Milestone #4

Rachael Baartmans, Lara Petalio, Christine Truong

11-14-22

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
##
## Attaching package: 'kableExtra'
```

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

```
## Rows: 1000 Columns: 24
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (24): ID, NERVOUS, WORRYING, PROBINTR, PROBDOWN, ASTHMA, HEARTDIS, DIABE...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 1000 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): smokstat, HOWMANY, SMOK6UNI, BUYCALIF, WHEREBUY
## dbl (2): psraid, SMOK6NUM
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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 into the new variable pack_years in our joined data set, joined_smoking_df. If necessary, please see the code chunks labeled "r pack-years calculation" and "r joining race_data_2 and smoker_data_3 together" in Milestone_4.Rmd to view these processes.

Visualizations

Table: Average Pack-years by Disease Status

```
#Table for avg 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 avg pack-years per disease for smokers who have a disease
#(produced below)
kable(t_avg_pack_years_disease,
booktabs=T,
col.names=c("Disease", "Average Pack-years"),
align='lcccc',
caption= 'Average Pack-years for Smokers Who Have a Disease') %>%
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 Pack-years for Smokers Who Have a Disease

Disease	Average 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 Pack-years by Disease Status Table:
This table demonstrates the average number of pack-years (i.e., average number of cigarette packs smoked per year) 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 Chart: Average Pack-years by Race and Mental Illness

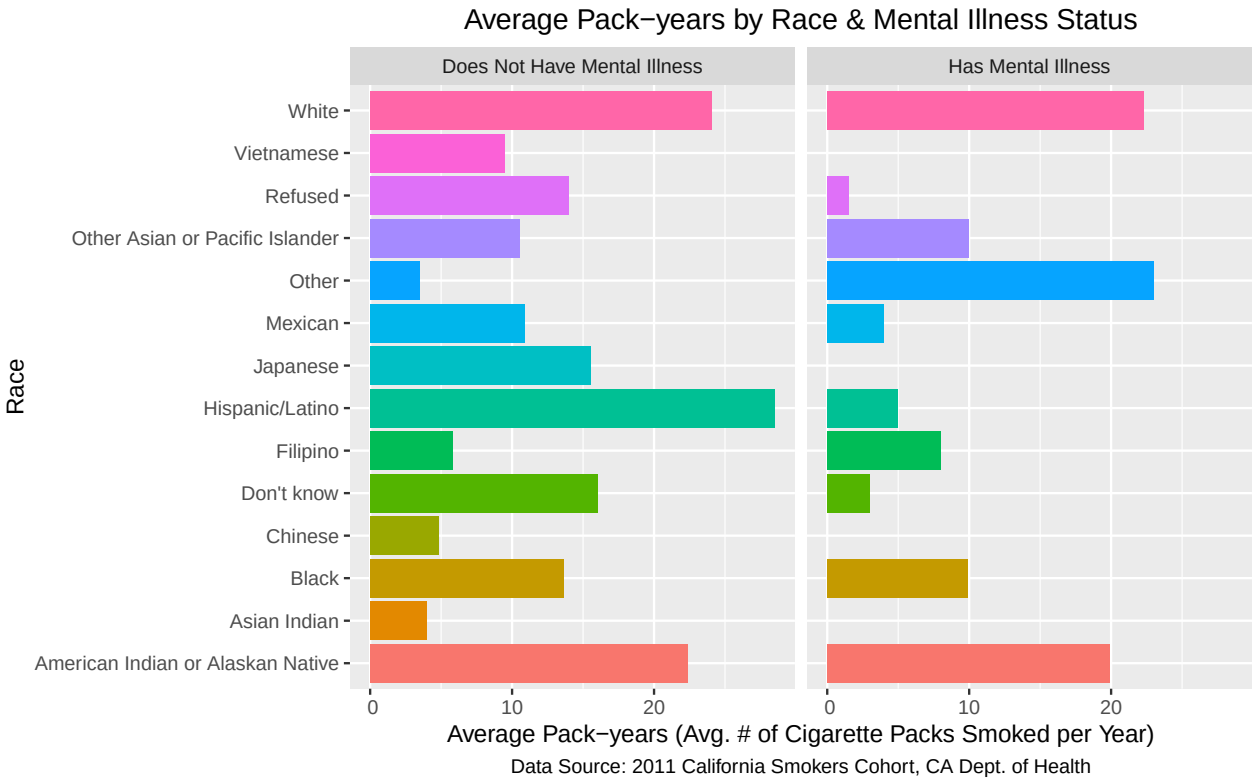
```
#We first created a subset of the data frame joined_smoking_df for our disease of
#interest, mental illness, called avg_pack_years_race_othmenill. This subset
#includes only the variable of `race` and average values of the variable
#`pack_years` pertaining to mental illness status. The purpose of 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())
```

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

```
#We then created a bar graph representing avg_pack_years_race_othmenill excluding
#NA values in the variables `othmenill` and `avg_pack_years` since we have
#determined that the NA values do not present valuable information for our
#analyses.

