# Milestone #6

Amber Morris & Ileah Rios

## Problem Statement

The California Department of Public Health (CDPH) is interested in investigating tobacco use and behaviors among smokers in California to better design and implement strategies in high-risk communities to increase quitting behaviors. Specifically, this report explores how race, income, and pack-years affect heart disease and diabetes outcomes among smokers in California.

## Methods

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.

To clean these two data sets, the tidyverse package was utilized. Moreover, these data were subsetted to only include all of the races identified in the survey (race 01-15), income, weight, height, education level, if participants believed smoking harms health status, days participants had 4 or more drinks in a row, how often participants were physically active for >= 10 minutes, if participants believed they were in good health or not, if participants reported having heart disease and/or diabetes, if participants smoked alone, how many people participants smoked with, sex, smoking status, cigarette brand, cigarettes per day, unit of time smoking daily, age when daily smoking began, age when first smoked, and how long smoking daily. Additionally, values that were input as "DO NOT READ" or "REFUSED" indicating that the participant did not answer the question were recoded as NA.

From there, these columns were renamed to better reflect the variables. For example, "race01" became "white", "race02" became "black", "veredu" became "school_level", "smok6num" became "how_long_smoking_daily", etc.. Because of small numbers for some race categories, we grouped certain races together such as Mexican and Hispanic/Latino were grouped together. A full list of the specific 29 variables that were renamed can be found in the RMD file. From there, the two data sets were joined together on a unique participant ID.

Furthermore, we created new variables from the existing variables to aid in the analysis. Specifically, the "bmi" variable was created from "wgtinlbs" and "htinfeet" variables. Next, the "packs_per_day" variable was created from using "cigs_per_day" / 20. Lastly, the "packs_per_year" variable was created after all of the values in "unit_of_time_smoking_daily" were recoded to be in years. We included these new variables in a descriptive statistics table using the kable package (Table 1).

In order to visualize the average number of pack-years on the two disease outcomes (figure 1), we created a boxplot using the boxplot function. Figure 1 compares occurrence of heart disease on the x-axis and cigarette pack-years on the y-axis. Figure 2 compares the occurrence of diabetes on the x-axis and cigarette pack-years on the y-axis.

To visualize and compare smoking status, race, and income we created Tabel 2 with the counts of participants by race, smoking status, and income. We also created a bubble plot (figure 3) using ggplot2 package and the geom_count function to obtain the proportion of counts. We used function geom_count to count the

number of observations at each location, because we have discrete data and overplotting. In this graph, we changed the counts to proportions. We started by recoding income levels as 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, and 7 = Over $150,000. Race was inputted as our x-axis and income as our y-axis, and we used the facet_wrap function to facet by our categorical variable smoking_status.

# Results

Through the analysis, it was discovered that the mean and standard deviation amount of pack-years in this study among participants was 21.46 and 17.94, respectively. Additionally, the visualizations provided in this report suggest that heart disease and diabetes outcomes were more commonly reported in participants who had a higher median number of pack-years when compared to participants who did not report heart disease and diabetes outcomes. Lastly, the analysis indicated that the participants who reported their race as white and income as $20,000 or less on an annual basis in this study were the largest proportion of current daily smokers. Additionally, no participants made over $100,001 dollars.

**Table 1**

Table 1: Descriptive Statistics 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 |

**Table 2**

Table 2: Counts of Participants by Race, Smoking Status, and Income from 2011 California Smokers Cohort

| Race | Smoking Status | $20,000 or Less | $20,001 to $30,000 | $30,001 to $50,000 |
|---|---|---|---|---|
| AmericanIndian/AlaskanNative | Current daily smoker | 6 | 9 | 6 |
| AmericanIndian/AlaskanNative | Current nondaily smoker | 2 | NA | 2 |
| Asian/PacificIslander | Current daily smoker | 4 | NA | 4 |
| Black | Current daily smoker | 18 | 14 | 9 |
| Black | Current nondaily smoker | 5 | 5 | 3 |
| Don't Know | Current nondaily smoker | 1 | NA | NA |
| Filipino | Current daily smoker | 2 | 2 | 1 |
| Hispanic/Latino | Current daily smoker | 7 | 4 | 2 |
| Unknown | Current daily smoker | 1 | 1 | NA |
| White | Current daily smoker | 168 | 90 | 119 |
| White | Current nondaily smoker | 23 | 12 | 31 |
| Don't Know | Current daily smoker | NA | 1 | 1 |
| Filipino | Current nondaily smoker | NA | 1 | NA |
| Hispanic/Latino | Current nondaily smoker | NA | NA | 2 |
| Asian/PacificIslander | Current nondaily smoker | NA | NA | NA |

| | | | | |
|---|---|---|---|---|
| Other | Current daily smoker | NA | NA | NA |

