PHW251 Data Project Milestone 6

Group W: Nakisa Golabi, Joanna Liang, Desmond Gill

12/06/2022

PROBLEM STATEMENT:

Tobacco use and smoking has been shown to be associated with negative health outcomes. It is important to further analyze the behavioral and social factors associated with tobacco use and smoking to identify the demographical and environmental factors of smokers in California. There is still a need for further research due to the fact that not every person is affected by specific factors in the same way. In order to reduce disease outcomes associated with smoking cigarettes, we need to understand the specific pathways that lead to high-risk individuals and communities. Our goal is to utilize the information found to help implement strategies that decrease smoking cigarettes and dependency.

METHODS:

The data source used was from the California Department of Public Health 2011, part of the UC San Diego library collections: (https://library.ucsd.edu/dc/object/bb6282371f). Creation date is 2011-2012 and it is a cross-sectional study looking at tobacco use and behaviors among Californians. The dataset provides the raw data needed to analyze the participant behaviors. We are interested in seeing the effects of certain individual behaviors & how this relates to smoking status as well as disease status (heart disease). Methods used include tidying and cleaning data sources to be able to combine the two datasets into one. We eliminated N/As, 0s, negative numbers, made all columns lower case with _ instead of spaces. We cleaned 11 variables; cigarette brand, smoking status, level of smoking status, where the participant bought cigarettes, smoking frequency compared to last year, heart disease, mental illness, diabetes, asthma, smoker onset, pack years. We also created two new variables; 1) Smoker onset (the difference between the age when they started smoking regularly and the age of the participant's first cigarette), 2) Pack years (the product of the number of packs of cigarettes smoked per day and the years a person has smoked). As a team, we worked cohesively to implement the tools and concepts we learned in this course to wrangle the data and create three visualizations that help address the problem statement.

RESULTS:

The results of Table 1 compare the average number of pack-years (the product of the number of packs of cigarettes smoked per day and the years a person has smoked) by the disease outcomes of asthma, heart disease, diabetes, and mental illness. Those with diabetes had the highest mean and those with mental illness had the lowest mean. Figure 1 compares the cigarette brands purchased by each smoker with their racial identification (self-reported mixed or single race), in the CDPH 2011 California Smokers Cohort. By knowing that out of 712 total smokers, Marlboro was the highest purchased cigarette brand for both self-reported mixed race smokers (at 14 counts) and self-reported single race smokers (at 290 counts), we can target this specific cigarette brand to help reduce certain disease outcomes (heart disease, asthma, mental illness, diabetes). Figure 2 compares smoker onset (the difference between the age when they started smoking regularly and the age of the participant's first cigarette) with their racial identification (self-reported mixed or single race), in the CDPH 2011 California Smokers Cohort. The medians for smoker onset were 2.00 for smokers with self-reported mixed and 2.00 for single race.

Table 1: Average Pack-Years Per Disease Outcome

Outcomes	Pack-Years
Asthma	23.93
Heart Disease	28.83
Mental Illness	18.94
Diabetes	30.10

TABLE COMPARING THE AVERAGE NUMBER OF PACK-YEARS AND THE DISEASE OUTCOMES

Table 1 compares the average number of pack-years (the product of the number of packs of cigarettes smoked per day and the years a person has smoked) by the disease outcomes of asthma, heart disease, diabetes, and mental illness. Those with diabetes had the highest mean (30.10) and those with mental illness had the lowest mean (18.94).

