Milestone #4

Ileah Rios & Amber Morris

Milestone #1

1. Project option selection: We pick the California Smokers' Cohort data set for this project.

2. Create git repository / Share a link to your group's git repository: https://github.com/ambernmorris/team_project

3. What is your team's preferred communication method - email, text, bcourse messaging? Text, email, Zoom calls

4. When will your team be holding meetings? How frequently will you meet? Are there times or days that work well for everyone? Once a week to every other week; we are already in contact frequently because we have other classes together and have worked together before. Our schedules vary, but typically Wednesday mornings work best for us to meet.

5. Discuss future non-academic commitments that might affect members' availability: Work, life events, partners, vacations. We communicate clearly when we have conflicts. Discuss meeting tempo: "checking in" at the beginning of meetings versus "just sticking to business". We like to check-in and then we usually get to work after chatting some.

6. How is your team going to keep track of progress? Who will be taking minutes, creating agendas, and contacting the course facilitators with questions? We both come to our meetings prepared with questions and a general idea of what we will be working on together. Amber will be the one usually to contact course instructors.

7. We encourage you to discuss potential dates and times to meet with a course facilitator during Weeks 3-5. Decide on the best date and time for all. We will meet with Lauren on Mondays via private office hour appointment if needed.

8. Determine a point person to submit each assignment for the team. Ileah

9. If a conflict arises, plan to solve the issue as soon as possible. This is best done using synchronous (Zoom, Google Hangouts) communication rather than asynchronous (email). If the group is unable to resolve the conflict, seek advice from the instructional team. Yes, will do.

10. Team's preferred communication method: Text, email, zoom

11. Team's preferred meeting times and frequency: Once a week to every other week; we take other classes together, so we are in contact often and are always willing to meet whenever.

12. Team's preferred method for tracking progress: We plan to have each milestone finished a week ahead of the due date, so we can make changes if we need to ahead of time.

13. Point person for contacting course facilitators with questions: Amber

Milestone #2

1. Description of dataset

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

The data source is from the California Smokers Cohort (CSC) 2011 through UC San Diego. These data are from a survey designed to investigate factors associated with tobacco quitting behaviors, sponsored/funded by CDPH.

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

Through these data, CDPH will better understand tobacco use and behaviors among smokers in California and be able to design an implementation strategies in high-risk communities to increase quitting behaviors. As we go through the milestones, the elements we use may change, but we foresee combining each data set and exploring how race, income, and cigarette brand affect heart disease outcomes. Specifically, creating a visualization to explore how these variables interact with each other and making recommendations to CDPH based on that.

2. Import statement

Utilize function arguments to control relevant components (i.e. change column types, column names, missing values, etc.)

```r
#change columns to all lowercase
clean_smoker_data <- clean_names(smoker_data)

#change column names (the elements we want to use only)
new_smoker_data <- rename(clean_smoker_data, sex = rightsex,
                          smoking_status = smokstat, cig_brand = smkbrand)

#subset the columns we want (elements we want to use only)
updated_smoker_data <- select(new_smoker_data, c("sex", "smoking_status",
                        "cig_brand"))

#change the DO NOT READ values to NA
updated_smoker_data[updated_smoker_data == "(DO NOT READ) Refused"] <- NA
updated_smoker_data[updated_smoker_data == "(DO NOT READ) Don't know"] <- NA

#change columns to all lowercase
clean_race_data <- clean_names(race_data)

#change column names (the elements we want to use only)
new_race_data <- rename(clean_race_data, heart_disease=heartdis,
white=race01, black=race02, japanese=race03, chinese=race04,
filipino=race05, korean=race06, asian_or_pacific_islander=race07,
mexican=race09,
hispanic_latino=race10, other=race11, vietnamese=race12,
asian_indian=race13, refused=race14, dont_know=race15,
americanindian_or_alaskannative=race08)
```

```r
#subset the columns we want (elements we want to use only)
updated_race_data <- select(new_race_data, c("heart_disease", "income", "white",
"black", "japanese", "chinese", "filipino", "korean",
"asian_or_pacific_islander", "americanindian_or_alaskannative", "mexican",
"hispanic_latino", "other", "vietnamese", "asian_indian", "refused",
  "dont_know"))

#change the DO NOT READ values to NA
updated_race_data[updated_race_data == "(DO NOT READ) Refused"] <- NA
updated_race_data[updated_race_data == "(DO NOT READ) Don't know"] <- NA
```

3. 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.