**Figure 1**

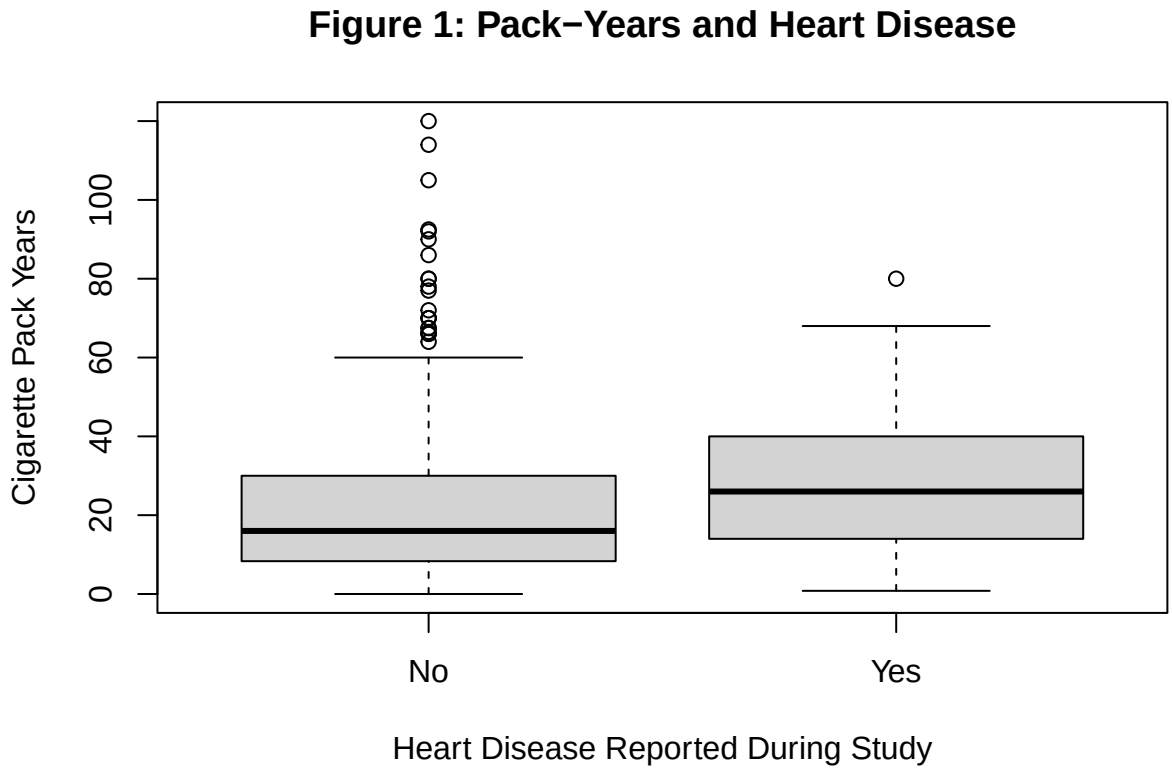


Figure 1: Based on this plot, a participant in this study who reported having Heart Disease, on average, had a wider range of cigarette pack years and higher median number of pack years when compared to another participant in the same study without heart disease.

