Group AA Milestone 4

2022-11-18

Problem statement

Smoking cigarettes damages almost all of the body's organs, increases the risk of several illnesses, and generally lowers people's health. Additionally, the most common preventable cause of mortality in the US is smoking cigarettes. Smoking harms almost all organs and is linked to numerous cancers, heart conditions, strokes, respiratory conditions such chronic obstructive pulmonary disease, diabetes, and other disorders. However, some smokers are more vulnerable to these negative effects than others. The question is what characteristics and behaviors among smokers in California lead to adverse health outcomes?

At any age, even among heavy and longtime smokers, quitting smoking can have a significant positive impact on both short- and long-term health. For the prevention and cessation of smoking, a range of therapies are successful. With the use of data from the 2011 California Smokers' Cohort (CSC), the California Tobacco Surveys (CTS), we may better understand the tobacco use and smoking habits of Californian smokers and put initiatives into place in high-risk areas to encourage quitting.

In the 2011 CSC study, 1,000 smokers completed the follow up survey and we had decided to clean and analyze the data for health characteristics. Our data sets were originally in two data sets, so they were combined by patient ID which correctly resulted in the total of 1,000 participants. Once that was finalized our group had cleaned and analyzed the data by factoring all variables that were characters and uniformed the units of pack years. Our study population did not have a very diverse spread, so we decided to exclude race in our analysis. Our team decided to investigate and focus on health outcomes and the stratification by income level, so we decided to subset our data accordingly to variables that would be of importance in our visualization analysis.

In our visualizations, when comparing all bar-graphs of health outcomes off asthma, heart disease, and diabetes stratified by income status, we can observe that the study population has a larger proportion of low-income patients, therefore we cannot assume that those with health outcomes are disproportionately low-income households. In patients who have heart disease, it is observed that they are the exact same patients that have diabetes, which we can interpret that heart disease and diabetes have some biological correlation.

In our outcomes table, those who were social smokers were about 75% of the study population, which may allude to a need for improved community public health intervention to reduce the prevalence on social smoking. Our outcomes table also includes prevalence percentages for the health outcomes of asthma, heart disease and other mental illnesses. If we wanted to investigate further, then we could you these percentages if we wanted to see if they are statistically significant enough for this study population.

In our distribution table, we can see that of those in the study population that did smoke, smoked an average of 13 cigarettes per day and have an average 21 pack years. Considering that 20-40 pack years are categorized as moderate smokers, these groups have higher odds in health outcomes and there could be effect modification on other confounders that we have not identified within the study.

Milestone 4: Visualizations

Visual 1: Table One of Pack Years and Health Outcomes

```
library(kableExtra)
```

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

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

```
library(tableone) #load table one library

#build new dataset with only necessary variables
visual3_df <- test_variable_df %>%
  select(new_asthma, new_diabetes, new_heartdis, new_pack_year, new_othmenill, pack_year_avg_level) %>%
  rename(Asthma=new_asthma,
         "Above/Below Average Pack Years" = pack_year_avg_level,
         Diabetes=new_diabetes,
         "Heart Disease"=new_heartdis,
         "Pack Years"=new_pack_year,
         "Other Mental Health" =new_othmenill)
#create table one
visual3table <- CreateTableOne(data=visual3_df,
  vars=c("Pack Years", "Asthma", "Diabetes", "Heart Disease", "Other Mental Health"),
  factorVars = c("Asthma", "Diabetes", "Heart Disease", "Other Mental Health", "Above/Below Average Pack
  strata="Above/Below Average Pack Years")
kable(print(visual3table, showAllLevels=TRUE),
      caption="Pack Years and Health Outcomes Stratified by Above/Below Average Pack Years")
```

```
##                           Stratified by Above/Below Average Pack Years
##                            level above average average        below average
##   n                                   317             5            503
##   Pack Years (mean (SD))          39.50 (16.86) 22.00 (0.00)   10.44 (6.10)
##   Asthma (%)               No       254 (80.1)      2 ( 40.0)    403 (80.1)
##                            Yes       63 (19.9)      3 ( 60.0)    100 (19.9)
##   Diabetes (%)             No       274 (87.3)      5 (100.0)    473 (94.0)
##                            Yes       40 (12.7)      0 (  0.0)     30 ( 6.0)
##   Heart Disease (%)        No       274 (87.3)      5 (100.0)    473 (94.0)
##                            Yes       40 (12.7)      0 (  0.0)     30 ( 6.0)
##   Other Mental Health (%)  No       255 (81.7)      4 ( 80.0)    406 (81.2)
##                            Yes       57 (18.3)      1 ( 20.0)     94 (18.8)
##                           Stratified by Above/Below Average Pack Years
##                            p      test
##   n
##   Pack Years (mean (SD))  <0.001
##   Asthma (%)               0.083
##
##   Diabetes (%)             0.003
```