This milestone is focused on: rightsex (sex), smokstat (smoking status), smkbrand (brand of cigarettes), race (race), income (income), and heartdis (heart disease status).

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

```r
#smoker data set
str(updated_smoker_data)

#race data set
str(updated_race_data)
```

Identify the desired type/format for each variable—will you need to convert any columns to numeric or another type?

Both data sets were assessed and it was determined there were not any columns that needed to be converted to another data type at this time.

4. Provide a basic description of the 5+ data elements

```r
#first data set
summary(updated_smoker_data)
#second data set
summary(updated_race_data)
```

Milestone #3

```r
raceclean <- clean_names(race_data)
smokerclean <- clean_names(smoker_data)
```

1. Subset row or columns as needed in race data

```r
#subset columns from race dataset
race <- raceclean %>% select(id, race01, race02, race03, race04, race05,
race06, race07, race08, race09, race10, race11, race12, race13, race14, race15,
income,vereduc, wgtinlbs, htinfeet, goodhlth, harmhlth, smokalone, acqsmoke,
act10min,drinkfiv, heartdis, diabetes)
```

```r
#rename columns
racedata <- race %>% rename(heart_disease=heartdis,
white=race01, black=race02, japanese=race03, chinese=race04,
filipino=race05, korean=race06, asian_or_pacific_islander=race07,
mexican=race09,
hispanic_latino=race10, other=race11, vietnamese=race12,
asian_indian=race13, refused=race14, dont_know=race15,
americanindian_or_alaskannative=race08, school_level = vereduc,
smoking_harms_health = harmhlth, how_many_people_smoke = acqsmoke,
physically_active_for_10min = act10min,
days_had_4ormore_drinks_inrow = drinkfiv)
```

```r
#recode race columns into one column
racedata1 <- racedata %>% mutate(race = case_when(white == "Yes" ~ "white",
black == "Yes" ~ "black", japanese == "Yes" ~ "japanese",
chinese == "Yes" ~ "chinese", filipino == "Yes" ~ "filipino",
korean == "Yes" ~ "korean",
asian_or_pacific_islander == "Yes" ~ "asian_or_pacific_islander",
mexican == "Yes" ~ "mexican", hispanic_latino == "Yes" ~ "hispanic_latino",
other == "Yes" ~ "other", vietnamese == "Yes" ~ "vietnamese",
asian_indian == "Yes" ~ "asian_indian", refused == "Yes" ~ "refused",
dont_know == "Yes" ~ "dont know",
americanindian_or_alaskannative == "Yes" ~ "americanindian_or_alaskannative"))
```

```r
#subset race data that is ready to combine with smoker data
recodedrace_data <- racedata1 %>% select(id, race, income, school_level, wgtinlbs,
htinfeet, goodhlth, smoking_harms_health, smokalone, how_many_people_smoke,
physically_active_for_10min,
days_had_4ormore_drinks_inrow, heart_disease,
diabetes)
```

2. Subset row or columns as needed in smoker data

```r
#subset columns from smoker data
smokerdata <- smokerclean %>% select(psraid, rightsex, smokstat, howmany,
smok6num, smok6uni, smkbrand, smk1age, smkage)
```

```r
#rename columns in smoker data
datasmoker <- smokerdata %>% rename(sex = rightsex, smoking_status = smokstat,
cig_brand = smkbrand, cigs_per_day = howmany,
how_long_smoking_daily = smok6num, unit_of_time_smoking_daily = smok6uni,
age_when_first_smoked = smk1age, age_when_daily_smoking_began = smkage,
id = psraid)
```

3. Combine both data sets

```r
#make id variables the same type in both datasets to combine datasets by id
newracedata <- cbind(recodedrace_data, i_d = datasmoker$id)
newracedata1 <- newracedata %>% select(i_d, race, income, school_level, wgtinlbs,
htinfeet, goodhlth, smoking_harms_health, smokalone, how_many_people_smoke,
physically_active_for_10min,
days_had_4ormore_drinks_inrow, heart_disease,
diabetes) %>% rename(id = i_d)
```

```r
#checking column types
typeof(datasmoker$id)
typeof(newracedata1$id)
```

```r
#combine datasets by ID
combineddata <- inner_join(newracedata1, datasmoker, by = c("id"))
```

```r
#change DO NOT READ values to NA data values because .......
combineddata[combineddata == "(DO NOT READ) Refused"] <- NA
combineddata[combineddata == "(DO NOT READ) Don't know"] <- NA
```

