PHW251 Data Project Milestone 3

Nakisa Golabi, Joanna Liang, Desmond Gill

11/3/2022

SUBSET ROWS OR COLUMNS

```
#Subset rows or columns

smoker_dataset_clean <- rename_with(smoker_dataset, ~ tolower(gsub(" ","_", .x,
                                                                   fixed=TRUE)))

smoker_dataset_race_clean <- rename_with(smoker_dataset_race, ~ tolower(gsub(" ","_", .x,
                                                                   fixed=TRUE)))

smoker_dataset_race_clean$id<-gsub("DIS","",as.character(smoker_dataset_race_clean$id))
smoker_dataset_race_clean$id<-gsub("STAT","",as.character(smoker_dataset_race_clean$id))

dataset1 <- rename(smoker_dataset_clean, id=psraid)

jointdataset <- merge(dataset1, smoker_dataset_race_clean, by = c('id'))
```

CREATE NEW VARIABLES NEEDED FOR ANALYSIS

```
#Create new variables needed for analysis (minimum 2)

jointdataset_2 <- na_if(jointdataset, "(DO NOT READ) Don't know")
jointdataset_3 <- na_if(jointdataset_2, "(DO NOT READ) Refused")

jointdataset_3$smkage[jointdataset_3$smkage=="5 years old or less"]<-"5"
jointdataset_3$smkage[jointdataset_3$smkage=="Never smoked regularly"]<-"0"
jointdataset_3$smk1age[jointdataset_3$smk1age=="5 years old or less"]<-"5"

jointdataset_4 <- drop_na(jointdataset_3, smk1age, smkage)

jointdataset_4$smkage <- as.numeric(as.character(jointdataset_4$smkage))
jointdataset_4$smk1age <- as.numeric(as.character(jointdataset_4$smk1age))

#Variable 1
jointdataset_5 <- jointdataset_4 %>%
   mutate(smoker_onset = (smkage - smk1age))

jointdataset_5$howmany[jointdataset_5$howmany=="100 or more cigarettes"]<-"100"
jointdataset_6 <- drop_na(jointdataset_5, howmany)

#Variable 2
clean_data <- jointdataset_6 %>% group_by(smokstat, howmany) %>%
  mutate(smoke_level= case_when(
    smokstat=="Current nondaily smoker" & howmany <= 10 ~ "nondaily low level",
    smokstat=="Current nondaily smoker" & howmany <= 30  ~ "nondaily medium level",
    smokstat=="Current nondaily smoker" ~ "nondaily high level",
    smokstat=="Current daily smoker" & howmany <= 10 ~ "daily low level",
    smokstat=="Current daily smoker" & howmany <= 30  ~ "daily medium level",
    smokstat=="Current daily smoker" ~ "daily high level"
  ))

#Variable 3
clean_data2 <- drop_na(clean_data, smok6num)
clean_data3 <- drop_na(clean_data2, smok6uni)

clean_data3$howmany <- as.numeric(as.character(clean_data3$howmany))
clean_data4 <- mutate(clean_data3, packs = howmany/20)

clean_data4$smok6num <- as.numeric(as.character(clean_data4$smok6num))

clean_data5 <- clean_data4 %>% mutate(smoking_years =
                              case_when(
               smok6uni == "Years" ~ smok6num,
               smok6uni == "Months" ~ smok6num/12,
               smok6uni == "Days" ~ smok6num/365,
               ))

clean_data6 <- clean_data5 %>% mutate(pack_years= round(packs*smoking_years, 2))
```

