Group AA Milestone 2

2022-10-03

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
hi <- "hi group!"
print(hi)

## [1] "hi group!"
```

Part 1: Describing the dataset

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

The data comes from the 2011 California Smokers' Cohort (CSC) and was the ninth of a series of triennial surveys called the California Tobacco Surveys (CTS) conducted since 1990. It was sponsored by the State of California's Department of Public Health through a contract with the University of California at San Diego (UCSD). Data collection for CLSS began on July 8, 2011 and was completed on December 8, 2011.

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

Smoking has been shown to lead to various poor health outcomes. However, some smokers are more prone than others to these adverse effects. The question is what characteristics and behaviors among smokers in California led to adverse health outcomes.

Part 2: Import statement

- 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

```
#import CA Smoker data set
ca_smoker_info <- read_csv("~/PHW251_Fall2022/phw251_projectdata/ca_csc_smoke
r_data.csv")

## Rows: 1000 Columns: 156
## ── Column specification ────────────────────────────────────────────────
──────
## Delimiter: ","
## chr (152): RIGHTSEX, smokstat, ACIG100, DOSMOKE, HOWMANY, SMOK6NUM, SMOK6U
NI...
## 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.

#tidying data
ca_smoker_selected <- ca_smoker_info %>% select(c(psraid,smokstat,HOWMANY, SM
OK6NUM, SMOK6UNI)) %>%
rename(ID = psraid, smoking_status = smokstat, howmany = HOWMANY, smok6num =
SMOK6NUM, smok6uni = SMOK6UNI) %>%
mutate(pack_year = howmany)
```

```
#import CA smoker disease outcome and race data set
ca_outcome_race <- read_csv("~/PHW251_Fall2022/phw251_projectdata/ca_csc_outc
ome_race_data.csv")
```

```
## Rows: 1000 Columns: 89
## ── Column specification ────────────────────────────────────────────────
──────
## Delimiter: ","
## chr (81): ID, INCARS, BANAGREE, CASINSMK, CASMOKES, HHSMOKNU, ACQSMOKE, LI
VE...
## 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.
```

```
#tidying data
ca_outcome_race_selected <- ca_outcome_race %>% select(c(ID, SOCIAL, ASTHMA,
                                                         HEARTDIS, DIABETES,
                                                         OTHMENILL, INCOME,
                                                         race01, race02, race
03,
                                                         race04, race05, race
06,
                                                         race07, race08, race
09,
                                                         race10, race11, race
12,
                                                         race13,race14, race1
5)) %>%
rename (social = SOCIAL, asthma = ASTHMA, heartdis = HEARTDIS, diabetes = DIA
BETES,
        othmenill = OTHMENILL, income = INCOME)
```

```
#remove "DIS" & "STAT" in the "ID" column
ca_outcome_race_selected$ID <- gsub("DIS","",as.character(ca_outcome_race_sel
ected$ID))
ca_outcome_race_selected$ID <- gsub("STAT","",as.character(ca_outcome_race_se
lected$ID))

#joining two data sets by participant's unique ID
ca_smoker_outcome <- merge(x = ca_smoker_selected, y = ca_outcome_race_select
ed, by = "ID")
```

```r
#use mutate to combine 15 binary race columns into one categorical variable c
alled "race"

ca_smoker_outcome <- ca_smoker_outcome %>%
  mutate(race = case_when(race01 == "Yes" ~ "White",
                          race02 == "Yes" ~ "Black",
                          race03 == "Yes" ~ "Japanese",
                          race04 == "Yes" ~ "Chinese",
                          race05 == "Yes" ~ "Filipino",
                          race06 == "Yes" ~ "Korean",
                          race12 == "Yes" ~ "Vietnamese",
                          race07 == "Yes" ~ "Other Asian Pacific Islander",
                          race08 == "Yes" ~ "American Indian Alaska Native",
                          race09 == "Yes" ~ "Mexican",
                          race10 == "Yes" ~ "Hispanic or Latino",
                          race11 == "Yes" ~ "Other",
                          race13 == "Yes" ~ "Asian Indian",
                          race14 == "Yes" ~ "Refused",
                          race15 == "Yes" ~ "Don't Know"))

#drop leftover binary race columns

ca_smoker_outcome <- select(ca_smoker_outcome,-race01,
                                -race02,
                                -race03,
                                -race04,
                                -race05,
                                -race06,
                                -race07,
                                -race08,
                                -race09,
                                -race10,
                                -race11,
                                -race12,
                                -race13,
                                -race14,
                                -race15)
```

Part 3: 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.
- Utilize functions or resources in RStudio to determine the types of each data element (i.e. character, numeric, factor)
- Identify the desired type/format for each variable—will you need to convert any columns to numeric or another type?

Five variables of interest: 1. Smoking status 2. Race 3. Income 4. Heart Disease 5. Pack years

```
#identify types of each data element
str(ca_smoker_outcome)