Table 1: Pack Years and Health Outcomes Stratified by Above/Below Average Pack Years

	level	above average	average	below average	p	test
n		317	5	503		
Pack Years (mean (SD))		39.50 (16.86)	22.00 (0.00)	10.44 (6.10)	<0.001	
Asthma (%)	No	254 (80.1)	2 (40.0)	403 (80.1)	0.083	
	Yes	63 (19.9)	3 (60.0)	100 (19.9)		
Diabetes (%)	No	274 (87.3)	5 (100.0)	473 (94.0)	0.003	
	Yes	40 (12.7)	0 (0.0)	30 (6.0)		
Heart Disease (%)	No	274 (87.3)	5 (100.0)	473 (94.0)	0.003	
	Yes	40 (12.7)	0 (0.0)	30 (6.0)		
Other Mental Health (%)	No	255 (81.7)	4 (80.0)	406 (81.2)	0.979	
	Yes	57 (18.3)	1 (20.0)	94 (18.8)		

```
##
##   Heart Disease (%)        0.003
##
##   Other Mental Health (%)  0.979
##
```

Interpretation:

This table is a "table one": a common table of descriptive statistics for a study sample. This table one displays descriptive statistics for our health outcomes of interest, as well as the pack years variable. All are stratified by one of our factor variables, "above/below pack years average", which displays whether the participant's pack years calculated in Milestone 3 were above, at, or below the average pack years of the entire sample.

Visual 2: Outcome and Income Cross Tabulations

```
# For asthma status data set
a <- test_variable_df %>%
  tabyl(new_asthma, income_levels, show_na = FALSE) %>%
  adorn_totals(c("col", "row")) %>%
  as.data.frame() %>%
  rename(Asthma_status = new_asthma)

a <- a[, c("Asthma_status", "High Income", "Middle Income", "Low Income", "Total")]

a %>%
  kbl(align = "l") %>%
  kable_styling(latex_options = "striped") %>%
  add_header_above(c(" " = 1, "Income Level" = 3, " " = 1)) %>%
  add_header_above(data.frame("Cross Tab of Asthma Status x Income Level", 5))
```

Cross Tab of Asthma Status x Income Level				
	Income Level			
Asthma_status	High Income	Middle Income	Low Income	Total
No	187	290	285	762
Yes	30	49	97	176
Total	217	339	382	938

Interpretation:

Of the 176 patients with asthma, 30 were reported as the high-income group, 49 as the middle-income group, and 97 as the low-income group. Of the 762 study participants without asthma, 187 were reported as the high-income group, 290 as the middle-income group, and 285 as the low-income group.

```
# For heart disease status data set
b <- test_variable_df %>%
  tabyl(new_heartdis, income_levels, show_na = FALSE) %>%
  adorn_totals(c("col", "row")) %>%
  as.data.frame() %>%
  rename(Heartdis_status = new_heartdis)

b <- b[-1, c("Heartdis_status", "High Income", "Middle Income", "Low Income", "Total")]

b %>%
  kbl(align = "l") %>%
  kable_styling(latex_options = "striped") %>%
  add_header_above(c(" " = 2, "Income Level" = 3, " " = 1)) %>%
  add_header_above(data.frame("Cross Tab of Heart Disease Status x Income Level", 6))
```

Cross Tab of Heart Disease Status x Income Level					
		Income Level			
	Heartdis_status	High Income	Middle Income	Low Income	Total
2	No	202	319	337	858
3	Yes	14	20	43	77
4	Total	216	339	380	935

Interpretation:

Of the 77 patients with heart disease, 14 were reported as the high-income group, 20 as the middle-income group, and 43 as the low-income group. Of the 858 study participants without heart disease, 202 were reported as the high-income group, 319 as the middle-income group, and 337 as the low-income group.

```
# For diabetes status data set
c <- test_variable_df %>%
  tabyl(new_diabetes, income_levels, show_na = FALSE) %>%
  adorn_totals(c("col", "row")) %>%
  as.data.frame() %>%
  rename(Diabetes_status = new_diabetes)

c <- c[-1, c("Diabetes_status", "High Income", "Middle Income", "Low Income", "Total")]

c %>%
  kbl(align = "l") %>%
  kable_styling(latex_options = "striped") %>%
  add_header_above(c(" " = 2, "Income Level" = 3, " " = 1)) %>%
  add_header_above(data.frame("Cross Tab of Diabetes Status x Income Level", 6))
```

