PHW251 Data Project Milestone 2

Nakisa Golabi, Joanna Liang, Desmond Gill

09/29/2022

#Description of dataset #What is the data source? (1-2 sentences on where the data is coming from, dates included, etc.) ca_csc_smoker_data.csv; ca_csc_outcome_race_data.csv This data source is from the UC San Diego library collections: (https://library.ucsd.edu/dc/object/bb6282371f). Creation date is 2011-2012 and it is a cross-sectional study looking at tobacco use and behaviors among Californians.

#How does the dataset relate to the group problem statement and question? The dataset provides the raw data needed to analyze the participant behaviors. We are interested in seeing the effects of certain individual behaviors & how this relates to smoking status as well as disease status (heart disease).

#Import statement NOTE: Please use datasets available in the PHW251 Project Data github repoLinks to an external site. (this is important to make sure everyone is using the same datasets) Use appropriate import function and package based on the type of file Utilize function arguments to control relevant components (i.e. change column types, column names, missing values, etc.) Document the import process

```
library(readr)
library(readxl)
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 dplyr   1.0.8
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tibble)
library(dplyr)

url_file = "https://raw.githubusercontent.com/PHW290/phw251_projectdata/main/ca_csc_smoker_data.csv"

smoker_dataset <- read_csv(url_file)
```

```
## Rows: 1000 Columns: 156
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (152): RIGHTSEX, smokstat, ACIG100, DOSMOKE, HOWMANY, SMOK6NUM, SMOK6UNI...
## dbl   (3): psraid, nosmknum1, quitoffn
## lgl   (1): QUITINTNFORM
##
## 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.
```

```
url_file_2 = "https://raw.githubusercontent.com/PHW290/phw251_projectdata/main/ca_csc_outcome_race_data

smoker_dataset_race <- read_csv(url_file_2)
```

```
## Rows: 1000 Columns: 89
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (81): ID, INCARS, BANAGREE, CASINSMK, CASMOKES, HHSMOKNU, ACQSMOKE, LIVE...
## dbl  (6): ACTIVHRS, ACTIVMIN, HTINFEET, HTINCHES, WGTINLBS, AGEUS
## lgl  (2): HTCENTIM, WGTINKILOS
##
## 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.
```

```
dataset_rename <- rename(smoker_dataset, ID = "psraid")

dataset_rename
```

```
## # A tibble: 1,000 x 156
##        ID RIGHTSEX smokstat   ACIG100 DOSMOKE HOWMANY SMOK6NUM SMOK6UNI MORNUNIT
##     <dbl> <chr>    <chr>      <chr>   <chr>   <chr>   <chr>    <chr>    <chr>
##  1 100099 Female   Current d~ Yes     Every ~ 30      36       Years    Minutes
##  2 100109 Female   Current d~ Yes     Every ~ 20      25       Years    Immedia~
##  3 100121 Male     Current n~ Yes     Some d~ 1       <NA>     <NA>     Immedia~
##  4 100191 Female   Current d~ Yes     Every ~ 15      20       Years    Minutes
##  5 100206 Male     Current d~ Yes     Every ~ 15      7        Years    Hours
##  6 100232 Female   Current d~ Yes     Every ~ 20      45       Years    Minutes
##  7 100256 Female   Current n~ Yes     Some d~ 3       <NA>     <NA>     Hours
##  8 100262 Female   Current d~ Yes     Every ~ 15      19       Years    Minutes
##  9 100317 Male     Current d~ Yes     Every ~ 7       2        Years    Minutes
## 10 100319 Female   Current d~ Yes     Every ~ 20      15       Years    Hours
## # ... with 990 more rows, and 147 more variables: mornnum <chr>, MUCHSMK <chr>,
## #   WHRUBUY <chr>, BRNDBUY <chr>, DESRQUIT <chr>, CIGCARTN <chr>,
## #   PAYCARTN <chr>, PAYPACK <chr>, BUYCALIF <chr>, WHEREBUY <chr>,
## #   SMKBRAND <chr>, SMO30MEN <chr>, WHYMENT1 <chr>, WHYMENT2 <chr>,
## #   WHYMENT3 <chr>, WHYMENT4 <chr>, smklotar <chr>, SMK1AGE <chr>,
## #   smkage <chr>, SMOKMORE <chr>, NOSMKUNI <chr>, nosmkuni2 <chr>,
## #   nosmknum1 <dbl>, nosmknum2 <chr>, QUITATPT <chr>, CONSIDER <chr>, ...
```

