Milestone #4

Rachael Baartmans, Lara Petalio, Christine Truong

11-21-22

NOTE: We calculated pack-years and joined our two data sets of race_data_2 and smoker_data_3 together before creating these visualizations below. The pack-years calculation was created as the new variable `pack_years` in our joined data set, joined_smoking_df. If necessary, please see the code chunks labeled "r joining race_data_2 and smoker_data_3 together" and "r pack-years calculation" in Milestone_4.Rmd to view these processes not mentioned in this pdf.

Visualizations

Table: Average Number of Pack-years by Disease Outcome Among Smokers

```
#Table for average number of pack-years per disease for smokers who have a disease
t_avg_pack_years_disease <- joined_smoking_df %>%
  mutate(disease = case_when(asthma == "Yes" ~ "Asthma",
                             heartdis == "Yes" ~ "Heart Disease",
                             diabetes == "Yes" ~ "Diabetes",
                             othmenill == "Yes" ~ "Mental Illness")) %>%
  select(disease, pack_years) %>%
  filter(!is.na(pack_years), !is.na(disease)) %>%
  group_by(disease) %>%
  summarize(avg_pack_years = round(sum(pack_years)/n(), 0))


#Kable table for average number of pack-years per disease outcome for smokers who
#have a disease (produced below)
kable(t_avg_pack_years_disease,
booktabs=T,
col.names=c("Disease", "Average Number of Pack-years"),
align='lcccc',
caption= 'Average Number of Pack-years by Disease Outcome Among Smokers') %>%
kable_styling(full_width = T) %>%
kable_styling(latex_options = "hold_position") %>%
footnote(general =
           "Data Source: 2011 California Smokers Cohort, CA Dept. of Health")
```

Table 1: Average Number of Pack-years by Disease Outcome Among Smokers

Disease	Average Number of Pack-years
Asthma	25
Diabetes	25
Heart Disease	28
Mental Illness	17

Note:
Data Source: 2011 California Smokers Cohort, CA Dept. of Health

Interpretation of Average Number of Pack-years by Disease Outcome Among Smokers Table: This table demonstrates the average number of pack-years per disease type for smokers who reported having asthma, diabetes, heart disease, and/or mental illness in the 2011 California Smokers Cohort study.

Among smokers who have reported having asthma, heart disease, diabetes, and/or mental illness, those with heart disease have the highest number of average pack-years (28), while those with mental illness have the lowest number of average pack-years (17).