4. Create and clean two new variables needed to analysis

```r
#variable 1, bmi
new_variables_data_1 <- combineddata %>%
  mutate(bmi = (wgtinlbs * 0.45359237) / (htinfeet * 0.304)^2)

#make packs per day variable to create packs per year variable
#packs per day is cigs_per_day was divided by 20 because there are 20 cigarettes in 1 pack
#variable 2
new_variables_data_2 <- new_variables_data_1 %>%
  mutate(cigs_per_day = as.numeric(cigs_per_day),
         packs_per_day = (cigs_per_day / 20))
```

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

```r
#variable 3
new_variables_data_2 <- new_variables_data_2 %>%
  mutate(how_long_smoking_daily = as.numeric(how_long_smoking_daily))

recoding_variables <- new_variables_data_2 %>%
mutate(how_long_smoking_daily = case_when(unit_of_time_smoking_daily == "Days" ~
how_long_smoking_daily/365, unit_of_time_smoking_daily == "Months" ~
how_long_smoking_daily/12, unit_of_time_smoking_daily == "Years" ~
  round(how_long_smoking_daily, 2)))

final_variables_data <- recoding_variables %>%
  mutate(how_long_smoking_daily = as.numeric(how_long_smoking_daily),
         pack_years = round((packs_per_day) * (how_long_smoking_daily), 2))

median_pack_years <- median(final_variables_data$pack_years, na.rm=T)
print(median_pack_years)
```

```
## [1] 17
```

```
mean_pack_years <- mean(final_variables_data$pack_years, na.rm=T)
print(mean_pack_years)
```

```
## [1] 21.45533
```

```
sd_pack_years <- sd(final_variables_data$pack_years, na.rm=T)
print(sd_pack_years)
```

```
## [1] 17.93978
```

5. Data dictionary for final dataset

- **Variable 1**
  - *Name:* cig_brand
  - *Data type:* Character
  - *Description:* This variable contains the name of the cigarette brand the participant reported smoking
- **Variable 2**
  - *Name:* age_when_daily_smoking_began
  - *Data type:* Character
  - *Description:* This variable contains the age of the participant when they began to smoke cigarettes on a regular basis
- **Variable 3**
  - *Name:* age_when_first_smoked
  - *Data type:* Character
  - *Description:* This variable contains the age when the participant first smoked their first whole cigarette
- **Variable 4**
  - *Name:* bmi
  - *Data type:* Double
  - *Description:* This variable contains the participant's BMI, calculated as kg/mˆ2

6. Table with descriptive statistics

```
#BMI
summary(final_variables_data$bmi, na.rm = T)
sd(final_variables_data$bmi, na.rm = T)
```

```
#Pack Years
summary(final_variables_data$pack_years, na.rm = T)
sd(final_variables_data$pack_years, na.rm = T)
```

```
#Packs Per Day
summary(final_variables_data$packs_per_day, na.rm = T)
sd(final_variables_data$packs_per_day, na.rm = T)
```

```
#Cigs Per Day
summary(final_variables_data$cigs_per_day, na.rm = T)
sd(final_variables_data$cigs_per_day, na.rm = T)
```

```
library(kableExtra)
```

```
##
## Attaching package: ’kableExtra’

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

```
df_ofdescriptive_stats <- data.frame("Variable" =
c("Pack Years", "Cigs Per Day", "BMI", "Packs Per Day"), "Mean" =
c(21.46, 13.89, 33.38, 0.69), "Standard Deviation" =
c(17.94, 9.30, 10.67, 0.47), "Range" = c(119.9, 59, 161.29, 2.95))

kable(df_ofdescriptive_stats, booktabs = T, longtable=T, col.names =
        c("Variable", "Mean", "SD", "Range"), align= 'lcccc',
caption= "Descriptive Statistics of Four Variables from 2011 California
Smokers Cohort")
```

Table 1: Descriptive Statistics of Four Variables from 2011 California Smokers Cohort

Variable	Mean	SD	Range
Pack Years	21.46	17.94	119.90
Cigs Per Day	13.89	9.30	59.00
BMI	33.38	10.67	161.29
Packs Per Day	0.69	0.47	2.95

Milestone #4

1. Plot that compares average number of pack-years on heart disease.