#Identify data types for 5+ data elements/columns/variables #Identify 5+ data elements required for your specified scenario. If <5 elements are required to complete the analysis, please choose additional variables of interest in the data set to explore in this milestone. Cigarette brands (SMKBRAND), current smoking status (smokstat), location of cigarette purchase (WHEREBUY), have you ever seriously considered quitting?

(CONSIDER), compared to a year ago, how often are you smoking? (SMOKMORE), race (RACE), heart disease (HEARTDIS)

#Utilize functions or resources in RStudio to determine the types of each data element (i.e. character, numeric, factor)

```
class(dataset_rename)
```

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

```
class(smoker_dataset_race)
```

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

```
typeof(dataset_rename)
```

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

```
typeof(smoker_dataset_race)
```

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

```
tibble(dataset_rename)
```

```
## # A tibble: 1,000 x 156
##        ID RIGHTSEX smokstat   ACIG100 DOSMOKE HOWMANY SMOK6NUM SMOK6UNI MORNUNIT
##     <dbl> <chr>    <chr>      <chr>   <chr>   <chr>   <chr>    <chr>    <chr>
##  1 100099 Female   Current d~ Yes     Every ~ 30      36       Years    Minutes
##  2 100109 Female   Current d~ Yes     Every ~ 20      25       Years    Immedia~
##  3 100121 Male     Current n~ Yes     Some d~ 1       <NA>     <NA>     Immedia~
##  4 100191 Female   Current d~ Yes     Every ~ 15      20       Years    Minutes
##  5 100206 Male     Current d~ Yes     Every ~ 15      7        Years    Hours
##  6 100232 Female   Current d~ Yes     Every ~ 20      45       Years    Minutes
##  7 100256 Female   Current n~ Yes     Some d~ 3       <NA>     <NA>     Hours
##  8 100262 Female   Current d~ Yes     Every ~ 15      19       Years    Minutes
##  9 100317 Male     Current d~ Yes     Every ~ 7       2        Years    Minutes
## 10 100319 Female   Current d~ Yes     Every ~ 20      15       Years    Hours
## # ... with 990 more rows, and 147 more variables: mornnum <chr>, MUCHSMK <chr>,
## #   WHRUBUY <chr>, BRNDBUY <chr>, DESRQUIT <chr>, CIGCARTN <chr>,
## #   PAYCARTN <chr>, PAYPACK <chr>, BUYCALIF <chr>, WHEREBUY <chr>,
## #   SMKBRAND <chr>, SMO30MEN <chr>, WHYMENT1 <chr>, WHYMENT2 <chr>,
## #   WHYMENT3 <chr>, WHYMENT4 <chr>, smklotar <chr>, SMK1AGE <chr>,
## #   smkage <chr>, SMOKMORE <chr>, NOSMKUNI <chr>, nosmkuni2 <chr>,
## #   nosmknum1 <dbl>, nosmknum2 <chr>, QUITATPT <chr>, CONSIDER <chr>, ...
```

```
tibble(smoker_dataset_race)
```

```
## # A tibble: 1,000 x 89
##    ID      INCARS BANAGREE CASINSMK CASMOKES HHSMOKNU ACQSMOKE LIVERELS OTHRRELS
##    <chr>   <chr>  <chr>    <chr>    <chr>    <chr>    <chr>    <chr>    <chr>
```

```
##  1 DIS100~ Not a~ Disagree Less li~ 40       1        3        Yes      Yes
##  2 DIS100~ Not a~ Disagree No diff~ 39       3        2        No       No
##  3 DIS100~ Not a~ Disagree Less li~ (DO NOT~ 1        10       No       Yes
##  4 DIS100~ Not a~ Disagree No diff~ 20       1        2        Yes      Yes
##  5 DIS100~ Not a~ Disagree No diff~ 20       1        4        No       No
##  6 DIS100~ Not a~ Disagree No diff~ (DO NOT~ 1        5        Yes      Yes
##  7 DIS100~ Not a~ Disagree No diff~ 95       2        9        Yes      Yes
##  8 DIS100~ Not a~ Disagree Less li~ 15       0        7        No       Yes
##  9 DIS100~ Not a~ Disagree No diff~ 40       3        5        Yes      Yes
## 10 DIS100~ Not a~ Disagree No diff~ 30       0        2        No       No
## # ... with 990 more rows, and 80 more variables: FRIENDS <chr>, SOCIAL <chr>,
## #   COWORKERS <chr>, SMOKALONE <chr>, ONLINEQTYN <chr>, RELGFREQ <chr>,
## #   CONSMOKE <chr>, WITHOTHR <chr>, SMKDRINK <chr>, SMKWKEND <chr>,
## #   HARMHLTH <chr>, AMADDICT <chr>, CAUSCANC <chr>, ADDICTIV <chr>,
## #   HELPDEPRESS <chr>, CALMANGRY <chr>, RELAXIRRIT <chr>, CALMNERVS <chr>,
## #   NERVOUS <chr>, WORRYING <chr>, PROBINTR <chr>, PROBDOWN <chr>,
## #   NOGOMYWAY <chr>, GOODGTBAD <chr>, CANDOALL <chr>, HELPLESS <chr>, ...
```