CLEAN VARIABLES NEEDED FOR ANALYSIS

```
#Clean variables needed for analysis (minimum 2)

#smkbrand
clean_data6$smkbrand[clean_data6$smkbrand=="Other (SPECIFY)"]<-"Other"
clean_data7 <- drop_na(clean_data6, smkbrand)

#wherebuy
clean_data7$wherebuy[clean_data7$wherebuy=="Somewhere else (SPECIFY)?"]<-"Somewhere else"
clean_data8 <- drop_na(clean_data7, wherebuy)

#not including consider in analysis

#smokmore
clean_data9 <- drop_na(clean_data8, smokmore)

#heartdis
clean_data10 <- drop_na(clean_data9, heartdis)

#othmenill
clean_data11 <- drop_na(clean_data10, othmenill)

#diabetes
clean_data12 <- drop_na(clean_data11, diabetes)

#asthma
clean_data13 <- drop_na(clean_data12, asthma)

#race
clean_data14 <- clean_data13
clean_data14$race01 <- if_else(clean_data13$race01 =="Yes",1,0)
clean_data14$race02 <- if_else(clean_data13$race02 =="Yes",1,0)
clean_data14$race03 <- if_else(clean_data13$race03 =="Yes",1,0)
clean_data14$race04 <- if_else(clean_data13$race04 =="Yes",1,0)
clean_data14$race05 <- if_else(clean_data13$race05 =="Yes",1,0)
clean_data14$race06 <- if_else(clean_data13$race06 =="Yes",1,0)
clean_data14$race12 <- if_else(clean_data13$race12 =="Yes",1,0)
clean_data14$race07 <- if_else(clean_data13$race07 =="Yes",1,0)
clean_data14$race08 <- if_else(clean_data13$race08 =="Yes",1,0)
clean_data14$race09 <- if_else(clean_data13$race09 =="Yes",1,0)
clean_data14$race10 <- if_else(clean_data13$race10 =="Yes",1,0)
clean_data14$race13 <- if_else(clean_data13$race13 =="Yes",1,0)
clean_data14$race11 <- if_else(clean_data13$race11 =="Yes",1,0)
clean_data14$race14 <- if_else(clean_data13$race14 =="Yes",1,0)
clean_data14$race15 <- if_else(clean_data13$race15 =="Yes",1,0)

clean_data14$race01[is.na(clean_data14$race01)] = 0
clean_data14$race02[is.na(clean_data14$race02)] = 0
clean_data14$race03[is.na(clean_data14$race03)] = 0
clean_data14$race04[is.na(clean_data14$race04)] = 0
clean_data14$race05[is.na(clean_data14$race05)] = 0
clean_data14$race06[is.na(clean_data14$race06)] = 0
clean_data14$race12[is.na(clean_data14$race12)] = 0
```

```
clean_data14$race07[is.na(clean_data14$race07)] = 0
clean_data14$race08[is.na(clean_data14$race08)] = 0
clean_data14$race09[is.na(clean_data14$race09)] = 0
clean_data14$race10[is.na(clean_data14$race10)] = 0
clean_data14$race13[is.na(clean_data14$race13)] = 0
clean_data14$race11[is.na(clean_data14$race11)] = 0
clean_data14$race14[is.na(clean_data14$race14)] = 0
clean_data14$race15[is.na(clean_data14$race15)] = 0

sum(clean_data14$race11)
```

```
## [1] 3
```

```
sum(clean_data14$race14)
```

```
## [1] 5
```

```
sum(clean_data14$race15)
```

```
## [1] 0
```

```
clean_data15 <- clean_data14 %>%
  mutate(race_num = race01+ race02+ race03+ race04+ race05 +race06+
           race12 + race07+ race08 + race09+ race10+ race13)

clean_data16 <- filter(clean_data15, id != 100099 & id != 100109 & id != 100191
                       & id != 100206 & id != 100232)
```

DATA DICTIONARY BASED ON CLEAN DATASET

VARIABLE NAME: smkbrand

DATA TYPE: character

DESCRIPTION: 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

VARIABLE NAME: smokstat

DATA TYPE: character

DESCRIPTION: Smoking status with unique values/categories: current Daily Smoker, Current Nondaily Smoker, Recent Quitter, Long-term quitter, Unspecified quitter, Never-Smoker, Unknown Smoking Status.

