added to code chunk above.

Document the import process: The import process is documented in the paragraph written above with details as to how and why actions were performed.

Identify 5+ data elements required for your specified scenario:
Instructions: "In addition to smoking status, CDPH would like for you to explore the impact of race and at least one social (income, education) or one behavioral (location of cigarette purchase, type of cigarette purchase, brand of cigarette) factor on disease status. Your team may pick your disease of interest (e.g. asthma, heart disease, diabetes, physical illness, and mental illness) to analyze further."

In smoker_data:
"smokstat" – assigns smoking status
"WHEREBUY" – where you buy cigarettes?
"BUYCALIF" – do/did you usually buy your cigarettes…
"NERVOUS" – felt nervous, anxious, or on edge?
"WORRYING" – not been able to stop or control worrying?
"PROBINTR" – felt little interest or pleasure in doing things?
"PROBDOWN" – felt down, depressed, or hopeless?
"OPHYSILL" – has a physician ever told you that you have any mental illness?

In race_data:
race01:race11 –> excluded race14 (REFUSED) and race15 (DON'T KNOW) because these don't specify race

Total: 8 + 11 = 19 data elements/variables

CHRISTINE WILL ASK MORE ABOUT THIS ON ED DISCUSSION; ALL MEMBERS WILL DISCUSS FINAL ANSWER TOGETHER ON SUNDAY

Utilize functions or resources in RStudio to determine the types of each data element:

```
typeof(race_data)
```

```
## [1] "list"
```

```
class(race_data)
```

```
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

```
typeof(smoker_data)
```

```
## [1] "list"
```

```
class(smoker_data)
```

```
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

Identify the desired type/format for each variable- will you need to convert any columns to numeric or other type?:

```
summary(race_data)
```

```
##                       NERVOUS                        WORRYING
##  Nearly every day        :180   Not at all              :504
##  Several days            :302   Several days            :232
##  Not at all              :376   Nearly every day        :163
##  More than half the days :126   More than half the days : 83
##  (DO NOT READ) Refused   :  6   (DO NOT READ) Don’t know: 10
##  (DO NOT READ) Don’t know: 10   (DO NOT READ) Refused   :  8
##                       PROBINTR                       PROBDOWN
##  Nearly every day        :128   Not at all              :620
##  Several days            :221   Several days            :221
##  Not at all              :550   (DO NOT READ) Don’t know:  7
##  More than half the days : 76   More than half the days : 52
##  (DO NOT READ) Don’t know: 14   Nearly every day        : 95
##  (DO NOT READ) Refused   : 11   (DO NOT READ) Refused   :  5
##                       OPHYSILL   race01    race02    race03    race04
##  No                      :623   Yes:804      :913      :991      :991
##  Yes                     :367      :196   Yes: 87   Yes:  9   Yes:  9
##  (DO NOT READ) Don’t know:  5
##  NA/Not Applicable       :  2
##  (DO NOT READ) Refused   :  3
##
##  race05    race06    race07    race08    race09    race10    race11
##     :988      :999      :992      :933      :977      :981      :996
##  Yes: 12   Yes:  1   Yes:  8   Yes: 67   Yes: 23   Yes: 19   Yes:  4
##
##
##
##
```

```
summary(smoker_data)
```

```
##                      smokstat     BUYCALIF           WHEREBUY
##  Current daily smoker   :837   Length:1000        Length:1000
##  Current nondaily smoker:163   Class :character   Class :character
##                                Mode  :character   Mode  :character
```

While importing our data we changed the data sets to the types of variables needed for our analysis, because of this no changes are needed at this point.

Provide a basic description of 5+ data elements(numeric/character/other descriptives):

```
#Character elements
unique(race_data$WORRYING)
```

```
## [1] Not at all               Several days             Nearly every day
## [4] More than half the days  (DO NOT READ) Don’t know (DO NOT READ) Refused
## 6 Levels: Not at all Several days Nearly every day ... (DO NOT READ) Refused
```

```
unique(race_data$NERVOUS)
```

```
## [1] Nearly every day         Several days             Not at all
## [4] More than half the days  (DO NOT READ) Refused    (DO NOT READ) Don’t know
## 6 Levels: Nearly every day Several days Not at all ... (DO NOT READ) Don’t know
```

```
unique(smoker_data$smokstat)
```

```
## [1] Current daily smoker    Current nondaily smoker
## Levels: Current daily smoker Current nondaily smoker
```

```
unique(smoker_data$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, or"
##  [6] "On Indian reservations"
##  [7] NA
##  [8] "At supermarkets"
##  [9] "Somewhere else (SPECIFY)?"
## [10] "(DO NOT READ) Don’t know"
## [11] "(DO NOT READ) Refused"
```

```
#number of rows and columns

nrow(race_data)
```

```
## [1] 1000
```

```
nrow(smoker_data)
```

```
## [1] 1000
```

```
ncol(race_data)
```

```
## [1] 16
```

```
ncol(smoker_data)
```

```
## [1] 3
```

```
#Frequency of race data
table(race_data$race01)
```

```
##
## Yes
## 804 196
```

```
table(race_data$race02)
```

```
##
##     Yes
## 913  87
```

```
table(race_data$race03)
```

```
##
##     Yes
## 991   9
```

```
table(race_data$race04)
```

```
##
##     Yes
## 991   9
```

```
table(race_data$race05)
```

```
##
##     Yes
## 988  12
```

```
table(race_data$race06)
```

```
##
##     Yes
## 999   1
```

```
table(race_data$race07)
```

```
##
##     Yes
## 992   8
```

```
table(race_data$race08)
```

```
##
##     Yes
## 933  67
```

```
table(race_data$race09)
```

```
##
##     Yes
## 977  23
```

```
table(race_data$race10)
```

```
##
##     Yes
## 981  19
```

```
table(race_data$race11)
```

```
##
##     Yes
## 996   4
```