PLOT COMPARING RACE AND BEHAVIORAL FACTOR OF CIGARETTE BRANDS PURCHASED

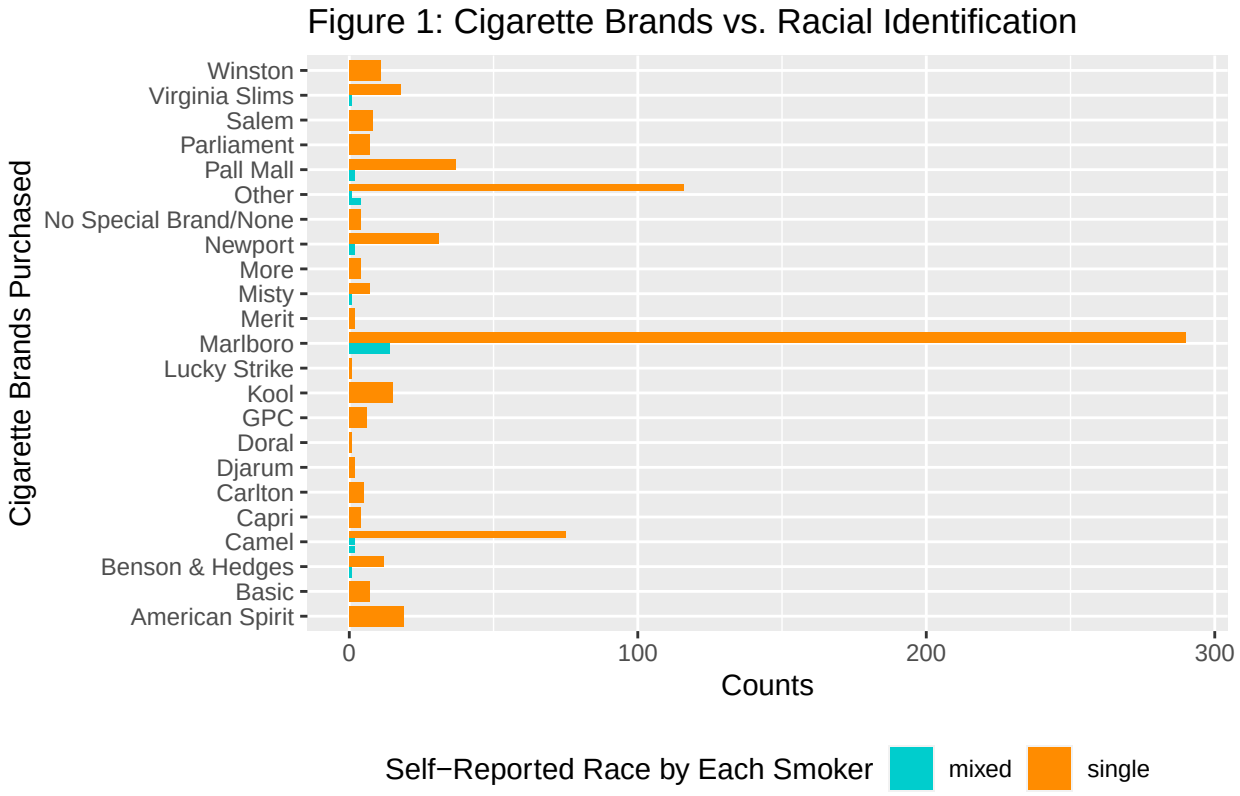

Figure 1 compares the cigarette brands purchased by each smoker with their racial identification (self-reported mixed or single race), in the CDPH 2011 California Smokers Cohort. By knowing that out of 712 total smokers, Marlboro was the highest purchased cigarette brand for both self-reported mixed race smokers (at 14 counts) and self-reported single race smokers (at 290 counts), we can target this specific cigarette brand to help reduce certain disease outcomes (heart disease, asthma, mental illness, diabetes).