avg_pack_years_race_othmenill %>%
  drop_na(c(othmenill, avg_pack_years)) %>%
  ggplot(aes(x = race, y = avg_pack_years)) +
  geom_bar(aes(fill=race), stat="identity", position = "dodge") +
  coord_flip() +
  guides(fill = "none") +
  labs(x = "Race",
       y = "Average Pack-years (Avg. # of Cigarette Packs Smoked per Year)",
  title = "Average Pack-years by Race & Mental Illness Status",
  caption = "Data Source: 2011 California Smokers Cohort, CA Dept. of Health") +
  scale_y_continuous(labels = function(x) format(x,big.mark=",",
                                                 scientific=FALSE)) +
  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 Disease Bar Graph:

This graph exhibits the number of average pack-years (i.e., average number of cigarette packs smoked per year) according to each race category and mental illness status of smokers in the 2011 California Smokers Cohort study.

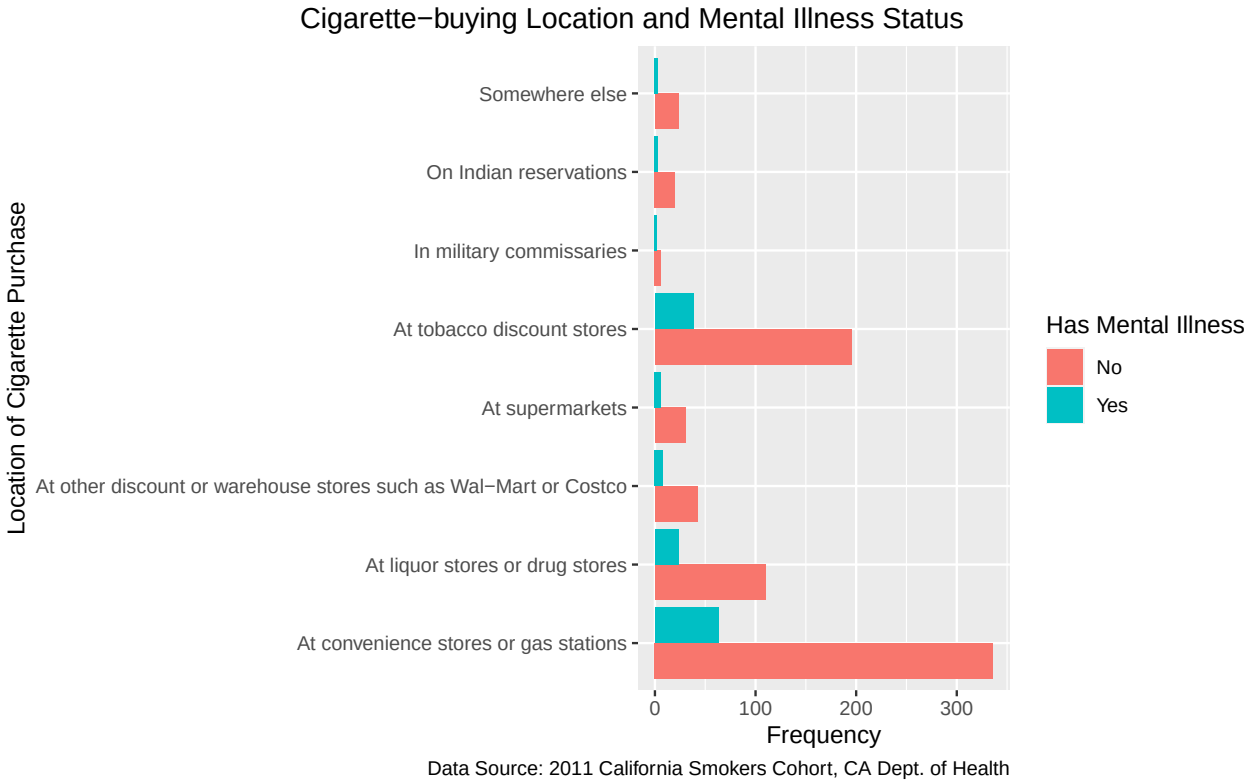
Among smokers who have reported having no mental illness, those who identified as "Hispanic/Latino" by race appear to have the greatest number of average pack-years, followed closely by "White" and "American Indian or Alaskan Native", compared to other races in the 2011 California Smokers Cohort.

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, followed closely by "White" and "American Indian or Alaskan Native", compared to other races in the 2011 California Smokers Cohort.

Bar Graph: Cigarette Purchase Location and Mental Illness

```
#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 stack bar graph in order to more easily compare the
#mental illness statuses reported by smokers for each cigarette
#purchase location, as well as showcase the total number of times (frequency)
#that cigarettes were purchased at each location by smokers.

joined_smoking_df %>%
  filter(!is.na(wherebuy), !is.na(othmenill)) %>%
  ggplot(aes(x = wherebuy)) +
  geom_bar(aes(fill=othmenill), position = "dodge") +
  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 and Mental Illness Status",
  caption = "Data Source: 2011 California Smokers Cohort, CA Dept. of Health")
```



Interpretation for Mental Illness Status by Cigarette Purchase Location Bar Graph:

This bar graph explores the relationship between 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 number of cigarette purchases made among smokers.