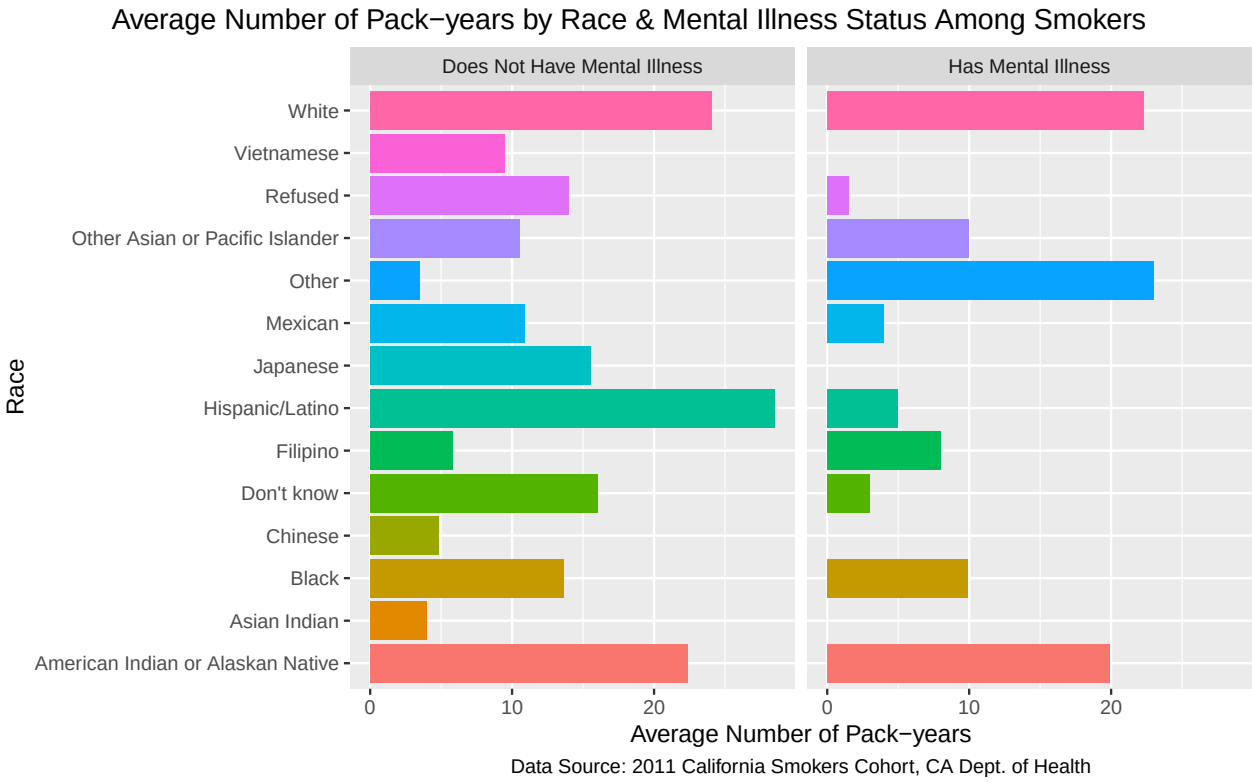
Bar Graph: Average Number of Pack-years by Race and Mental Illness Status Among Smokers

```
#We first created a subset of the data frame joined_smoking_df for our disease of
#interest, mental illness, called avg_pack_years_race_othmenill. This subset
#includes only the variable of `race` and average values of the variable
#`pack_years` pertaining to mental illness status. The purpose of creating this
#subset is to simplify the process of creating a graph in the next step by
#showing only the relevant information we need.

avg_pack_years_race_othmenill <- joined_smoking_df %>%
  filter(!is.na(pack_years)) %>%
  group_by(race, othmenill) %>%
  summarize(avg_pack_years = sum(pack_years)/n())


#We then created a bar graph representing avg_pack_years_race_othmenill,
#excluding NA values in the variable `othmenill` since we have determined that
#the NA values do not present valuable information for our analyses; the NA
#values had already been dropped for `avg_pack_years` in the process of creating
#the subset of avg_pack_years_race_othmenill in the previous step.

avg_pack_years_race_othmenill %>%
  drop_na(othmenill) %>%
  ggplot(aes(x = race, y = avg_pack_years)) +
  geom_bar(aes(fill = race), stat = "identity", position = "dodge") +
  coord_flip() +
  guides(fill = "none") +
  theme(plot.title.position = "plot",
        plot.title = element_text(hjust = 0.5)) +
  labs(x = "Race",
       y = "Average Number of Pack-years",
  title = "Average Number of Pack-years by Race & Mental Illness Status Among Smokers",
  caption = "Data Source: 2011 California Smokers Cohort, CA Dept. of Health") +
  facet_wrap(~ othmenill, labeller = labeller(othmenill =
                                        c("No" = "Does Not Have Mental Illness",
                                          "Yes" = "Has Mental Illness"))) +
  theme(plot.title = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5))
```



Interpretation of Average Pack-years by Race and Mental Illness Status Among Smokers Bar Graph:

This graph exhibits the number of average pack-years for each race category and by mental illness status of smokers in the 2011 California Smokers Cohort study.

