Group AA Milestone 3

2022-11-01

Milestone 3

- Subset rows or columns, as needed
- Create new variables needed for analysis (minimum 2)
- New variables should be created based on existing columns; for example
- Calculating a rate
- Combining character strings
- Reordering income to CA - low, med, high
- Pack years -> find avg, and categorize those below "low" and above "high"
- If no new values are needed for final tables/graphs, please create 2 new variables anyway

taking out values that include don't know to NA

```
library(tidyverse)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
is.na(ca_smoker_outcome) <- ca_smoker_outcome == "(DO NOT READ) Don't know"
is.na(ca_smoker_outcome) <- ca_smoker_outcome == "(DO NOT READ) Refused"
is.na(ca_smoker_outcome) <- ca_smoker_outcome == "NA/Not Applicable"

#subsetting rows  and columns to include only "new" columns, rows include NA

str(ca_smoker_outcome)
```

```
## 'data.frame':    1000 obs. of  26 variables:
##  $ ID                : num  1e+05 1e+05 1e+05 1e+05 1e+05 ...
##  $ smoking_status    : chr  "Current daily smoker" "Current daily smoker" "Current nondaily smoker" "
##  $ howmany           : num  30 20 1 15 15 20 3 15 7 20 ...
##  $ smok6num          : num  36 25 NA 20 7 45 NA 19 2 15 ...
##  $ smok6uni          : chr  "Years" "Years" NA "Years" ...
##  $ temp_var          : num  1 1 NA 1 1 1 NA 1 1 1 ...
##  $ smok6num_inyears  : num  36 25 NA 20 7 45 NA 19 2 15 ...
##  $ pack_year         : num  54 25 NA 15 5.25 ...
##  $ 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 $
##  $ race              : chr  "White" "White" "White" "White" ...
##  $ new_howmany       : num  30 20 1 15 15 20 3 15 7 20 ...
##  $ new_smoking_status: Factor w/ 2 levels "Current daily smoker",..: 1 1 2 1 1 1 2 1 1 1 ...
##  $ new_smok6num      : num  36 25 NA 20 7 45 NA 19 2 15 ...
##  $ new_smok6uni      : Factor w/ 5 levels "(DO NOT READ) Don't know",..: 5 5 NA 5 5 5 NA 5 5 5 ...
##  $ new_pack_year     : num  54 25 NA 15 5.25 ...
```

```
##  $ new_social        : Factor w/ 4 levels "(DO NOT READ) Don't know",..: 3 4 4 4 4 3 4 4 4 4 ...
##  $ new_asthma        : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 2 1 1 1 ...
##  $ new_heartdis      : Factor w/ 3 levels "(DO NOT READ) Don't know",..: 3 2 2 2 2 3 3 2 2 2 ...
##  $ new_diabetes      : Factor w/ 3 levels "(DO NOT READ) Don't know",..: 3 2 2 2 2 3 3 2 2 2 ...
##  $ new_othmenill     : Factor w/ 4 levels "(DO NOT READ) Don't know",..: 3 3 3 3 3 3 3 3 4 3 ...
##  $ new_income        : Factor w/ 9 levels "(DO NOT READ) Don't know",..: 6 4 6 5 3 5 4 7 4 9 ...
```

```
ca_smoker_outcome <- select(ca_smoker_outcome,
                            c('ID', 'new_smoking_status', 'new_howmany',
                              'new_smok6num', 'new_smok6uni', 'new_pack_year',
                              'new_social', 'new_asthma', 'new_heartdis',
                              'new_diabetes', 'new_othmenill', 'new_income'))
```

creating new variable 1: standardizing numbers used to calculate pack-years variable to years unit

```
#create temporary variable for calculation
ca_smoker_outcome <- ca_smoker_outcome %>%
  mutate(temp_var = case_when(new_smok6uni=="Months" ~ 12,
                              new_smok6uni=="Years" ~ 1,
                              new_smok6uni=="Days" ~ 365))

#do calculation
ca_smoker_outcome <- ca_smoker_outcome %>%
  mutate(new_smok6num_inyears = new_smok6num/temp_var)

#drop unnecessary temporary variable
ca_smoker_outcome <-  subset(ca_smoker_outcome, select = -c(temp_var))
```

creating new variable 2: finding average pack years and creating low, average, high values

```
library(dplyr)
summary(ca_smoker_outcome$new_pack_year)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##   0.00438   8.75000  17.00000  21.45533  30.00000 120.00000       175
```

```
ca_smoker_outcome <- ca_smoker_outcome %>%
  mutate(pack_year_avg = mean(ca_smoker_outcome$new_pack_year, na.rm = TRUE))
ca_smoker_outcome$pack_year_avg <- round(ca_smoker_outcome$pack_year_avg ,digit= 0)
ca_smoker_outcome <- ca_smoker_outcome %>% mutate(pack_year_avg_level = case_when
                                                  (new_pack_year > pack_year_avg ~ "above average",
                                                    new_pack_year < pack_year_avg ~ "below average",
                                                    new_pack_year == pack_year_avg ~ "average"))
```