```
data_visual_one <- final_variables_data %>%
  select(c("pack_years", "heart_disease", "diabetes"))

#Box plot of pack years by heart disease
boxplot_one <- boxplot(pack_years~heart_disease, data = data_visual_one,
main="Figure 1: Pack-Years and Heart Disease",
xlab= "Heart Disease Reported During Study", ylab="Cigarette Pack Years")
```

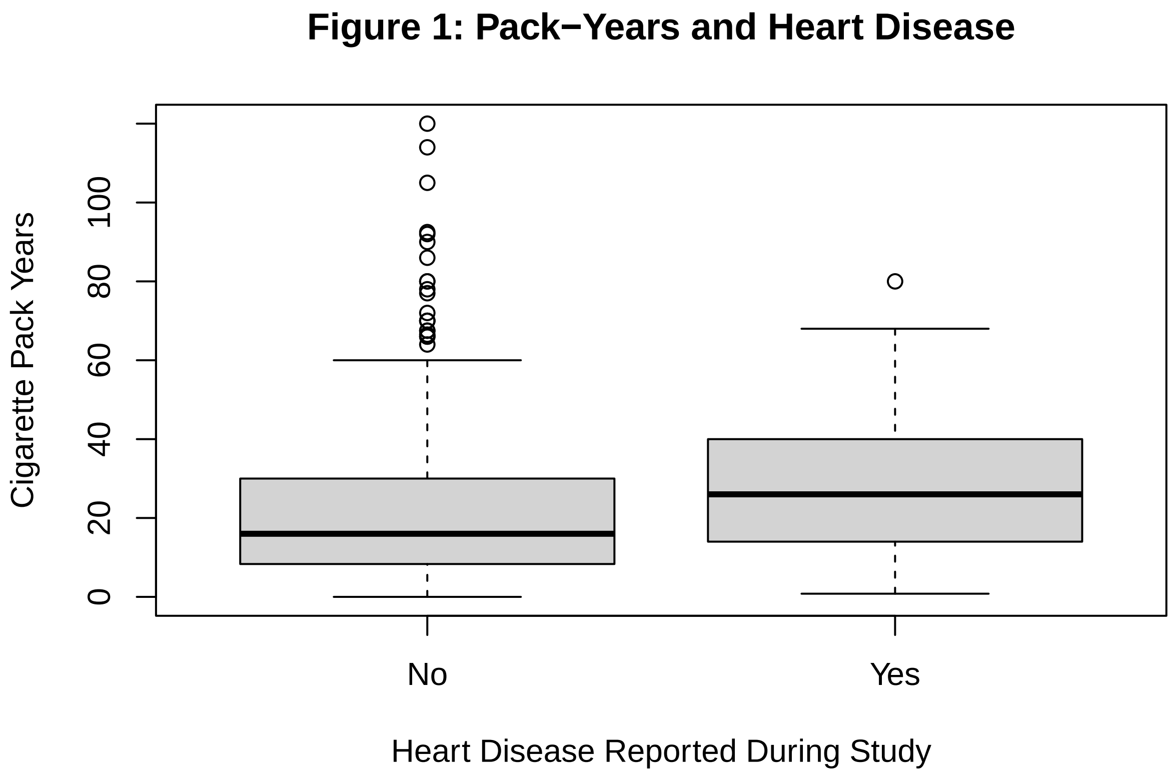

Figure 1. Based on this plot, a participant in this study who reported having Heart Disease, on average, had a wider range of cigarette pack years and higher median number of pack years when compared to another participant in the same study without heart disease.

2. Plot that compares average number of pack-years on diabetes.

