Milestone #4 - complete by FRIDAY NOV 18

Milestone #1 1. Project option selection: We pick the California Smokers' Cohort data set for this project.

2. Create git repository / Share a link to your group's git repository: https://github.com/ ambernmorris/team_project
3. What is your team's preferred communication method - email, text, bcourse messaging? Text, email, Zoom calls
4. When will your team be holding meetings? How frequently will you meet? Are there times or days that work well for everyone? Once a week to every other week; we are already in contact frequently because we have other classes together and have worked together before. Our schedules vary, but typically Wednesday mornings work best for us to meet.
5. Discuss future non-academic commitments that might affect members' availability: Work, life events, partners, vacations. We communicate clearly when we have conflicts. Discuss meeting tempo: "checking in" at the beginning of meetings versus "just sticking to business". We like to check-in and then we usually get to work after chatting some.
6. How is your team going to keep track of progress? Who will be taking minutes, creating agendas, and contacting the course facilitators with questions? We both come to our meetings prepared with questions and a general idea of what we will be working on together. Amber will be the one usually to contact course instructors.
7. We encourage you to discuss potential dates and times to meet with a course facilitator during Weeks 3-5. Decide on the best date and time for all. We will meet with Lauren on Mondays via private office hour appointment if needed.
8. Determine a point person to submit each assignment for the team. Ileah
9. If a conflict arises, plan to solve the issue as soon as possible. This is best done using synchronous (Zoom, Google Hangouts) communication rather than asynchronous (email). If the group is unable to resolve the conflict, seek advice from the instructional team. Yes, will do.
10. Team's preferred communication method: Text, email, zoom
11. Team's preferred meeting times and frequency: Once a week to every other week; we take other classes together, so we are in contact often and are always willing to meet whenever.
12. Team's preferred method for tracking progress: We plan to have each milestone finished a week ahead of the due date, so we can make changes if we need to ahead of time.
13. Point person for contacting course facilitators with questions: Amber

Milestone 2

What is the data source? (1-2 sentences on where the data is coming from, dates included, etc.) The data source is from the California Smokers Cohort (CSC) 2011 through UC San Diego. These data are from a survey designed to investigate factors associated with tobacco quitting behaviors, sponsored/funded by CDPH.

How does the dataset relate to the group problem statement and question? Through these data, CDPH will better understand tobacco use and behaviors among smokers in California and be able to design an implementation strategies in high-risk communities to increase quitting behaviors. As we go through the milestones, the elements we use may change, but we foresee combining each data set and exploring how race, income, and cigarette brand affect heart disease outcomes. Specifically, creating a visualization to explore how these variables interact with each other and making recommendations to CDPH based on that.

2. Import statement

```
#adding libraries

library(tidyverse)
```

```
## 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()
```

```
library(readr)
library(readxl)
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(stringr)

#creating a file path for both data sets
file_path_smoker <- "https://raw.githubusercontent.com/PHW290/phw251_projectdata/main/ca_csc_smoker_data
file_path_race <- "https://raw.githubusercontent.com/PHW290/phw251_projectdata/main/ca_csc_outcome_race_

#importing full file of both data sets
smoker_data <- read_csv(file_path_smoker)
```

```
## Warning in readRDS(responseFile): invalid or incomplete compressed data

## Rows: 1000 Columns: 156

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (152): RIGHTSEX, smokstat, ACIG100, DOSMOKE, HOWMANY, SMOK6NUM, SMOK6UNI...
## dbl   (3): psraid, nosmknum1, quitoffn
## lgl   (1): QUITINTNFORM
##
## 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.
```

```
race_data <- read_csv(file_path_race)
```

```
## Rows: 1000 Columns: 89
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (81): ID, INCARS, BANAGREE, CASINSMK, CASMOKES, HHSMOKNU, ACQSMOKE, LIVE...
## dbl  (6): ACTIVHRS, ACTIVMIN, HTINFEET, HTINCHES, WGTINLBS, AGEUS
## lgl  (2): HTCENTIM, WGTINKILOS
##
## 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.
```

