# Milestone #2

Rachael Baartmans, Lara Petalio, Christine Truong

10-03-22

## 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 characters 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_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.
```

```
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.
```

**Utilize function arguments to control relevant components:**
Argument of col_select was added to read_csv() function above to import only specific columns/variables of interest, as well as na for R to read in missing values as "NA."

**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 used col_select to read in only a subset of the columns from the original data sets, which were our variables of interest for the this project. In addition, we included the argument of na in our read_csv()

functions for both data sets in order to indicate that any missing values that show up as blank or "(DO NOT READ)"/"n/a"/any variations of N/A are identified as NA values in R. During the importing process, each subsetted data set was assigned to two separate objects, which are called race_data and smoker_data. In the data set containing demographic/behavioral/race information about the study participants (i.e., data set labeled race_data), our variables of interest are NERVOUS, WORRYING, PROBINTR, PROBDOWN, ASTHMA, HEARTDIS, DIABETES, OTHMENILL, 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 (i.e., data set labeled as smoker_data), our variables of interest are smokstat, WHEREBUY, BUYCALIF, HOWMANY, SMOK6NUM, and SMOK6UNI.

## Identify types for 5+ data elements/columns/variables

### Identify 5+ data elements required for your specified scenario:

*In race_data:*
NERVOUS – "Over the past 2 weeks, how often have you felt nervous, anxious, or on edge?"
WORRYING – "Over the past 2 weeks, how often have you not been able to stop or control worrying?"
PROBINTR – "Over the past 2 weeks, how often have you felt little interest or pleasure in doing things?"
PROBDOWN – "Over the past 2 weeks, how often have you felt down, depressed, or hopeless?"
ASTHMA – "Has a physician ever told you that you have asthma?"
DIABETES – "Has a physician ever told you that you have diabetes?"
HEARTDIS – "Has a physician ever told you that you have heart disease?"
OTHMENILL – "Has a physician ever told you that you have any mental illness?"
race01:race15 – "Which of the following categories best describes your racial background?"

*In smoker_data:*
smokstat – Assigns smoking status
WHEREBUY – "Where do/did you usually buy your cigarettes?"
BUYCALIF – "Do/did you usually buy your cigarettes. . . "
HOWMANY – "During the past 30 days, on the days that you did smoke, about how many cigarettes did you usually smoke per day?" [100 = 100 OR MORE CIGARETTES]
SMOK6NUM – "How long have you been smoking on a daily basis?" (AMOUNT OF TIME)
SMOK6UNI – "How long have you been smoking on a daily basis?" (UNIT OF TIME)

Total: 23 + 6 = 29 data elements/variables

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

```
typeof(race_data$NERVOUS)
```

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

```
typeof(race_data$WORRYING)
```

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

```
typeof(race_data$PROBINTR)
```

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

```
typeof(race_data$PROBDOWN)
```

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

```
typeof(race_data$OTHMENILL)
```

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

```
typeof(race_data$ASTHMA)
```

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

```
typeof(race_data$HEARTDIS)
```

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

```
typeof(race_data$DIABETES)
```

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

```
typeof(race_data$race01)
```

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

```
typeof(race_data$race02)
```

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

```
typeof(race_data$race03)
```

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

```
typeof(race_data$race04)
```

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

```
typeof(race_data$race05)
```

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

```
typeof(race_data$race06)
```

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

```
typeof(race_data$race07)
```

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

```
typeof(race_data$race08)
```

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

```
typeof(race_data$race09)
```

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

```
typeof(race_data$race10)
```

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

```
typeof(race_data$race11)
```

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

```
typeof(race_data$race12)
```

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

```
typeof(race_data$race13)
```

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

```
typeof(race_data$race14)
```

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

```
typeof(race_data$race15)
```

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

```
typeof(smoker_data$smokstat)
```

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

```
typeof(smoker_data$WHEREBUY)
```

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

```
typeof(smoker_data$BUYCALIF)
```

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

```
typeof(smoker_data$HOWMANY)
```

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

```
typeof(smoker_data$SMOK6NUM)
```

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

```
typeof(smoker_data$SMOK6UNI)
```

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

**Identify the desired type/format for each variable - will you need to convert any columns to numeric or other type?:**
The desired type/format for the variables NERVOUS, WORRYING, PROBINTR, PROBDOWN, smokstat, WHEREBUY, BUYCALIF, HOWMANY, and SMOK6UNI is character. On the other hand, the desired type/format for the variables OTHMENILL, ASTHMA, HEARTDIS, DIABETES, and race01 through race15 is logical (i.e., "yes" or "no"). Also, the desired type/format for the variable SMOK6NUM is numeric. Because all variables except for SMOK6NUM (which is appropriately a numeric data type) are the character data type at the moment, we will need to convert the variables OTHMENILL, ASTHMA, HEARTDIS, DIABETES, and race01 through race15 to logical. Additionally, we will not need to convert the variable HOWMANY to any other data type either; although HOWMANY mostly consists of integers ranging from 0 to 100, the value of "100 or more cigarettes," a character string, under the variable could represent numeric values over 100 for every participant that has this value. This means that converting the data type from character to numeric for HOWMANY would not make the variable useful for any number summaries, in that it skews the results we see. For example, we can convert "100 or more cigarettes" to "100" to convey the same meaning, but in the form of a number. We can then convert the data type from character to numeric using as.numeric() and use the summary() function to find the minimum and maximum values of the variable. In this case, the range could be much greater for the variable than we see it here because the values labeled "100" could be greater than 100.

### Provide a Basic Description of the 5+ Data Elements

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