Cross Tab of Diabetes Status x Income Level					
		Income Level			
	Diabetes_status	High Income	Middle Income	Low Income	Total
2	No	202	319	337	858
3	Yes	14	20	43	77
4	Total	216	339	380	935

Interpretation:

Of the 77 patients with heart disease, 14 were reported as the high-income group, 20 as the middle-income group, and 43 as the low-income group. Of the 858 study participants without heart disease, 202 were reported as the high-income group, 319 as the middle-income group, and 337 as the low-income group.

Visual 3: Of those who do have health outcomes, how are they distributed by income status. How are those that do not have health outcomes distributed by income?

Asthma data set and ggplot

```
asthma_by_income <- test_variable_df %>% select(new_asthma, income_levels) %>%
  group_by(new_asthma, income_levels) %>%  mutate(asthma_count = n()) %>%
  distinct(income_levels,.keep_all = TRUE) %>% na.omit()

A <- ggplot(data = na.omit(asthma_by_income), aes(x = new_asthma, y= asthma_count)) +
  geom_bar(aes(fill = income_levels), stat = "identity", position = position_dodge()) +
    scale_fill_discrete(name = "Income Level") +
  labs(x= "Prevalence of Asthama", y = "# of Patients",
       title = "Patients and Their Asthma Status Distributed by Income Status",
       caption = "Note: Patients who did not respond were excluded")
```

Heart Disease data set and ggplot

```
heartdis_by_income <- test_variable_df %>% select(new_heartdis, income_levels) %>%
  group_by(new_heartdis, income_levels) %>%  mutate(heartdis_count = n()) %>%
  distinct(income_levels,.keep_all = TRUE) %>% na.omit()

B <- ggplot(data = heartdis_by_income, aes(x = new_heartdis, y= heartdis_count)) +
  geom_bar(aes(fill = income_levels), stat = "identity", position = position_dodge()) +
  scale_fill_discrete(name = "Income Level") +
  labs(x= "Prevalence of Heart Disease", y = "# of Patients",
       title = "Patients and Their Heart Disease Status Distributed by Income Status",
       caption = "Note: Patients who did not respond were excluded")
```

Diabetes data set and ggplot

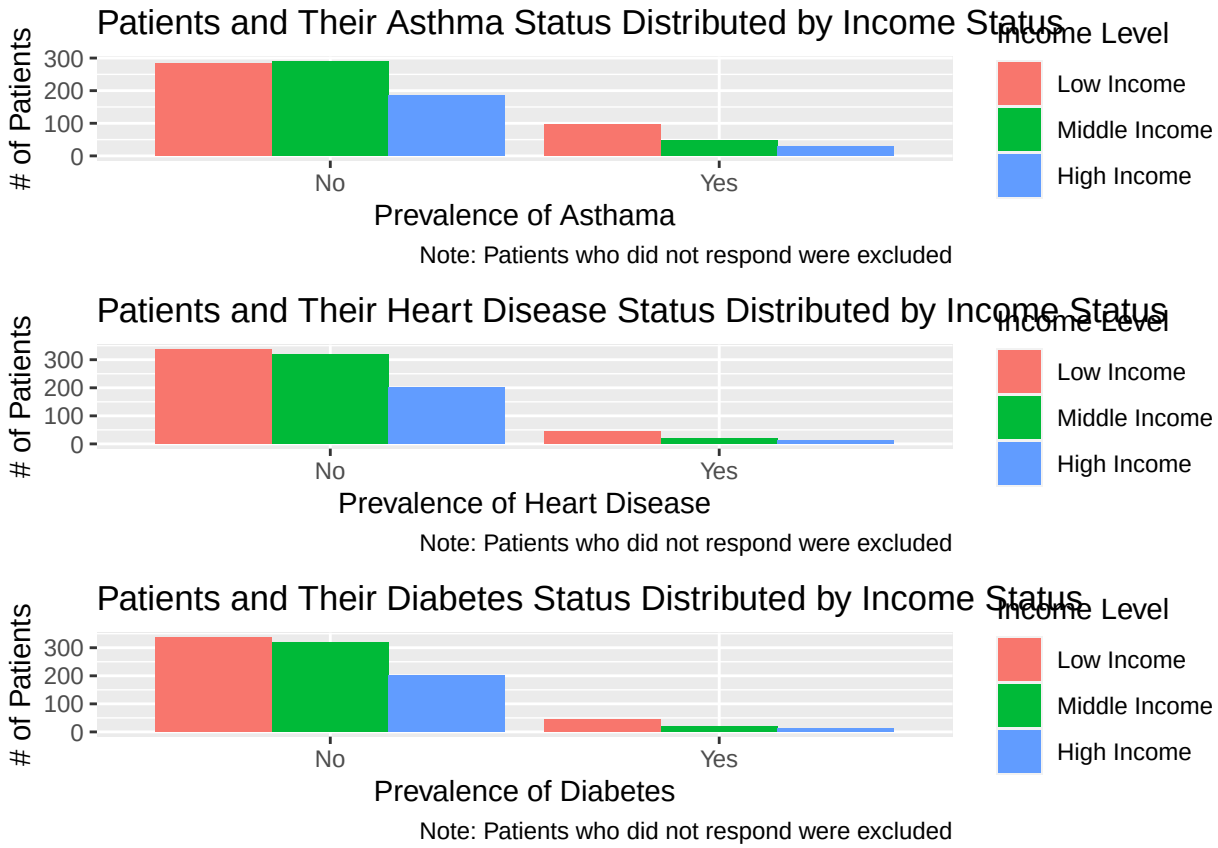
```
diabetes_by_income <- test_variable_df %>% select(new_diabetes, income_levels) %>%
  group_by(new_diabetes, income_levels) %>%  mutate(diabetes_count = n()) %>%
  distinct(income_levels,.keep_all = TRUE) %>% na.omit()

C<- ggplot(data = diabetes_by_income, aes(x = new_diabetes, y= diabetes_count)) +
  geom_bar(aes(fill = income_levels), stat = "identity", position = position_dodge()) +
  scale_fill_discrete(name = "Income Level") +
  labs(x= "Prevalence of Diabetes", y = "# of Patients",
       title = "Patients and Their Diabetes Status Distributed by Income Status",
       caption = "Note: Patients who did not respond were excluded")
```