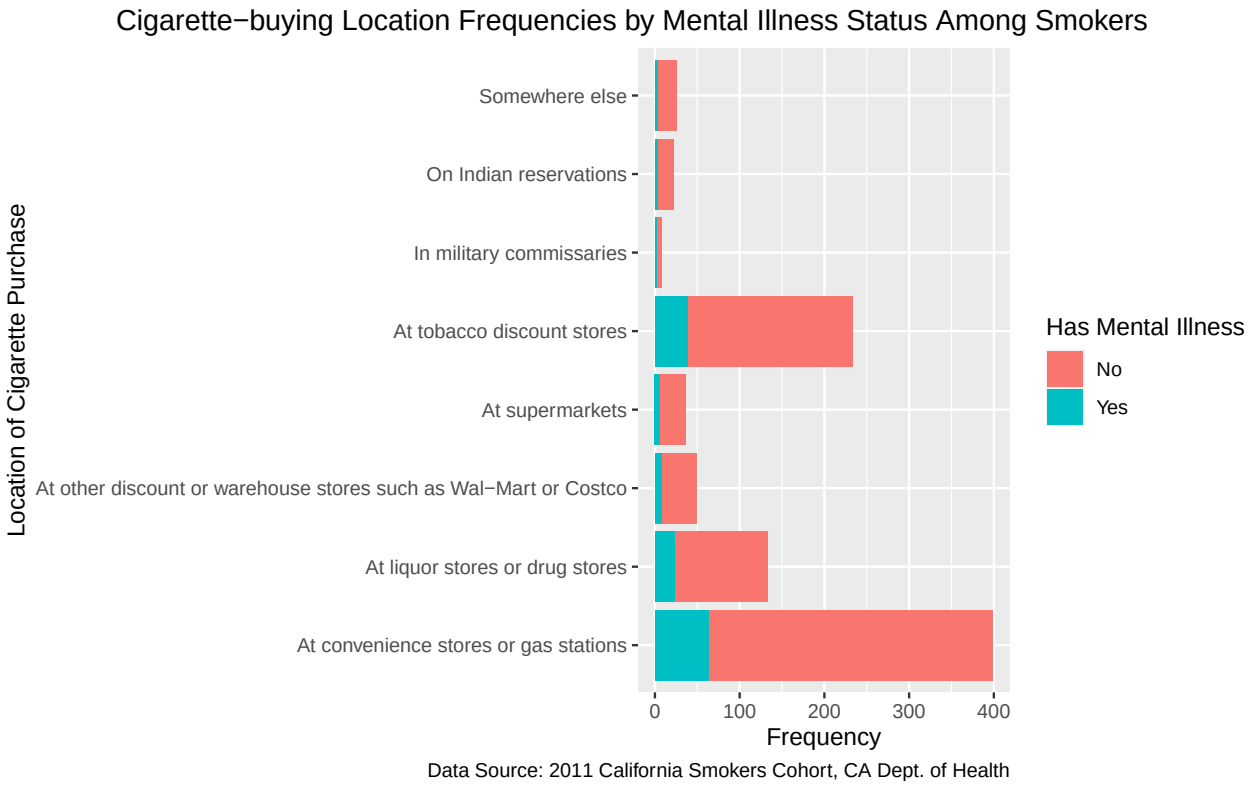
Among smokers who have reported having no mental illness, those who identified as "Hispanic/Latino" by race appear to have the greatest number of average pack-years, followed by "White and"American Indian or Alaskan Native", out of all race categories in the 2011 California Smokers Cohort.

Among smokers who have reported having mental illness, those who identified as "Other" by race appear to have the greatest number of average pack-years compared to other races in the 2011 California Smokers Cohort, with "White" and "American Indian or Alaskan Native" following closely behind.

Bar Graph: Cigarette-buying Location Frequencies by Mental Illness Status Among Smokers

```
#As similarly performed for the graph above, we also excluded NA values for the
#variables of `wherebuy` and `othmenill` prior to creating this graph because
#we did not believe NA values would be telling us any valuable information.
#We chose to create a stacked bar graph instead of a dodged bar graph in order
#to facilitate total frequency comparisons between different cigarette purchase
#locations regardless of mental illness status.

joined_smoking_df %>%
  filter(!is.na(wherebuy), !is.na(othmenill)) %>%
  ggplot(aes(x = wherebuy)) +
  geom_bar(aes(fill = othmenill), position = "stack") +
  coord_flip() +
  theme(plot.title.position = "plot",
        plot.title = element_text(hjust = 0.5)) +
  scale_fill_discrete(name = "Has Mental Illness") +
  labs(x = "Location of Cigarette Purchase",
       y = "Frequency",
  title = "Cigarette-buying Location Frequencies by Mental Illness Status Among Smokers",
  caption = "Data Source: 2011 California Smokers Cohort, CA Dept. of Health")
```



Interpretation of Cigarette-buying Location Frequencies by Mental Illness Status Among Smokers Bar Graph:

This bar graph explores the relationship between frequencies per cigarette purchase location and mental illness status among smokers in the 2011 California Smokers Cohort study.

Mental illness was not reported by the majority of the smokers for each cigarette purchase location. However, mental illness was reported in the greatest number by those who purchased cigarettes at convenience stores or gas stations, followed by those who purchased cigarettes at tobacco discount stores; these are the two locations that also have the highest frequencies among smokers for making cigarette purchases at.