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.
```

Calculating Pack-years

```
#We calculated pack-years, which is given by the formula of
#pack-years = # of packs of cigarettes smoked per day * years a person has smoked.
#This calculation led to the creation of a new variable called `pack_years`.
#`pack_years` was created conditionally based on the three different time units
#as determined by the existing variable `smok6uni`, which are: "Days", "Months",
#and "Years". To change the unit of "Days" to years, we set up a conditional
#statement in the code to divide `smok6num` by 365 before multiplying the result
#by `packs_per_day` to get pack-years. Similarly, to change the unit of "Months"
#to years, we set up a conditional statement in the code to divide `smok6num`
#by 12 before multiplying the result by `packs_per_day` to get pack-years. For
#`smok6uni` observations that have the value "Years", we just multiplied
#`smok6num` by `packs_per_day` to get pack-years directly. We assigned
#this overall change in the data frame smoker_data_2 to a new data frame
#called smoker_data_3, which includes the new variable `pack_years`.

smoker_data_3 <- smoker_data_2 %>%
  mutate(pack_years =
           case_when(smok6uni == "Days" ~ packs_per_day*(smok6num/365),
                     smok6uni == "Months" ~ packs_per_day*(smok6num/12),
                     smok6uni == "Years" ~ packs_per_day*(smok6num)))

#We then rounded `pack_years` to the nearest whole number for all observations.

smoker_data_3$pack_years <- round(smoker_data_3$pack_years, 0)
```

Joining Cleaned Data Sets Together

```
#In order to join our two cleaned data sets together, we first had to remove the
#strings of 'DIS' and 'STAT' from the `id` column of race_data_2 by using gsub().
#We overwrote these changes in the race_data_2 data frame and viewed these new
#changes to make sure the `id` variable only contains numbers and no
#characters.

race_data_2$id <- gsub('[DISSTAT]', '', race_data_2$id)

#Next, looking at the smoker_data_3 data frame, we see that the `psraid`
#variable contains each study participant's unique ID number, but the variable
#is a numeric data type. On the other hand, `id` from the race_data_2
#data frame is a character data type. We needed to convert `psraid` then from
#character to numeric data type because 1) `psraid` is an identifier rather than
#a numeric value to mathematically manipulate even if it does contain numbers
#and 2) in order to perform a join, the two variables must be the same data
#type.

smoker_data_3$psraid <- as.character(smoker_data_3$psraid)

#Afterward, we performed an inner join between race_data_2 and
#smoker_data_3 by each study participant's unique ID number, which is
#represented by `id` in race_data_2 and `psraid` in smoker_data_3. We
#chose to do an inner join because we wanted to select participants that
#exist in each of our two data sets for our final data frame.
#We assigned this join to a new data frame called joined_smoking_df.

joined_smoking_df <- inner_join(x = race_data_2, y = smoker_data_3,
                                by=c("id" = "psraid"))
```

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

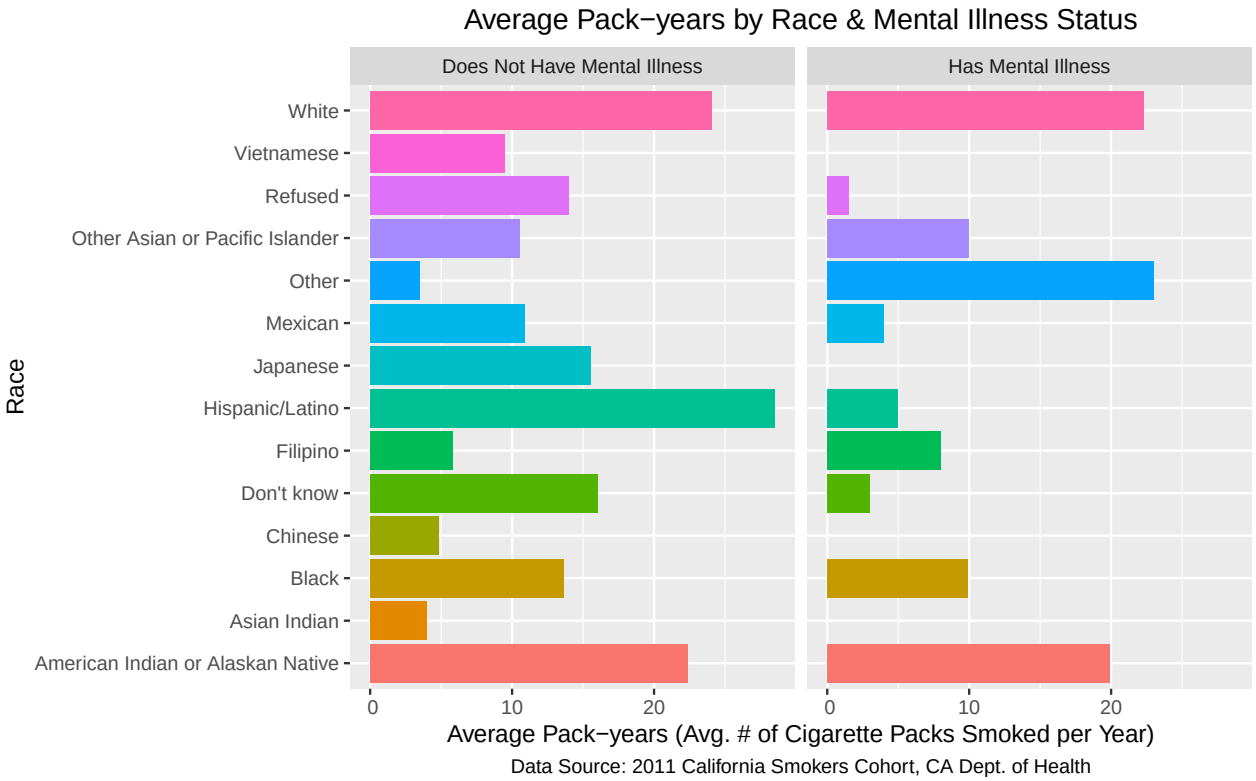
```
#We first created a subset of the data frame joined_smoking_df for our disease of
#interest, mental illness, called avg_pack_years_race_othmenill. This subset
#includes only the variable of `race` and average values of the variable
#`pack_years` pertaining to mental illness status. The purpose of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

FOR LARA:

one additional table or plot With ggplot: dodged bar chart with (x = cigarette purchase location, aes(fill = othmenill) to view cigarette purchase location and mental illness status