Utilize function arguments to control relevant components (i.e. change column types, column names, missing values, etc.)

```
#first data set work below
#change columns to all lowercase
clean_smoker_data <- clean_names(smoker_data)

#change column names (the elements we want to use only)
new_smoker_data <- rename(clean_smoker_data, sex = rightsex,
                          smoking_status = smokstat, cig_brand = smkbrand)

#subset the columns we want (elements we want to use only)
updated_smoker_data <- select(new_smoker_data, c("sex", "smoking_status",
                        "cig_brand"))

#change the DO NOT READ values to NA
updated_smoker_data[updated_smoker_data == "(DO NOT READ) Refused"] <- NA
updated_smoker_data[updated_smoker_data == "(DO NOT READ) Don't know"] <- NA

#change columns to all lowercase
clean_race_data <- clean_names(race_data)

#change column names (the elements we want to use only)
new_race_data <- rename(clean_race_data, heart_disease=heartdis,
white=race01, black=race02, japanese=race03, chinese=race04,
filipino=race05, korean=race06, asian_or_pacific_islander=race07,
mexican=race09,
hispanic_latino=race10, other=race11, vietnamese=race12,
asian_indian=race13, refused=race14, dont_know=race15,
americanindian_or_alaskannative=race08)

#subset the columns we want (elements we want to use only)
updated_race_data <- select(new_race_data, c("heart_disease", "income", "white",
"black", "japanese", "chinese", "filipino", "korean",
"asian_or_pacific_islander", "americanindian_or_alaskannative", "mexican",
"hispanic_latino", "other", "vietnamese", "asian_indian", "refused",
  "dont_know"))

#change the DO NOT READ values to NA
updated_race_data[updated_race_data == "(DO NOT READ) Refused"] <- NA
updated_race_data[updated_race_data == "(DO NOT READ) Don't know"] <- NA
```

3. Identify 5+ data elements required for your specified scenario. If <5 elements are required to complete the analysis, please choose additional variables of interest in the data set to explore in this milestone.

This milestone is focused on: rightsex (sex), smokstat (smoking status), smkbrand (brand of cigarettes), race (race), income (income), and heartdis (heart disease status).

Utilize functions or resources in RStudio to determine the types of each data element (i.e. character, numeric, factor)

```
#smoker data set
str(updated_smoker_data)
```

```
## tibble [1,000 x 3] (S3: tbl_df/tbl/data.frame)
##  $ sex           : chr [1:1000] "Female" "Female" "Male" "Female" ...
##  $ smoking_status: chr [1:1000] "Current daily smoker" "Current daily smoker" "Current nondaily smoke
##  $ cig_brand     : chr [1:1000] "Virginia Slims" "Marlboro" "Camel" "Marlboro" ...
```

```
#race data set
str(updated_race_data)
```

```
## tibble [1,000 x 17] (S3: tbl_df/tbl/data.frame)
##  $ heart_disease                  : chr [1:1000] "Yes" "No" "No" "No" ...
##  $ income                         : chr [1:1000] "$30,001 to $50,000" "$20,000 or less" "$30,001 to $
##  $ white                          : chr [1:1000] "Yes" "Yes" "Yes" "Yes" ...
##  $ black                          : chr [1:1000] NA NA NA NA ...
##  $ japanese                       : chr [1:1000] NA NA NA NA ...
##  $ chinese                        : chr [1:1000] NA NA NA NA ...
##  $ filipino                       : chr [1:1000] NA NA NA NA ...
##  $ korean                         : chr [1:1000] NA NA NA NA ...
##  $ asian_or_pacific_islander      : chr [1:1000] NA NA NA NA ...
##  $ americanindian_or_alaskannative: chr [1:1000] NA NA NA NA ...
##  $ mexican                        : chr [1:1000] NA NA NA NA ...
##  $ hispanic_latino                : chr [1:1000] NA NA NA NA ...
##  $ other                          : chr [1:1000] NA NA NA NA ...
##  $ vietnamese                     : chr [1:1000] NA NA NA NA ...
##  $ asian_indian                   : chr [1:1000] NA NA NA NA ...
##  $ refused                        : chr [1:1000] NA NA NA NA ...
##  $ dont_know                      : chr [1:1000] NA NA NA NA ...
```

Identify the desired type/format for each variable—will you need to convert any columns to numeric or another type?