**Figure 2**

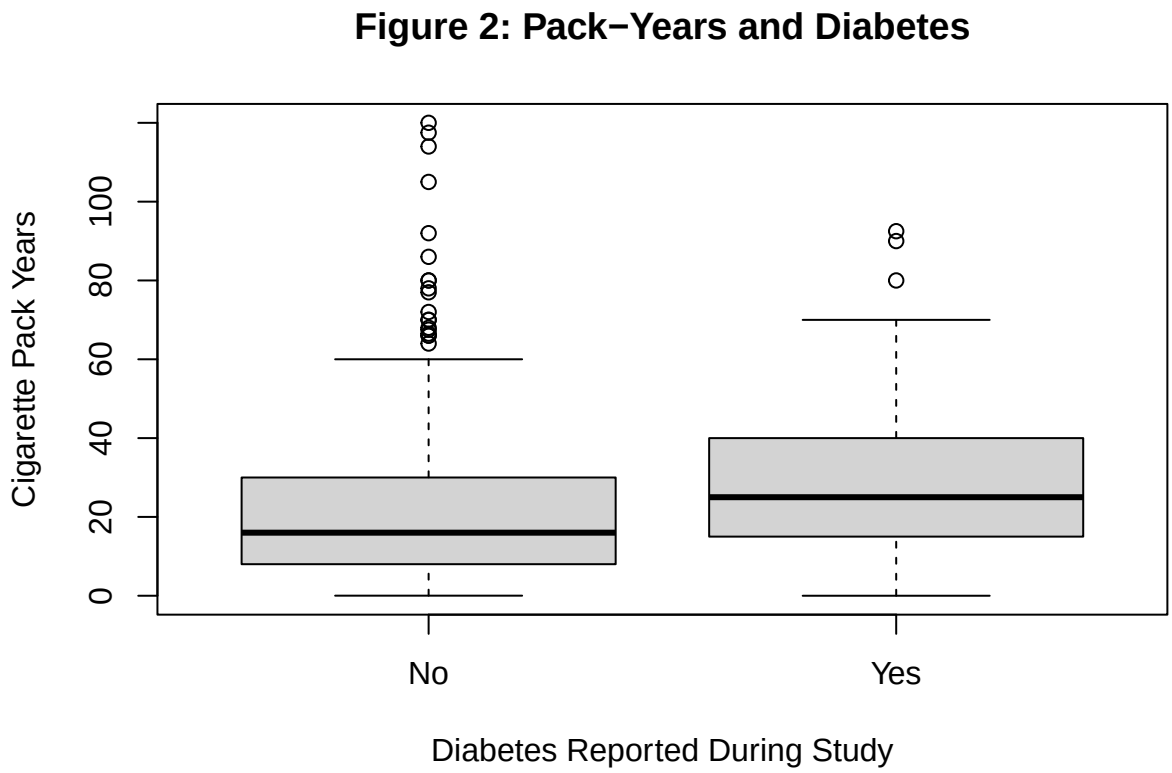


Figure 2. Based on this plot, a participant in this study who reported having Diabetes, on average, had a wider range of cigarette pack years and higher median number of pack years when compared to another participant in the same study without Diabetes.

