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(), WORR
smoker_data <- read_csv("ca_csc_smoker_data.csv", col_types = cols_only(smokstat = col_factor(), WHEREBU

#Description of import process: We utilized the read_csv() function from the R readr package to import/

#can delete below if group is ok with current code above
# read_csv(file,
#          col_names = TRUE,
#          col_types = NULL,
#          locale = default_locale(),
#          na = c("", "NA"),
#          quoted_na = TRUE,
#          quote = "\"",
#          comment = "",
#          trim_ws = TRUE,
#          skip = 0,
#          n_max = Inf,
#          guess_max = min(1000, n_max),
#          progress = show_progress(),
#          skip_empty_rows = TRUE)
#CHRISTINE
```

Utilize function arguments to control relevant components: Add read_csv additional arguments later. **CHRISTINE** - added arguments to code chunk above (09/28/22)

Document the import process: Write up a hashtagged paragraph under importing code to describe what we did and why. **CHRISTINE** - wrote hashtagged descriptive paragraph in code chunk above (09/28/22)

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: class(),

typeof() functions used on variables in each relevant dataset. **Rachael**

Identify the desired type/format for each variable- will you need to convert any columns to numeric or other type?: as.numeric(), as.Date(), as.factor(), etc. functions if needed after viewing data types using str() or summary(). **Rachael We already changed the data type during the importing process of the data sets to the types we desire per variable – there shouldn't be a need to convert any columns to numeric or other type at this point - Christine**

Provide a basic description of 5+ data elements(numeric/character/other descriptives): Numeric – use summarize() function Character – unique() For other descriptives, will think of later when we come to them.