Both data sets were assessed and it was determined there were not any columns that needed to be converted to another data type at this time.

4. Provide a basic description of the 5+ data elements

```
#first data set
summary(updated_smoker_data)
```

```
##      sex            smoking_status      cig_brand
##  Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

```
#second data set
summary(updated_race_data)
```

```
##  heart_disease         income             white              black
##  Length:1000        Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##    japanese           chinese            filipino           korean
##  Length:1000        Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##  asian_or_pacific_islander americanindian_or_alaskannative   mexican
##  Length:1000               Length:1000                     Length:1000
##  Class :character          Class :character                Class :character
##  Mode  :character          Mode  :character                Mode  :character
##  hispanic_latino       other            vietnamese        asian_indian
##  Length:1000        Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##    refused            dont_know
##  Length:1000        Length:1000
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

Milestone #3

```
raceclean <- clean_names(race_data)
smokerclean <- clean_names(smoker_data)
```

1. Subset row or columns as needed in race data

```
#subset columns from race dataset
race <- raceclean %>% select(id, race01, race02, race03, race04, race05,
race06, race07, race08, race09, race10, race11, race12, race13, race14, race15,
income,vereduc, wgtinlbs, htinfeet, goodhlth, harmhlth, smokalone, acqsmoke,
act10min,drinkfiv, heartdis, diabetes)
```

```
#rename columns
racedata <- race %>% rename(heart_disease=heartdis,
white=race01, black=race02, japanese=race03, chinese=race04,
filipino=race05, korean=race06, asian_or_pacific_islander=race07,
mexican=race09,
hispanic_latino=race10, other=race11, vietnamese=race12,
asian_indian=race13, refused=race14, dont_know=race15,
americanindian_or_alaskannative=race08, school_level = vereduc,
smoking_harms_health = harmhlth, how_many_people_smoke = acqsmoke,
physically_active_for_10min = act10min,
days_had_4ormore_drinks_inrow = drinkfiv)
```

```
#recode race columns into one column
racedata1 <- racedata %>% mutate(race = case_when(white == "Yes" ~ "white",
black == "Yes" ~ "black", japanese == "Yes" ~ "japanese",
chinese == "Yes" ~ "chinese", filipino == "Yes" ~ "filipino",
korean == "Yes" ~ "korean",
asian_or_pacific_islander == "Yes" ~ "asian_or_pacific_islander",
mexican == "Yes" ~ "mexican", hispanic_latino == "Yes" ~ "hispanic_latino",
other == "Yes" ~ "other", vietnamese == "Yes" ~ "vietnamese",
asian_indian == "Yes" ~ "asian_indian", refused == "Yes" ~ "refused",
dont_know == "Yes" ~ "dont know",
americanindian_or_alaskannative == "Yes" ~ "americanindian_or_alaskannative"))
```

```
#subset race data that is ready to combine with smoker data
recodedrace_data <- racedata1 %>% select(id, race, income, school_level, wgtinlbs,
htinfeet, goodhlth, smoking_harms_health, smokalone, how_many_people_smoke,
physically_active_for_10min,
days_had_4ormore_drinks_inrow, heart_disease,
diabetes)
```

2. Subset row or columns as needed in smoker data

```
#subset columns from smoker data
smokerdata <- smokerclean %>% select(psraid, rightsex, smokstat, howmany,
smok6num, smok6uni, smkbrand, smk1age, smkage)
```

```
#rename columns in smoker data
datasmoker <- smokerdata %>% rename(sex = rightsex, smoking_status = smokstat,
cig_brand = smkbrand, cigs_per_day = howmany,
how_long_smoking_daily = smok6num, unit_of_time_smoking_daily = smok6uni,
age_when_first_smoked = smk1age, age_when_daily_smoking_began = smkage,
id = psraid)
```