**Figure 3**

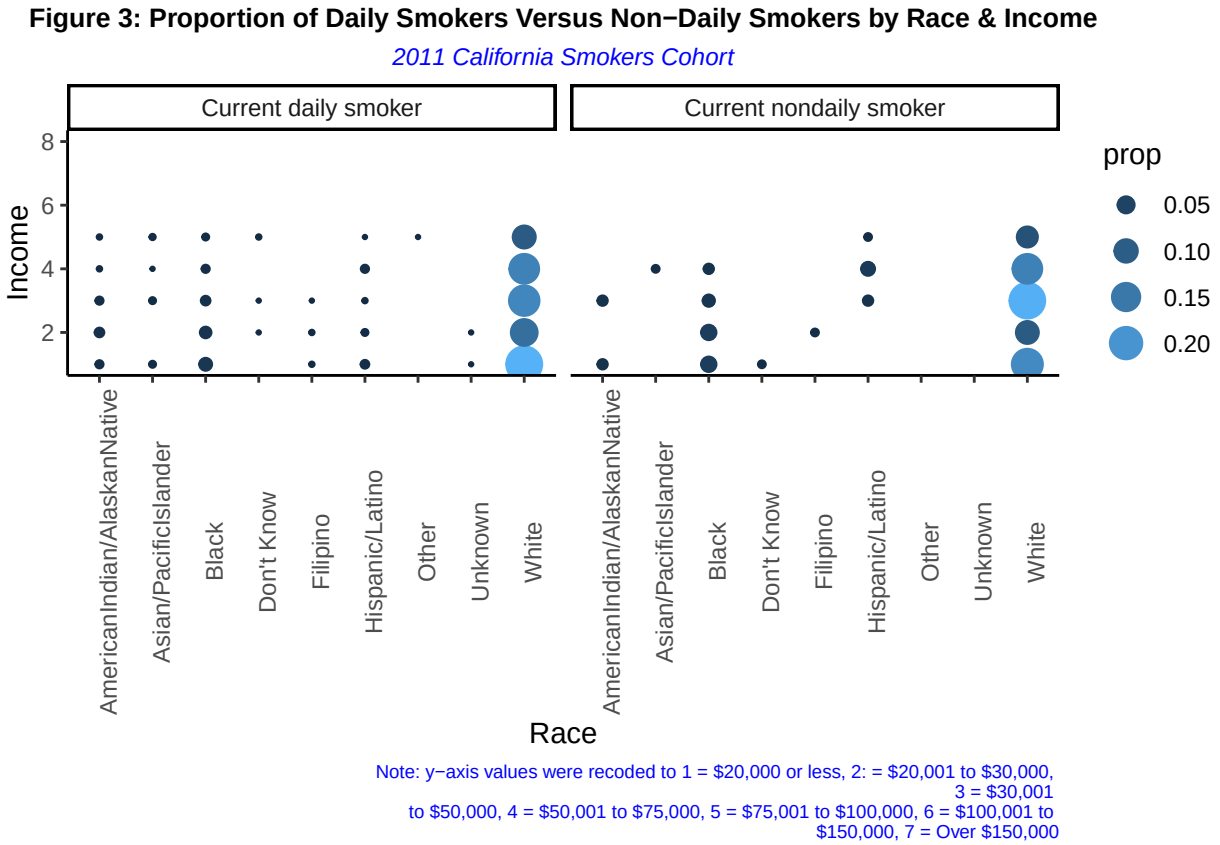


Figure 3. Based on this plot, the largest proportion of current daily smokers in this study are white and earn $20,000 dollars or less on an annual basis.

## Discussion

Based on the results, CDPH should focus on developing policies that help regulate the price of tobacco products. Since it was discovered that the largest proportion of current daily smokers are low-income white people, a policy in specific-counties establishing minimum-pack and minimum-price for tobacco retailers could help prevent the tobacco industry from targeting these price-sensitive customers. If CDPH were to help institute the passage of county-wide price regulation policies, low-income people would be less likely to buy tobacco products because the products would not be as readily accessible to them due to the price point. Additionally, CDPH should work on developing anti-tobacco messaging in communities that are composed of mostly white people after reviewing commonly-used tobacco ad tactics in these communities. In other words, the tobacco industry typically targets white, usually rural, communities with ads showing images of farms, cowboys, blue-collar workers, firefighters, police officers, etc. all using tobacco products because 'it's a way of life' which can be quite influential in initiating smoking among these community members. If CDPH were to design anti-tobacco messaging on local radios, highway billboards, tv commercials (based on geofencing in these predominantly white communities), it would draw attention to the predatory ad behavior used by the tobacco industry and white community members would be more likely to recognize it after and less likely to 'fall' for the messaging and initiate usage of these deadly tobacco products. Lastly, after examining cigarette pack-years calculation and uncovering the median consumption of these products, the participants in this study who reported having Diabetes and/or Heart Disease also consume more cigarettes. Again, this suggests that this higher consumption of these products could potentially cause and/or increase risk

for Diabetes and/or Heart Disease. As such, CDPH should consider mentioning this relationship in future anti-tobacco messaging campaigns in order to pique and further inform communities impacted.

These data have limitations as participant home and buying locations were not collected during the study. In other words, the place a participant lives and buys tobacco products can greatly influence the prevalence and quantity of tobacco consumption and specific products (cigarettes, chewing tobacco, synthetic tobacco, vapes, cigars, etc). It is well documented in various pieces of peer-reviewed literature that rural communities specifically are targeted by tobacco industry couponing, ads, and other influences which increase tobacco consumption in these areas, so it makes sense from an analysis point of view that white, low-income participants in this study were the largest proportion of smokers.