```
#Box plot of pack years by diabetes
boxplot_two <- boxplot(pack_years~diabetes, data = data_visual_one,
main="Figure 2: Pack-Years and Diabetes",
xlab= "Diabetes Reported During Study", ylab="Cigarette Pack Years")
```

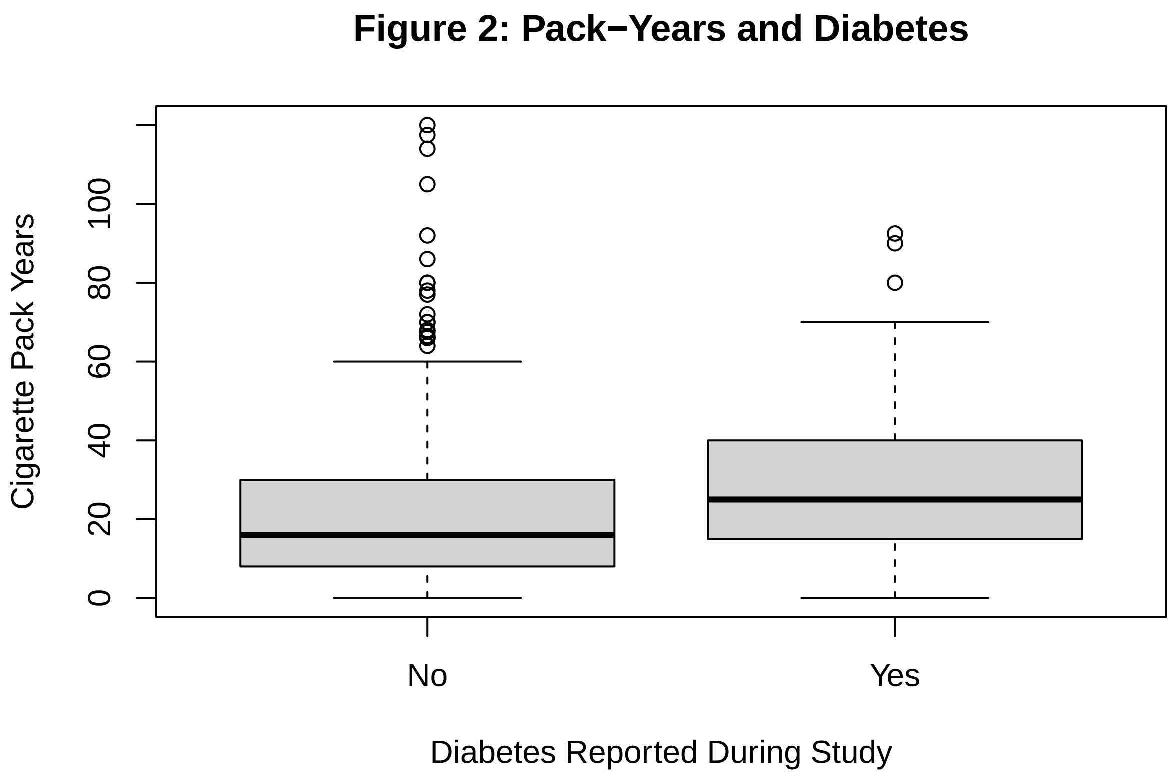

Figure 2. Based on this plot, a participant in this study who reported having Diabetes, on average, had a wider range of cigarette pack years and higher median number of pack years when compared to another participant in the same study without Diabetes.

3. Plot that compares smoking status, race, and income.

```
visualization2 <- final_variables_data %>% select(id, race, income,
heart_disease, diabetes, sex, smoking_status, age_when_first_smoked)

visualization2 <- visualization2 %>% mutate(income = case_when(income == "$20,000 or less" ~ 1,
income == "$20,001 to $30,000" ~ 2, income == "$30,001 to $50,000" ~ 3,
income == "$50,001 to $75,000" ~ 4, income == "$75,001 to $100,000" ~ 5,
income == "$100,001 to $150,000 or" ~ 6, income == "Over $150,000?" ~ 7)) %>%
mutate(income = as.numeric(unlist(income)))

library(ggplot2)
ggplot(visualization2, aes(x = race, y = income)) +
geom_count(aes(color = ..prop.., size = ..prop.., group = 1)) +
guides(color = 'legend') + theme_classic() + labs(x = "Race",
y = "Income",
title = "Figure 3: Proportion of Daily Smokers Versus Non-Daily Smokers by Race & Income",
subtitle = "2011 California Smokers Cohort",
caption = "Note: y-axis values were recoded to 1 = $20,000 or less, 2: = $20,001 to $30,000,
3 = $30,001
to $50,000, 4 = $50,001 to $75,000, 5 = $75,001 to $100,000, 6 = $100,001 to
$150,000, 7 = Over $150,000") + facet_wrap(~smoking_status) +
theme(axis.text.x = element_text(angle = 90)) + scale_size_area() + theme(
plot.title = element_text(color = "black", size = 10, face = "bold",
hjust = 0.5), plot.caption = element_text(size = 7, hjust = 1,
color = "blue"), plot.subtitle = element_text(size = 9, color = "blue", face = "italic",
hjust = 0.5)) + theme(legend.position = "right")
```