creating new variable 3: characterizing income as low, middle, high

```
unique(ca_smoker_outcome$new_income)
```

```
## [1] $30,001 to $50,000   $20,000 or less      $20,001 to $30,000
## [4] $100,001 to $150,000 $50,001 to $75,000   Over $150,000
## [7] $75,001 to $100,000  <NA>
## 9 Levels: (DO NOT READ) Don't know ... Over $150,000
```

```
ca_smoker_outcome <- ca_smoker_outcome %>% mutate(income_levels = case_when(
  new_income %in% c("$20,000 or less", "$20,001 to $30,000") ~ "Low income",
new_income %in% c("$30,001 to $50,000", "$50,001 to $75,000") ~ "Middle Income",
new_income %in% c("$75,001 to $100,000", "$100,001 to $150,000", "Over $150,000") ~ "High Income"))
```

Cleaning - NA values using tables to ensure no weird values

```
table(ca_smoker_outcome$new_smoking_status, useNA = "always")
```

```
##
##    Current daily smoker Current nondaily smoker                    <NA>
##                     837                     163                       0
```

```
table(ca_smoker_outcome$new_smok6uni, useNA = "always")
```

```
##
## (DO NOT READ) Don't know    (DO NOT READ) Refused                    Days
##                        0                        0                       4
##                   Months                   Years                    <NA>
##                        5                      825                     166
```

```
table(ca_smoker_outcome$new_social, useNA = "always")
```

```
##
## (DO NOT READ) Don't know        NA/Not Applicable                      No
##                        0                        0                     241
##                      Yes                     <NA>
##                      748                       11
```

```
table(ca_smoker_outcome$new_asthma, useNA = "always")
```

```
##
##   No  Yes <NA>
##  809  191    0
```

```
table(ca_smoker_outcome$new_heartdis, useNA = "always")
```

```
##
## (DO NOT READ) Don't know                      No                     Yes
##                        0                     916                      81
##                     <NA>
##                        3
```

```
table(ca_smoker_outcome$new_diabetes, useNA = "always")
```

```
##
## (DO NOT READ) Don't know                      No                     Yes
##                        0                     916                      81
##                     <NA>
##                        3
```

```
table(ca_smoker_outcome$new_othmenill, useNA = "always")
```

```
##
## (DO NOT READ) Don't know    (DO NOT READ) Refused                      No
##                         0                       0                     820
##                       Yes                    <NA>
##                       171                       9
```

```
table(ca_smoker_outcome$new_income, useNA = "always")
```

```
##
## (DO NOT READ) Don't know    (DO NOT READ) Refused    $100,001 to $150,000
##                         0                       0                      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
##                     <NA>
##                       62
```

Data dictionary based on clean dataset (minimum 4 data elements), including:

- Variable name
- Data type
- Description

```
variable_name <- c("ID", "new_smoking_status", "new_howmany", "new_smok6num",
                   "new_somk6uni", "new_pack_year", "new_social", "new_asthma",
                   "new_heartdis", "new_diabetes", "new_othmenill", "new_income", "smok_daily",
                   "pack_year_avg", "pack_year_avg_level", "income_levels")

data_type <- c("numeric", "character", "numeric", "numeric", "character", "numeric",
               "character", "character", "character", "character", "character",
               "character", "numeric", "numeric", "character", "character")

description <- c("Participant identification number",
                 "Current Smoking Status: Current Daily Smoker (Smoked >99 and smokes every day),
                 Current Nondaily Smoker(Smoked >99 and smokes some days)",
              "During the past 30 days, on the days that you did smoke, about
              'HOWMANY' cigarettes did you usually smoke per day?",
              "How long have you been smoking on a daily basis?", "The unit for smol6num variable",
              "A pack-year is used to describe how many cigarettes smoked in a person's lifetime,
              with a pack equal to 20 cigarettes",
              "yes/no the participant identifies as a social smoker",
              "yes/no the participant has asthma",
              "yes/no the participant has heart disease",
              "yes/no the participant has diabetes",
              "yes/no the participant reports 'other mental illness'",
              "participant household income",
              "cigarettes smoked daily",
              "average pack years for entire cohort, created for our average variable",
              "indicator for whether participant is below, at, or above the average pack years",
              "categorical income levels defined using California household income data")

data_dictionary <- data.frame(variable_name, data_type, description)
data_dictionary <- data_dictionary %>%
  rename(
    "Variable Name" = variable_name,
    "Data Type" = data_type,
    "Description" = description)

#load kable library
library(knitr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
kable(data_dictionary)
```

