PHW251 Data Project Milestone 6

Nakisa Golabi, Joanna Liang, Desmond Gill

11/28/2022

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

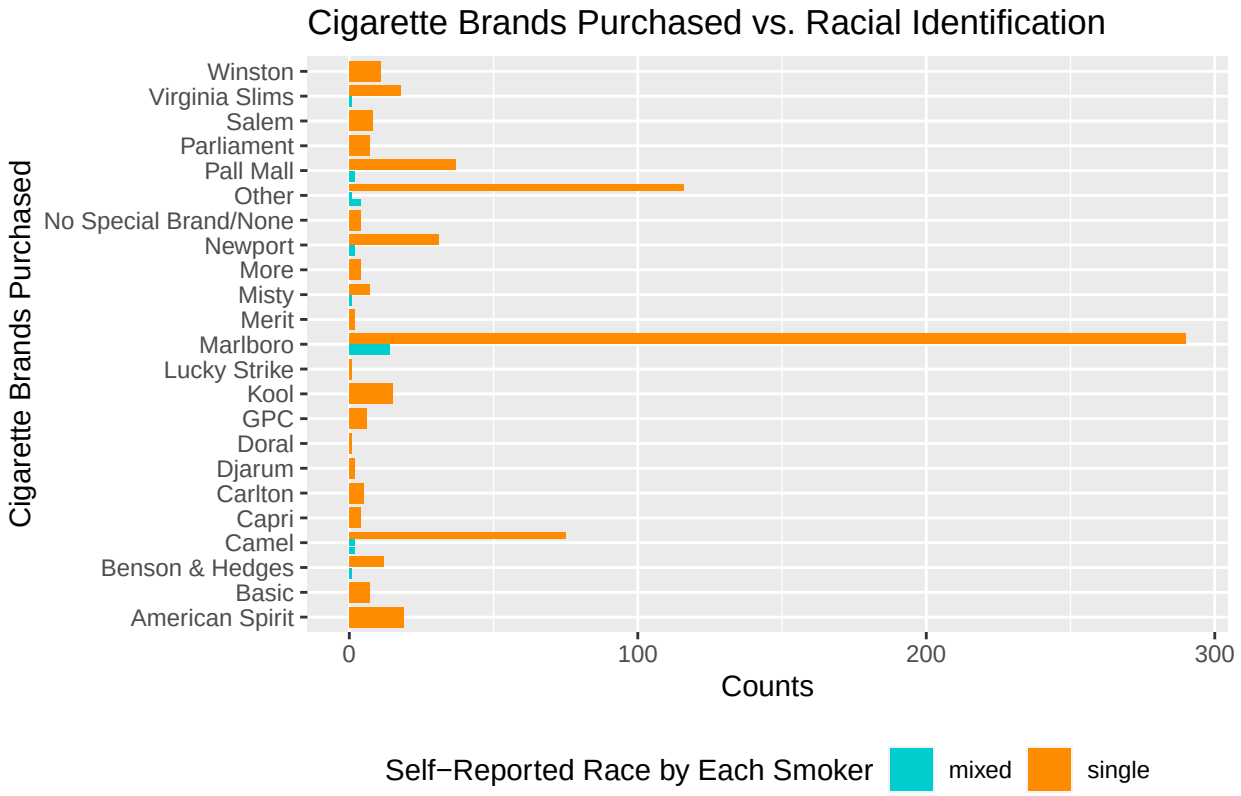
PRINT-QUALITY TABLE THAT SHOWS THE AVERAGE NUMBER OF PACK-YEARS AND THE DISEASE OUTCOMES

```
kable(tibble_py_disease, booktabs = T, digits = 2, escape=F,
                    col.names=c("Outcomes", "Pack-Years"),
                    align="cc",
                    caption = "Average Pack-Years Per Disease Outcome")
```

The table 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 and those with mental illness had the lowest mean.