```
## Warning: Removed 189 rows containing non-finite values (stat_sum).
```

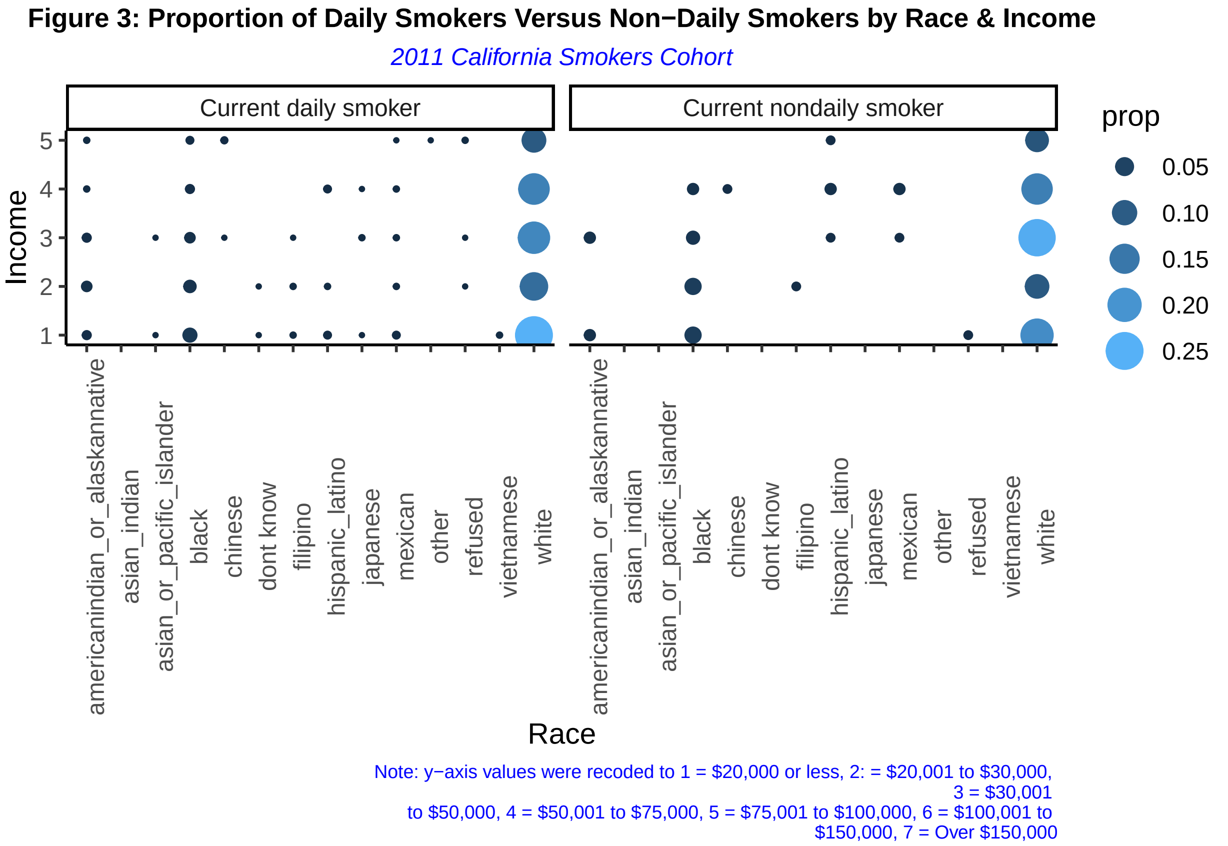

Figure 3: Proportion of Daily Smokers Versus Non−Daily Smokers by Race & Income

2011 California Smokers Cohort

Note: y−axis values were recoded to 1 = $20,000 or less, 2: = $20,001 to $30,000, 3 = $30,001 to $50,000, 4 = $50,001 to $75,000, 5 = $75,001 to $100,000, 6 = $100,001 to $150,000, 7 = Over $150,000

Figure 3. Based on this plot, the largest proportion of current daily smokers in this study are white and earn $20,000 dollars or less on an annual basis.

4. Plot that compares diabetes, race, and income.