## 'data.frame':    1000 obs. of  13 variables:
##  $ ID            : num  1e+05 1e+05 1e+05 1e+05 1e+05 ...
##  $ smoking_status: chr  "Current daily smoker" "Current daily smoker" "Cur
rent nondaily smoker" "Current daily smoker" ...
##  $ howmany       : chr  "30" "20" "1" "15" ...
##  $ smok6num      : chr  "36" "25" NA "20" ...
##  $ smok6uni      : chr  "Years" "Years" NA "Years" ...
##  $ pack_year     : chr  "30" "20" "1" "15" ...
##  $ social        : chr  "No" "Yes" "Yes" "Yes" ...
##  $ asthma        : chr  "No" "No" "No" "Yes" ...
##  $ heartdis      : chr  "Yes" "No" "No" "No" ...
##  $ diabetes      : chr  "No" "No" "No" "No" ...
##  $ othmenill     : chr  "No" "No" "No" "No" ...
##  $ income        : chr  "$30,001 to $50,000" "$20,000 or less" "$30,001 to
$50,000" "$20,001 to $30,000" ...
##  $ race          : chr  "White" "White" "White" "White" ...

#convert data types to appropriate type in new column such as as.factor; as.n
umeric; as.character
ca_smoker_outcome <- ca_smoker_outcome %>% mutate(new_howmany = as.numeric(ho
wmany)) %>%
  mutate(new_smoking_status= as.factor(smoking_status)) %>%
  mutate(new_smok6num = as.numeric(smok6num)) %>%
  mutate(new_smok6uni = as.factor(smok6uni)) %>%
  mutate(new_pack_year = as.numeric(pack_year)) %>%
  mutate(new_social = as.factor(social)) %>%
  mutate(new_asthma = as.factor(asthma)) %>%
  mutate(new_heartdis = as.factor(heartdis)) %>%
  mutate(new_diabetes = as.factor(heartdis)) %>%
  mutate(new_othmenill = as.factor(othmenill)) %>%
  mutate(new_income = as.factor(income))

## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion

## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion

summary(ca_smoker_outcome$new_howmany)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    7.00   12.00   13.89   20.00   60.00      10
```

Part 4: 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

Smoking status:

```
#number of unique categories
ca_smoker_outcome %>% summarize(n_distinct(smoking_status))

##   n_distinct(smoking_status)
## 1                          2

#names of unique categories
ca_smoker_outcome %>% summarize(unique(smoking_status))

##    unique(smoking_status)
## 1    Current daily smoker
## 2 Current nondaily smoker

#tabulate smoking status
table(ca_smoker_outcome$smoking_status)

##
##    Current daily smoker Current nondaily smoker
##                     837                     163
```

Race:

```
#number of unique categories
ca_smoker_outcome %>% summarize(n_distinct(race))

##   n_distinct(race)
## 1               14

#names of unique categories
ca_smoker_outcome %>% summarize(unique(race))

##                     unique(race)
## 1                          White
## 2                          Black
## 3   Other Asian Pacific Islander
## 4  American Indian Alaska Native
## 5             Hispanic or Latino
## 6                        Mexican
## 7                   Asian Indian
## 8                       Filipino
## 9                       Japanese
## 10                    Don't Know
## 11                        Chinese
## 12                        Refused
```

```
## 13                          Other
## 14                     Vietnamese
```

```
#tabulate smoking status
table(ca_smoker_outcome$race)
```

```
##
## American Indian Alaska Native                  Asian Indian
##                            40                             1
##                         Black                       Chinese
##                            78                             7
##                    Don't Know                      Filipino
##                             2                             8
##            Hispanic or Latino                      Japanese
##                            17                             6
##                       Mexican                         Other
##                            19                             3
##  Other Asian Pacific Islander                       Refused
##                             6                             7
##                    Vietnamese                         White
##                             2                           804
```

Income:

```
#number of unique categories
ca_smoker_outcome %>% summarize(n_distinct(income))
```

```
##   n_distinct(income)
## 1                  9
```

```
#names of unique categories
ca_smoker_outcome %>% summarize(unique(income))
```

```
##             unique(income)
## 1       $30,001 to $50,000
## 2          $20,000 or less
## 3       $20,001 to $30,000
## 4     $100,001 to $150,000
## 5       $50,001 to $75,000
## 6            Over $150,000
## 7      $75,001 to $100,000
## 8    (DO NOT READ) Refused
## 9 (DO NOT READ) Don't know
```

```
#tabulate income
table(ca_smoker_outcome$income)
```

```
##
## (DO NOT READ) Don't know    (DO NOT READ) Refused     $100,001 to $150,000
##                       14                       48                       83
##          $20,000 or less       $20,001 to $30,000       $30,001 to $50,000
##                      243                      139                      182
```

```
##       $50,001 to $75,000      $75,001 to $100,000            Over $150,000
##                      157                       90                       44
```

Heart Disease:

```
#number of unique categories
ca_smoker_outcome %>% summarize(n_distinct(heartdis))

##   n_distinct(heartdis)
## 1                    3

#names of unique categories
ca_smoker_outcome %>% summarize(unique(heartdis))

##           unique(heartdis)
## 1                      Yes
## 2                       No
## 3 (DO NOT READ) Don't know

#tabulate income
table(ca_smoker_outcome$heartdis)

##
## (DO NOT READ) Don't know                       No                      Yes
##                        3                      916                       81
```

Pack Years:

```
#look at minimum, median, mean, maximum, and # of NAs in pack year
summary(ca_smoker_outcome$new_pack_year)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    7.00   12.00   13.89   20.00   60.00      10
```