Combined Dodged Barcharts displaying Prevalence of Asthma, Heart Disease and Diabetes Stratified by Income Status

```
# Hidden code allows download of ggpubr to utilize the ggarrange function to
# combine all seperate ggplots together.

figure_one <- ggarrange(A, B, C, ncol = 1, nrow = 3)
figure_one
```



Interpretation In Patients and Their Asthma Status Distributed by Income Status, of the 762 patients that reported no asthma, 285 patients came from low-income households, 290 patients came from middle-income households, and 187 patients came from high-income households. Of the 176 patients that reported having asthma, 97 patients came from low-income households, 49 patients came from middle-income households, and 30 patients came from high-income households.

In Patients and Their Heart Disease Status Distributed by Income Status, of the 858 patients that reported no heart disease, 337 patients came from low-income households, 319 patients came from middle-income households, and 202 patients came from high-income households. Of the 77 patients that reported having heart disease, 43 patients came from low-income households, 20 patients came from middle-income households, and 14 patients came from high-income households.

In Patients and Their Diabetes Status Distributed by Income Status, of the 858 patients that reported no diabetes, 337 patients came from low-income households, 319 patients came from middle-income households, and 202 patients came from high-income households. Of the 77 patients that reported having diabetes, 43 patients came from low-income households, 20 patients came from middle-income households, and 14 patients came from high-income households.

Bonus Table from Milestone 3: Smoker Outcomes Distributions

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

#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 N/A"))
outcomes_table
```

Condition	Yes Count	No Count	Not Applicable	Percent Yes	Percent No	Percent N/A
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%

Interpretation:

This table is a set of descriptive statistics related to our variables of interest, and the prevalence of each outcome in our sample.

Bonus Table from Milestone 3: 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
##    0.10    8.80   17.00   21.68   30.00  120.00     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

Interpretation:

This table displays the spread of the variables "Average Cigarettes Smoked per Day" and "Pack Years" in our sample.

.

Not for grading, but also including our data dictionary from Milestone 3 with adjustment recommended by Will, which helped with page runoff. Thank you for this suggestion!

```
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)

kable(data_dictionary, "latex") %>%
  kable_styling(full_width=TRUE)
```

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), Current Nondaily Smoker(Smoked >99 and smokes some days)
new_howmany	numeric	During the past 30 days, on the days that you did smoke, about 'HOWMANY' cigarettes did you usually smoke per day?
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, with a pack equal to 20 cigarettes
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