```
visualization3 <- visualization2 %>% drop_na() %>% mutate(income =
recode(income, "1" = "$20,000 or Less", "2" = "$20,001 to $30,000",
"3" = "$30,001 to $50,000", "4" = "$50,001 to $75,000", "5" = "$75,001 to
$100,000"))
```

```
ggplot(visualization3, aes(y = race, fill = diabetes)) + geom_bar() +
facet_wrap(~income, scales = "free_x") + theme_minimal() + labs(x = "Count",
y = "Race", fill = "Diabetes",
title = "Figure 4: Counts of Diabetes Outcomes Per Race and Income Status",
subtitle = "2011 California Smokers Cohort") +
  scale_fill_manual(values = c("#d8b365", "#5ab4ac")) + theme(
plot.title = element_text(color = "black", size = 11, face = "bold",
hjust = 0.5), plot.subtitle = element_text(size = 9, color = "#5ab4ac",
                                           face = "bold", hjust = 0.5))
```

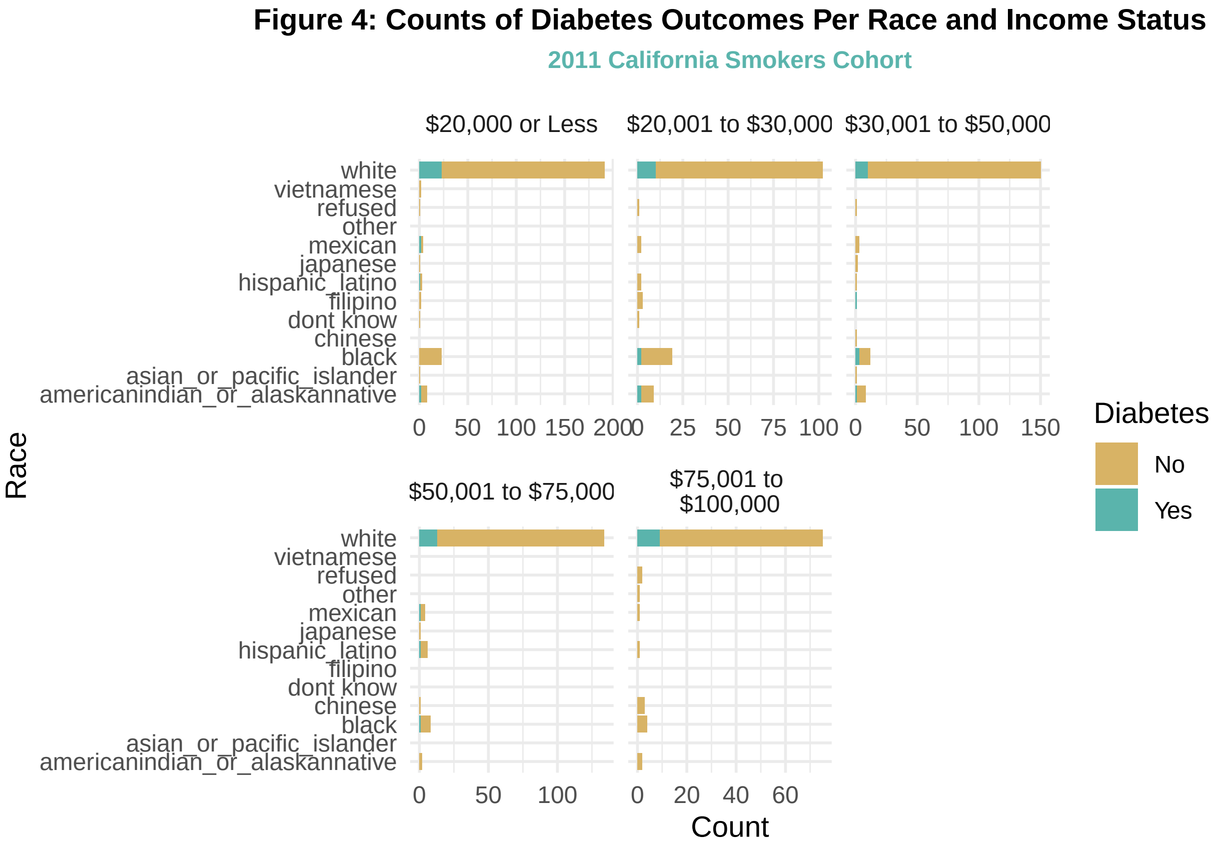

Figure 4: Based on this plot, white participants were more likely to have diabetes than any other race; however, diabetes was not a prevalent disease in this group of particiapnts.

5. Plot that compares heart disease, race, and income.

