This is a team assignment; each team should complete and turn in a PDF created from an Rmd via Github. Please include code and output for the following components

Description of dataset 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 eachother and making recommendations to CDPH based on that.

Import statement NOTE: Please use datasets available in the PHW251 Project Data github repo (Links to an external site.) (this is important to make sure everyone is using the same datasets) Use appropriate import function and package based on the type of file

```
#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)
```

```r
#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 to see what it looks like
smoker_data <- read_csv(file_path_smoker)
```

```
## 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.
```

```r
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.)

```r
#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[29,3] <- "NA"
updated_smoker_data[65,3] <- "NA"
updated_smoker_data[116,3] <- "NA"
updated_smoker_data[232,3] <- "NA"
updated_smoker_data[246,3] <- "NA"
updated_smoker_data[247,3] <- "NA"
updated_smoker_data[272,3] <- "NA"
```

```
updated_smoker_data[278,3] <- "NA"
updated_smoker_data[281,3] <- "NA"
updated_smoker_data[339,3] <- "NA"
updated_smoker_data[344,3] <- "NA"
updated_smoker_data[386,3] <- "NA"
updated_smoker_data[534,3] <- "NA"
updated_smoker_data[556,3] <- "NA"
updated_smoker_data[643,3] <- "NA"
updated_smoker_data[681,3] <- "NA"
updated_smoker_data[855,3] <- "NA"
updated_smoker_data[878,3] <- "NA"
updated_smoker_data[949,3] <- "NA"

head(updated_smoker_data)
```

```
## # A tibble: 6 x 3
##   sex    smoking_status          cig_brand
##   <chr>  <chr>                   <chr>
## 1 Female Current daily smoker    Virginia Slims
## 2 Female Current daily smoker    Marlboro
## 3 Male   Current nondaily smoker Camel
## 4 Female Current daily smoker    Marlboro
## 5 Male   Current daily smoker    Camel
## 6 Female Current daily smoker    Pall Mall
```

```
#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,   americanindian_or_alaskannative=race
mexican=race09,
hispanic_latino=race10, other=race11, vietnamese=race12,
asian_indian=race13, refused=race14, dont_know=race15)

#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

head(updated_race_data)
```

```
## # A tibble: 6 x 17
##   heart_disease income              white black japanese chinese filipino korean
##   <chr>         <chr>               <chr> <chr> <chr>    <chr>   <chr>    <chr>
```

```
## 1 Yes           $30,001 to $50,000  Yes   <NA>  <NA>     <NA>    <NA>     <NA>
## 2 No            $20,000 or less     Yes   <NA>  <NA>     <NA>    <NA>     <NA>
## 3 No            $30,001 to $50,000  Yes   <NA>  <NA>     <NA>    <NA>     <NA>
## 4 No            $20,001 to $30,000  Yes   <NA>  <NA>     <NA>    <NA>     <NA>
## 5 No            $100,001 to $150,0~ Yes   <NA>  <NA>     <NA>    <NA>     <NA>
## 6 Yes           $20,001 to $30,000  Yes   <NA>  <NA>     <NA>    <NA>     <NA>
## # ... with 9 more variables: asian_or_pacific_islander <chr>,
## #   americanindian_or_alaskannative <chr>, mexican <chr>,
## #   hispanic_latino <chr>, other <chr>, vietnamese <chr>, asian_indian <chr>,
## #   refused <chr>, dont_know <chr>
```

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
typeof(updated_smoker_data$sex)
```

```
## [1] "character"
```

```
typeof(updated_smoker_data$smoking_status)
```

```
## [1] "character"
```

```
typeof(updated_smoker_data$cig_brand)
```

```
## [1] "character"
```

```
#race data set
typeof(updated_race_data$income)
```

```
## [1] "character"
```

```
typeof(updated_race_data$heart_disease)
```

```
## [1] "character"
```

```
typeof(updated_race_data$white)
```

```
## [1] "character"
```

```
typeof(updated_race_data$black)
```

```
## [1] "character"
```

```
typeof(updated_race_data$japanese)
```

```
## [1] "character"
```

```
typeof(updated_race_data$chinese)
```

```
## [1] "character"
```

```
typeof(updated_race_data$filipino)
```

```
## [1] "character"
```

```
typeof(updated_race_data$korean)
```

```
## [1] "character"
```

```
typeof(updated_race_data$asian_or_pacific_islander)
```

```
## [1] "character"
```

```
typeof(updated_race_data$american_indian_or_alaskan_native)
```

```
## Warning: Unknown or uninitialised column: ‘american_indian_or_alaskan_native‘.
```

```
## [1] "NULL"
```

```
typeof(updated_race_data$mexican)
```

```
## [1] "character"
```

```
typeof(updated_race_data$hispanic_latino)
```

```
## [1] "character"
```

```
typeof(updated_race_data$other)
```

```
## [1] "character"
```

```
typeof(updated_race_data$vietnamese)
```

```
## [1] "character"
```

```
typeof(updated_race_data$asian_indian)
```

```
## [1] "character"
```

```
typeof(updated_race_data$refused)
```

```
## [1] "character"
```

```
typeof(updated_race_data$dont_know)
```

```
## [1] "character"
```

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 determined there were not any columns that needed to be converted to another data type at this time.

Provide a basic description of the 5+ data elements Numeric: mean, median, range Character: unique values/categories Or any other descriptives that will be useful to the analysis

```
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
```

```
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 ...
```