PLOT COMPARING RACE AND SMOKER ONSET

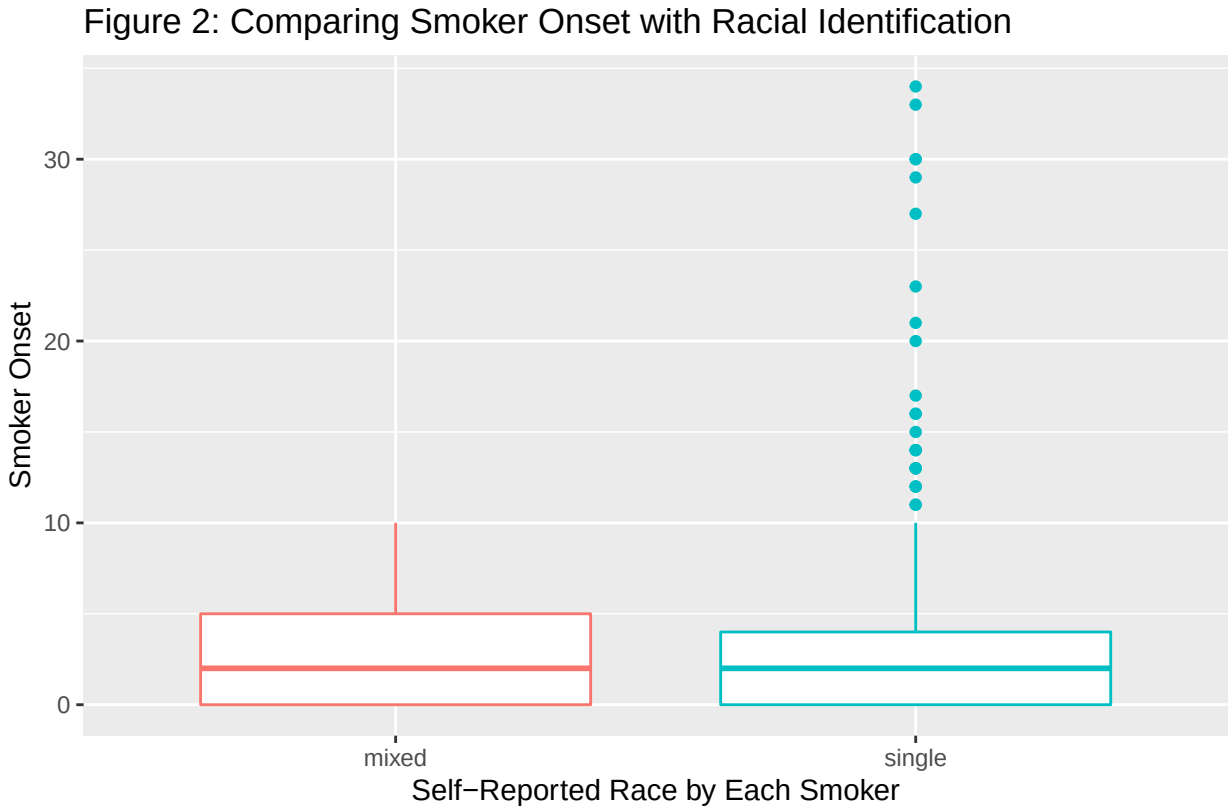

Figure 2 compares smoker onset (the difference between the age when they started smoking regularly and the age of the participant's first cigarette) with their racial identification (self-reported mixed or single race) in the CDPH 2011 California Smokers Cohort, showing the median for smoker onset is 2.00 for both self-reported mixed and single race.

Due to the difference in sub-population size of self-reported mixed at 30 counts and single race at 677 counts, the outliers with higher onset of the single race group did not significantly affect the mean/median.

DISCUSSION:

Our findings in Table 1 suggest that Diabetes has the highest mean pack-years out of our four assessed health outcomes (heart disease, asthma, mental illness, diabetes). Although we are able to conclude this measurement from our analyzed data, this finding is limited and tells us that there is a need to further investigate the association between tobacco use and Diabetes. From Figure 1, we determined that Marlboro is the most popular cigarette brand regardless of racial identification (self-reported mixed or single race). More information around how Marlboro advertises their products may help provide additional context on why this cigarette brand is so popular among smokers. Similarly, regardless of racial identification, the median smoker onset was two years among both self-reported mixed and single race. As shown by these certain behaviors being constant, more insight into smoker behaviors would be beneficial in order to determine why this is occurring. The findings of Figure 2 suggest that further analysis into additional behavioral factors beyond racial identification is needed to help reduce our assessed disease outcomes (heart disease, asthma, mental illness, diabetes). It would also be beneficial to look into social factors that could be affecting these behavioral choices, such as SES, Income, Education, etc.

We chose to include racial identification through self-reported mixed or single race because our goal was to better understand the population to determine which subgroups to target tobacco cessation interventions. Since the data is from a California cohort, we believed this was indicative of a diverse population. However upon analyzing the data, we saw in Figure 1 and Figure 2 an unequal representation of self-reported mixed (at 30 counts) vs single race (at 677 counts). More data is needed to determine the target populations to effectively direct CDPH resources for further development of strategic interventions to increase quitting behaviors and reduce tobacco consumption.