```
ggplot(visualization3, aes(y = race, fill = heart_disease)) + geom_bar() +
  facet_wrap(~income, scales = "free_x") + theme_minimal() + labs(x = "Count",
y = "Race", fill = "Heart Disease",
title = "Figure 5: Counts of Heart Disease Outcomes Per Race and Income Status",
subtitle = "2011 California Smokers Cohort") +
  scale_fill_manual(values = c("#C3D7A4", "#52854C")) + theme(
plot.title = element_text(color = "black", size = 11, face = "bold",
hjust = 0.5), plot.subtitle = element_text(size = 9, color = "#52854C",
                                           face = "bold", hjust = 0.5))
```

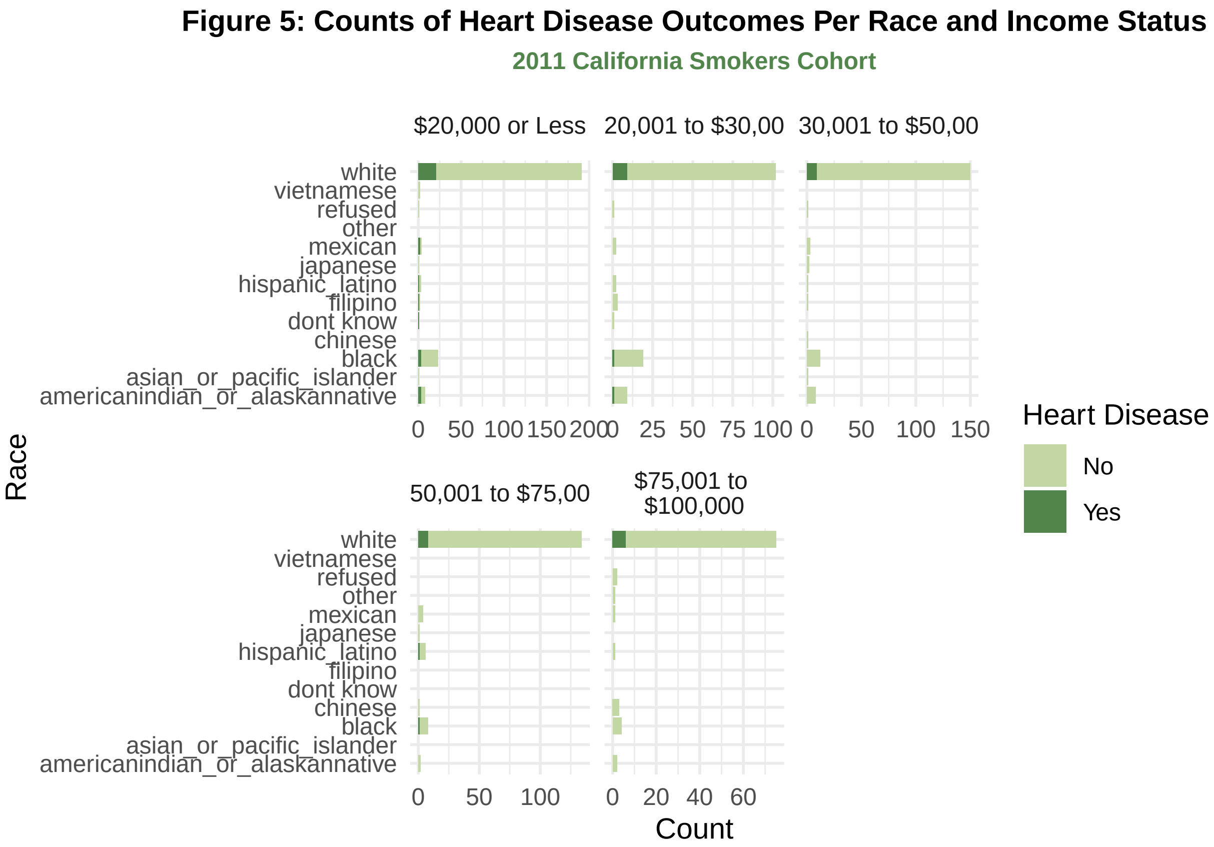

Figure 5: Based on this plot, participants who were more likely to have heart disease were white and made below $30,001 on an annual basis. Heart disease was not a prevalent disease in this group of participants.