Variable Name	Data Type	Description
ID	numeric	Participant identification number
new_smoking_status	character	Current Smoking Status: Current Daily Smoker (Smoked >99 and smokes every day
new_howmany	numeric	During the past 30 days, on the days that you did smoke, about 'HOWMANY' ciga
new_smok6num	numeric	How long have you been smoking on a daily basis?
new_somk6uni	character	The unit for smol6num variable
new_pack_year	numeric	A pack-year is used to describe how many cigarettes smoked in a person's lifetime, v
new_social	character	yes/no the participant identifies as a social smoker
new_asthma	character	yes/no the participant has asthma
new_heartdis	character	yes/no the participant has heart disease
new_diabetes	character	yes/no the participant has diabetes
new_othmenill	character	yes/no the participant reports 'other mental illness'
new_income	character	participant household income
smok_daily	numeric	cigarettes smoked daily
pack_year_avg	numeric	average pack years for entire cohort, created for our average variable
pack_year_avg_level	character	indicator for whether participant is below, at, or above the average pack years
income_levels	character	categorical income levels defined using California household income data

One or more tables with descriptive statistics for 4 data element

Table 1: Smooker Outcomes Distributions

```
#calculate data for proportion table
summary(ca_smoker_outcome$new_social)
```

```
## (DO NOT READ) Don't know        NA/Not Applicable                        No
##                        0                        0                       241
##                      Yes                      NA's
##                      748                       11
```

```
summary(ca_smoker_outcome$new_asthma)
```

```
##  No Yes
## 809 191
```

```
summary(ca_smoker_outcome$new_heartdis)
```

```
## (DO NOT READ) Don't know                       No                      Yes
##                        0                      916                       81
##                     NA's
##                        3
```

```
summary(ca_smoker_outcome$new_diabetes)
```

```
## (DO NOT READ) Don't know                       No                      Yes
##                        0                      916                       81
##                     NA's
##                        3
```

```
summary(ca_smoker_outcome$new_othmenill)
```

```
## (DO NOT READ) Don't know    (DO NOT READ) Refused                       No
##                        0                        0                      820
##                      Yes                      NA's
##                      171                        9
```

```
#build dataframe
outcome_variable <- c("Social Smoker", "Asthma", "Heart Disease", "Diabetes", "Other Mental Illness")
yes_count_variable <- c(748, 191, 81, 81, 171)
no_count_variable <- c(241, 809, 916, 916, 820)
NA_count_variable <- c(11, 0, 3, 3, 9)
yes_prop_variable <- c("74.8%", "19.1%", "8.1%", "8.0%", "17.1%")
no_prop_variable <- c("24.1%", "80.9%", "91.6%", "91.6%", "82.0%")
NA_prop_variable <- c("1.1%", "0", "0.3%", "0.3%", "0.9%")

proportion_df <- data.frame(outcome_variable, yes_count_variable,
                            no_count_variable, NA_count_variable,
```

```
                              yes_prop_variable, no_prop_variable,
                              NA_prop_variable)

#create table
outcomes_table <- kable(proportion_df, booktabs=T, align="l",
                        col.names=c("Condition", "Yes Count",
                                    "No Count", "Not Applicable",
                                    "Percent Yes", "Percent No",
                                    "Percent Not Applicable"))
outcomes_table
```

Condition	Yes Count	No Count	Not Applicable	Percent Yes	Percent No	Percent Not Applicable
Social Smoker	748	241	11	74.8%	24.1%	1.1%
Asthma	191	809	0	19.1%	80.9%	0
Heart Disease	81	916	3	8.1%	91.6%	0.3%
Diabetes	81	916	3	8.0%	91.6%	0.3%
Other Mental Illness	171	820	9	17.1%	82.0%	0.9%

Table 2: Spread of Continuous Variables

```
#get data
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
```

```
summary(ca_smoker_outcome$new_pack_year)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##   0.00438   8.75000  17.00000  21.45533  30.00000 120.00000       175
```

```
#build dataframe
measure <- c("Average Cigarettes Smoked Per Day", "Pack Years")
minimum <- c(1, 0.10)
median <- c(12, 17)
mean <- c(13.89, 21.68)
maximum <- c(60, 120)
distribution_df <- data.frame(measure, minimum, median, mean, maximum)

#create table
distribution_table <- kable(distribution_df, booktabs=T, align="lcccc",
                        col.names=c("Measure", "Minimum Value",
                                    "Median Value", "Mean",
                                    "Maximum Value"))
distribution_table
```

Measure	Minimum Value	Median Value	Mean	Maximum Value
Average Cigarettes Smoked Per Day	1.0	12	13.89	60
Pack Years	0.1	17	21.68	120

```
end <- "The End"

print(end)
```

```
## [1] "The End"
```