#Identify the desired type/format for each variable—will you need to convert any columns to numeric or another type? Will organize columns/rename/tidy as needed.

#Provide a basic description of the 5+ data elements #Numeric: mean, median, range #Character: unique values/categories #Or any other descriptives that will be useful to the analysis

SMKBRAND: Character variable; types of cigarette brands the individuals smoke (American Spirit, Basic, Benson & Hedges, Camel, Capri, Carlton, Djarum, Doral, Generic, GPC, Kent, Kool, Lucky Strike, Marlboro, Merit, Misty, More, Newport, Pall Mall, Parliament, Philip Morris, Raleigh, Salem, Virginia Slims, Winston, No special brand, Other (SPECIFY), REFUSED, DON'T KNOW)

smokstat: Character variable; unique values/categories: current Daily Smoker, Current Nondaily Smoker, Recent Quitter, Long-term quitter, Unspecified quitter, Never-Smoker, Unknown Smoking Status.

WHEREBUY: Character variable; unique values/categories: At convenience stores or gas stations, At super markets, At liquor stores or drug stores, At tobacco discount stores, At other discount or warehouse stores such as Wal-Mart or Costco, On Indian reservations, In military commissaries, Somewhere else?, (Specify), REFUSED, DON'T KNOW

CONSIDER: Character variable; unique values/categories: Yes, No, REFUSED, DON'T KNOW

SMOKMORE: Character variable; unique values/categories: The same as you were before, More than you were before, Less than you were before, REFUSED, DON'T KNOW

RACE: Character variable; unique values/categories: White, Black, Japanese, Chinese, Filipino, Korean, Vietnamese, Other Asian or Pacific Islander, American Indian or Alaskan Native, Mexican, Hispanic/Latino, Asian Indian, OTHER (Specify), REFUSED, DON'T KNOW

HEARTDIS: Character variable; unique values/categories: Yes, No, REFUSED, DON'T KNOW