Milestone #2

Rachael Baartmans, Lara Petalio, Christine Truong

2022-10-03

Description of Dataset

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

Sponsored by the State of California's Department of Public Health (CDPH), the data source for this project is the 2011 California Smokers Cohort (CSC), which is a part of the California Tobacco Surveys (CTS) that collected information on the prevalence of tobacco use in California and behaviors among smokers in order to inform tobacco prevention efforts. To help CDPH better assess the effectiveness of smoking cessation strategies, the 2011 CSC data specifically investigates factors associated with quitting behavior among only smokers identified through telephone contacts purchased from data brokers, as well as through the California Health Interview Survey Longitudinal Smokers Survey (CLSS) between July 8, 2011 and December 8, 2011.

How does the data set relate to the group problem statement and 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. To better understand the relationships between these variables, the survey data would need to be cleaned up and present only our variables of interest (i.e., smoking status of each participant, location of cigarette purchase, race, and disease outcomes). The two separate data sets would also be joined by the unique ID of each participant, with race being re-coded into a single field from fifteen indicator variables to facilitate our subsequent analyses. Visualization of this cleaned, joined data through graphs or tables will then allow us to better assess the relationships between our variables of interest. In addition, we are interested in utilizing the 2011 CSC data to help us understand tobacco consumption in terms of "pack-years," which is the product of the number of packs of cigarettes smoked per day and the years a person has smoked, and comparing the average number of pack-years to asthma, heart disease, diabetes, and mental illness; this relationship can also be better understood through graphical or table visualization. Based on our observations from our analyses, 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 our project.

Import Statement

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

```
library(readr)
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(),
                                           ASTHMA = col_factor(),
                                           HEARTDIS = col_factor(),
                                           DIABETES = 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(),
                                           race12 = col_factor(),
                                           race13 = col_factor(),
                                           race14 = col_factor(),
                                           race15 = col_factor()))
smoker_data <- read_csv("ca_csc_smoker_data.csv",
                      col_types = cols_only(smokstat = col_factor(),
                                            WHEREBUY = col_character(),
                                            BUYCALIF = col_character(),
                                            HOWMANY = col_character(),
                                            SMOK6NUM = col_character(),
                                            SMOK6UNI = col_character()))
```

Utilize function arguments to control relevant components: Arguments added to code chunk above to specify data types and import specific columns/variables of interest.

Document the 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 outcomes, 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, ASTHMA, HEARTDIS, DIABETES, race01, race02, race03, race04, race05, race06, race07, race08, race09, race10, race11, race12, race13, race14, and race15. In the data set containing smoking status/location of cigarette purchase/disease outcomes, our

variables of interest are smokstat, WHEREBUY, BUYCALIF, HOWMANY, SMOK6NUM, and SMO6UNI. 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.

Identify types for 5+ data elements/columns/variables

Identify 5+ data elements required for your specified scenario:

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   ASTHMA
##  Nearly every day        :128   Not at all              :620   No :809
##  Several days            :221   Several days            :221   Yes:191
##  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
##                       HEARTDIS                         DIABETES
##  Yes                     : 81   No                      :899
##  No                      :916   Yes                     : 98
##  (DO NOT READ) Don't know:  3   (DO NOT READ) Don't know:  3
##
##
##
##                       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    race12    race07    race08    race09    race10
##     :988      :999      :997      :992      :933      :977      :981
##  Yes: 12   Yes:  1   Yes:  3   Yes:  8   Yes: 67   Yes: 23   Yes: 19
##
##
##
##
##  race13    race11    race14    race15
##     :998      :996      :993      :998
##  Yes:  2   Yes:  4   Yes:  7   Yes:  2
##
##
##
##
```

```
summary(smoker_data)
```

```
##                       smokstat     HOWMANY            SMOK6NUM
##  Current daily smoker   :837   Length:1000        Length:1000
##  Current nondaily smoker:163   Class :character   Class :character
```

```
##                                 Mode  :character   Mode  :character
##    SMOK6UNI           BUYCALIF           WHEREBUY
##  Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character
##  Mode  :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] 23
```

```
ncol(smoker_data)
```

```
## [1] 6
```

```
#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
```