PLOT COMPARING RACE AND BEHAVIORAL FACTOR OF CIGARETTE BRANDS PURCHASED

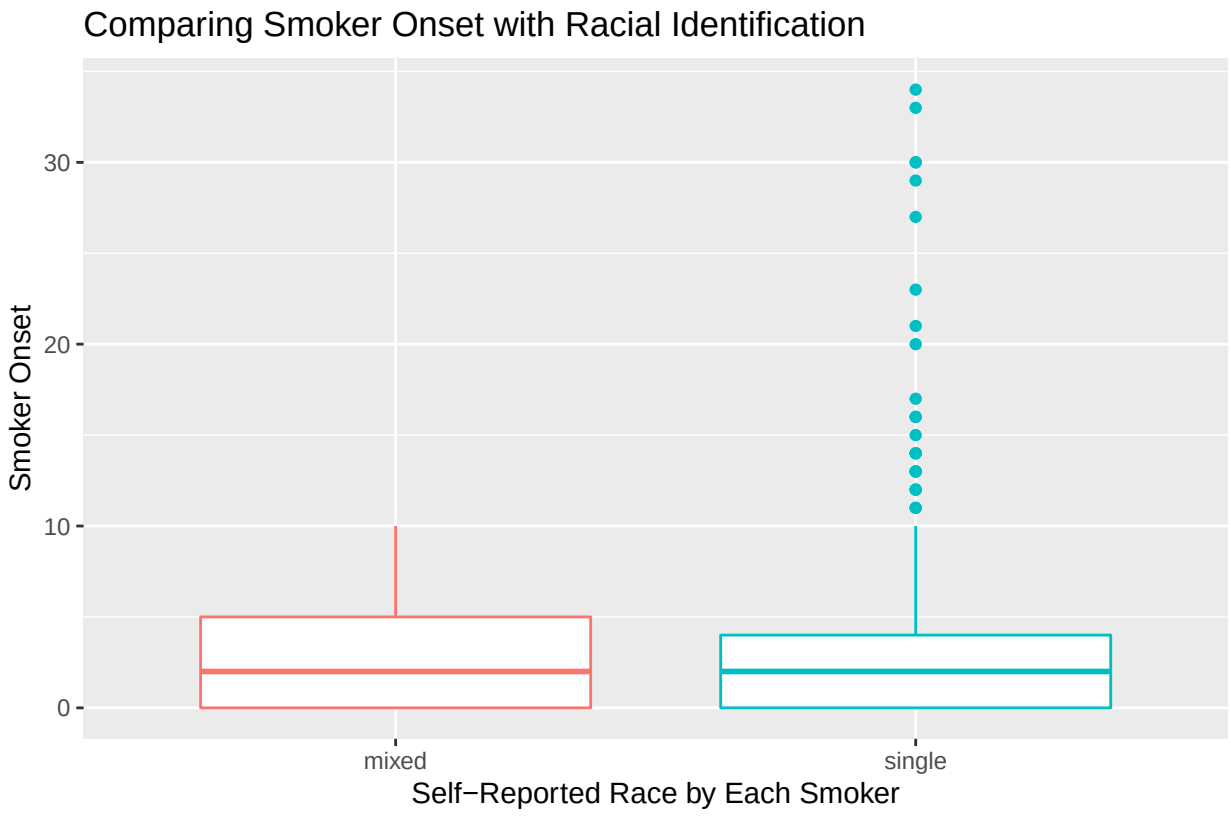
```
ggplot(data=df_rb, aes(x = counts, y=factor(smkbrand), fill=race_binary)) +
  geom_bar(aes(fill=race_binary), stat="identity", position=position_dodge2()) +
  labs(x="Counts",y="Cigarette Brands Purchased",
       title="Cigarette Brands Purchased vs. Racial Identification",
      caption="Data Source: CDPH 2011 California Smokers Cohort.")+
  scale_fill_manual(name="Self-Reported Race by Each Smoker",
                    values=c("cyan3","darkorange", "darkgreen"))+
  theme(legend.position="bottom")
```



This visualization 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

```
ggplot(data=df_rs, aes(x=race_binary,y=smoker_onset)) +
  geom_boxplot(aes(colour=race_binary)) +
  theme(legend.position="none") +
  labs(x="Self-Reported Race by Each Smoker",y="Smoker Onset",
       title="Comparing Smoker Onset with Racial Identification",
       caption= "Data Source: CDPH 2011 California Smokers Cohort.")
```



This box plot 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, thus further analysis into additional behavioral factors beyond racial identification is needed to help reduce certain disease outcomes (heart disease, asthma, mental illness, diabetes).

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 tobacco use, 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 tobacco use and dependency.

Methods:

The data source used was from the California Department of Public Health Smokers Cohort 2011. 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 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). We worked through various components and tidied the dataset accordingly.

Results:

The results of the Kable table 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 and those with mental illness had the lowest mean. The bar graph 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). The box plot 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, thus further analysis into additional behavioral factors beyond racial identification is needed to help reduce certain disease outcomes (heart disease, asthma, mental illness, diabetes).

Discussion:

Our findings suggest that Diabetes has the highest mean pack-years out of our four assessed health outcomes. 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 our bar graph visualization we determined that Marlboro is the most popular cigarette brand regardless of race. More information around how Marlboro advertises their products may help provide additional context on why this cigarette brand is so popular among smokers. Like the Marlboro cigarette brand popularity shared by all races studied in this dataset, smoker onset also shares certain similarities- the median smoker onset was two years among both single and mixed races. This shows us that certain behaviors are constant regardless of race and more insight into smoker behaviors would be beneficial in order to determine why this is occurring.