```
## [1] "Nearly every day"        "Several days"
## [3] "Not at all"              "More than half the days"
## [5] NA
```

```
unique(race_data$WORRYING)
```

```
## [1] "Not at all"              "Several days"
## [3] "Nearly every day"        "More than half the days"
## [5] NA
```

```
unique(race_data$PROBINTR)
```

```
## [1] "Nearly every day"        "Several days"
## [3] "Not at all"              "More than half the days"
## [5] NA
```

```
unique(race_data$PROBDOWN)
```

```
## [1] "Not at all"              "Several days"
## [3] NA                        "More than half the days"
## [5] "Nearly every day"
```

```
unique(race_data$OTHMENILL)
```

```
## [1] "No"  "Yes" NA
```

```
unique(race_data$ASTHMA)
```

```
## [1] "No"  "Yes"
```

```
unique(race_data$HEARTDIS)
```

```
## [1] "Yes" "No"  NA
```

```
unique(race_data$DIABETES)
```

```
## [1] "No"  "Yes" NA
```

```
unique(race_data$race01)
```

```
## [1] "Yes" NA
```

```
unique(race_data$race02)
```

```
## [1] NA    "Yes"
```

```
unique(race_data$race03)
```

```
## [1] NA    "Yes"
```

```
unique(race_data$race04)
```

```
## [1] NA    "Yes"
```

```
unique(race_data$race05)
```

```
## [1] NA    "Yes"
```

```
unique(race_data$race06)
```

```
## [1] NA    "Yes"
```

```
unique(race_data$race07)
```

```
## [1] NA    "Yes"
```

```
unique(race_data$race08)
```

```
## [1] NA    "Yes"
```

```
unique(race_data$race09)
```

```
## [1] NA    "Yes"
```

```
unique(race_data$race10)
```

```
## [1] NA    "Yes"
```

```
unique(race_data$race11)
```

```
## [1] NA    "Yes"
```

```
unique(race_data$race12)
```

```
## [1] NA    "Yes"
```

```
unique(race_data$race13)
```

```
## [1] NA    "Yes"
```

```
unique(race_data$race14)
```

```
## [1] NA    "Yes"
```

```
unique(race_data$race15)
```

```
## [1] NA    "Yes"
```

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

```
## [1] "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)?"
```

```
unique(smoker_data$HOWMANY)
```

```
##  [1] "30"                    "20"                    "1"
##  [4] "15"                    "3"                     "7"
##  [7] "10"                    "4"                     "5"
## [10] "6"                     "60"                    "8"
## [13] "18"                    NA                      "25"
## [16] "40"                    "100 or more cigarettes" "2"
## [19] "24"                    "9"                     "12"
## [22] "35"                    "11"                    "48"
## [25] "50"                    "19"                    "13"
## [28] "21"                    "17"                    "14"
## [31] "29"                    "16"
```

```
unique(smoker_data$SMOK6UNI)
```

```
## [1] "Years"  NA       "Days"   "Months"
```

**Basic description for character variables:**
For variables associated with the feelings of participants over the past 2 weeks before the time of the survey data collection (i.e., NERVOUS, WORRYING, PROBINTR, PROBDOWN), the unique values are a small number of phrases that include "Nearly every day," "Several days," "More than half the days," and "Not at all."
For variables associated with physical and mental illnesses (i.e., OTHMENILL, ASTHMA, HEARTDIS, DIABETES), the unique values are "No," and "Yes."
For variables associated with race (i.e., race01 through race15), the unique value is mainly "Yes," as well as "No" for some of these variables.
For the variable associated with smoker status (i.e., smokstat), the unique values are "Current daily smoker" and "Current nondaily smoker."
For the variable associated with location of cigarette purchase (i.e., WHEREBUY), the unique values include a variety of geographical locations, such as "At other discount or warehouse stores such as Wal-Mart or Costco," "At tobacco discount stores," "At convenience stores or gas stations," "At liquor stores or drug stores," "In military commissaries, or," "On Indian reservations," "At supermarkets," and "Somewhere else (SPECIFY)?"
For the variable associated with how many cigarettes participants smoke per day (i.e., HOWMANY), the unique values are a range of numbers in their numeric forms presented as character strings, along with "100 or more cigarettes."
For the variable associated with the unit of time for how long participants have been smoking on a daily basis (i.e., SMOK6UNI), the unique values are "Years," "Days," and "Months."

*Note: NA values are present for every variable of interest except for smokstat.*

```
summary(smoker_data$SMOK6NUM)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   15.00   30.00   26.05   36.00  120.00     168
```

**Basic description for numeric variables:**
For the variable associated with the amount of time participants have been smoking on a daily basis (i.e., SMOK6NUM), the mean is 26.05, the median is 30.00, and the range is 120.00 - 1.00 = 119.00.