3. Combine both data sets

```
#make id variables the same type in both datasets to combine datasets by id
newracedata <- cbind(recodedrace_data, i_d = datasmoker$id)
newracedata1 <- newracedata %>% select(i_d, race, income, school_level, wgtinlbs,
htinfeet, goodhlth, smoking_harms_health, smokalone, how_many_people_smoke,
physically_active_for_10min,
days_had_4ormore_drinks_inrow, heart_disease,
diabetes) %>% rename(id = i_d)
```

```
#checking column types
typeof(datasmoker$id)
```

```
## [1] "double"
```

```
typeof(newracedata1$id)
```

```
## [1] "double"
```

```
#combine datasets by ID
combineddata <- inner_join(newracedata1, datasmoker, by = c("id"))
```

```
#change DO NOT READ values to NA data values because .......
combineddata[combineddata == "(DO NOT READ) Refused"] <- NA
combineddata[combineddata == "(DO NOT READ) Don't know"] <- NA
```

4. Create and clean two new variables needed to analysis

```
#variable 1, bmi
new_variables_data_1 <- combineddata %>%
  mutate(bmi = (wgtinlbs * 0.45359237) / (htinfeet * 0.304)^2)

#first need to make packs per day variable to create packs per year variable
#packs per day is cigs_per_day was divided by 20 because there are 20 cigarettes in 1 pack
#variable 2
new_variables_data_2 <- new_variables_data_1 %>%
  mutate(cigs_per_day = as.numeric(cigs_per_day),
         packs_per_day = (cigs_per_day / 20))
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```r
#variable 3, pack years calculation
final_variables_data <- new_variables_data_2 %>%
  mutate(how_long_smoking_daily = as.numeric(how_long_smoking_daily),
         pack_years = (packs_per_day) * (how_long_smoking_daily))
```

5. Data dictionary for final dataset

minimum of 4 variables

variable 1 Variable name: cig_brand Data type: Character Data description: This variable contains the name of the cigarette brand the participant reported smoking.

Variable 2 Variable name: age_when_daily_smoking_began Data type: Character Data description: This variable contains the age of the participant when they began to smoke cigarettes on a regular basis.

variable 3 Variable name: age_when_first_smoked Data type: Character Data description: This variable contains the age when the participant first smoked their first whole cigarette.

variable 4 Variable name: bmi Data type: Double Data description: This variable contains the participant's BMI, calculated as kg/m^2.

6. Table with descriptive statistics

```r
library(kableExtra)
```

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

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

```r
df_ofdescriptive_stats <- data.frame("Variable" =
c("Pack Years", "Cigs Per Day", "BMI", "Packs Per Day"), "Mean" =
c(21.68, 13.89, 33.38, 0.69), "Standard Deviation" =
c(18.19, 9.30, 10.67, 0.47), "Range" = c(119.9, 59, 161.29, 2.95))
kable(df_ofdescriptive_stats, booktabs = T, longtable=T, col.names =
        c("Variable", "Mean", "SD", "Range"), align= 'lcccc',
caption= "Descriptive Statistics of Four Variables from 2011 California
Smokers Cohort")
```

Table 1: Descriptive Statistics of Four Variables from 2011 California Smokers Cohort

Variable	Mean	SD	Range
Pack Years	21.68	18.19	119.90
Cigs Per Day	13.89	9.30	59.00
BMI	33.38	10.67	161.29
Packs Per Day	0.69	0.47	2.95