VARIABLE NAME: smoke_level

DATA TYPE: character

DESCRIPTION: level of smoking status with unique values/categories: nondaily medium level, nondaily high level, daily low level, daily medium level, daily high level

VARIABLE NAME: wherebuy

DATA TYPE: character

DESCRIPTION: where the participant bought cigarettes, with 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

VARIABLE NAME: smokmore

DATA TYPE: character

DESCRIPTION: compared to last year, was the participant smoking more, with unique values/categories: The same as you were before, More than you were before, Less than you were before

VARIABLE NAME: heartdis

DATA TYPE: character

DESCRIPTION: has a physician told the participant if they have heart disease, with unique values/categories: Yes, No

VARIABLE NAME: othmenill

DATA TYPE: character

DESCRIPTION: if the participant had any mental illness, with unique values/categories: Yes, No

VARIABLE NAME: diabetes

DATA TYPE: character

DESCRIPTION: if the participant had diabetes, with unique values/categories: Yes, No

VARIABLE NAME: asthma

DATA TYPE: character

DESCRIPTION: if the participant had asthma, with unique values/categories: Yes, No

VARIABLE NAME: smoker_onset

DATA TYPE: numeric

DESCRIPTION: the difference between the age of the participant's first cigarette and the age when they started smoking regularly

VARIABLE NAME: pack_years

DATA TYPE: numeric

DESCRIPTION: the product of the number of packs of cigarettes smoked per day and the years a person has smoked

TABLES WITH DESCRIPTIVE STATISTICS

```
#packs
mean(clean_data13$packs)
```

```
## [1] 0.7845304
```

```
max(clean_data13$packs)
```

```
## [1] 5
```

```
min(clean_data13$packs)
```

```
## [1] 0.05
```

```
#smoking_years
mean(clean_data13$smoking_years)
```

```
## [1] 25.44923
```

```
max(clean_data13$smoking_years)
```

```
## [1] 53
```

```
min(clean_data13$smoking_years)
```

```
## [1] 0.0109589
```

```
#pack_years
mean(clean_data13$pack_years)
```

```
## [1] 21.30007
```

```
max(clean_data13$pack_years)
```

```
## [1] 210
```

```
min(clean_data13$pack_years)
```

```
## [1] 0
```

```
tibble_pack_years <- tibble(
  statistics = c("mean","minimum","maximum","range"),
  packs = c(0.7845304, 0.5, 5, 4.5),
  smoking_years = c(25.44923, 0.0109589, 53, 52.98904),
  pack_years = c(21.30007, 0,210,210),
)
tibble_pack_years
```

```
## # A tibble: 4 x 4
##   statistics packs smoking_years pack_years
##   <chr>      <dbl>         <dbl>      <dbl>
## 1 mean       0.785       25.4          21.3
## 2 minimum    0.5          0.0110        0
## 3 maximum    5           53            210
## 4 range      4.5         53.0          210
```

```r
as.data.frame(table(clean_data13$heartdis))
```

```
##   Var1 Freq
## 1   No  673
## 2  Yes   51
```

```r
as.data.frame(table(clean_data13$asthma))
```

```
##   Var1 Freq
## 1   No  587
## 2  Yes  137
```

```r
as.data.frame(table(clean_data13$othmenill))
```

```
##   Var1 Freq
## 1   No  604
## 2  Yes  120
```

```r
as.data.frame(table(clean_data13$diabetes))
```

```
##   Var1 Freq
## 1   No  656
## 2  Yes   68
```

```r
tibble_diseases <- tibble(
  counts = c("yes", "no"),
  heart_disease = c(51, 673),
  asthma = c(137, 587),
  mental_illness = c(120, 604),
  diabetes = c(68,656)
)
tibble_diseases
```

```
## # A tibble: 2 x 5
##   counts heart_disease asthma mental_illness diabetes
##   <chr>          <dbl>  <dbl>          <dbl>    <dbl>
## 1 yes               51    137            120       68
## 2 no               673    587            604      656
```