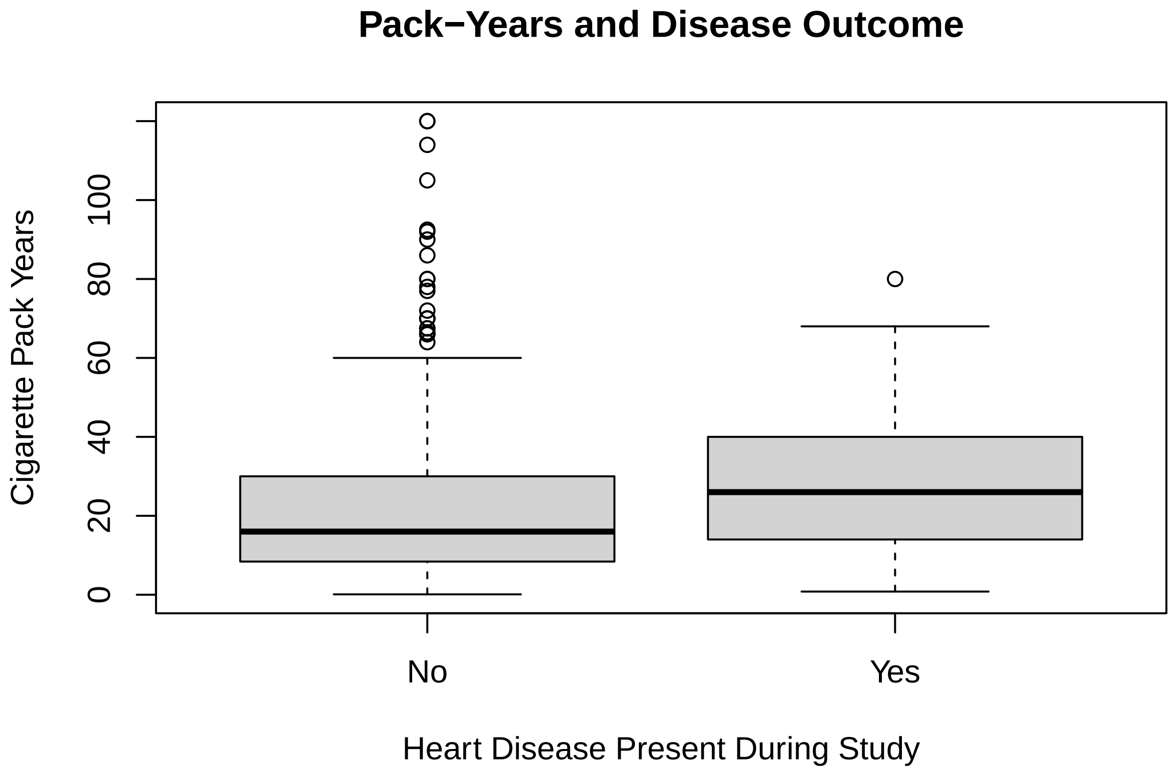
Milestone 4

For each visual, include: Code used to generate visual Legend (if necessary) 1-2 sentence interpretation

One print-quality plot or chart that shows the average number of pack-years and the disease outcomes Using variables: pack_years, heart_disease, diabetes

```
data_visual_one <- final_variables_data %>%
  select(c("pack_years", "heart_disease", "diabetes"))

#Box plot of pack years by heart disease
visual_one <- boxplot(pack_years~heart_disease, data = data_visual_one,
main="Pack-Years and Disease Outcome",
xlab= "Heart Disease Present During Study", ylab="Cigarette Pack Years")
```



```
print(visual_one)
```

```
## $stats
##      [,1] [,2]
## [1,]  0.1  0.8
## [2,]  8.4 14.0
## [3,] 16.0 26.0
## [4,] 30.0 40.0
## [5,] 60.0 68.0
##
## $n
## [1] 752  70
##
## $conf
##          [,1]  [,2]
## [1,] 14.75548 21.09
## [2,] 17.24452 30.91
##
## $out
##  [1] 114.0  66.0 120.0  70.0  90.0  77.0  86.0  92.0 120.0  80.0  67.5  67.5
## [13]  92.5  64.0  72.0  78.0  80.0  70.0  66.0  70.0  66.5 105.0  80.0
##
## $group
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2
##
## $names
## [1] "No"  "Yes"
```

Based on this visual, a participant in this study with heart disease had a wider range of cigarette pack years and higher median number of pack years when compared to another participant in the same study without heart disease.

One print-quality plot or chart that compares race and income on disease status Using variables: race, income, heart_disease, diabetes