Capstone Project Exploratoru Analysis

Steve Uzupis

2024-07-07

*Packages Load*

# load packages  
  
library(readr) # load csv files  
library(readxl) # load excel files  
library(dplyr) # data manipulation

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(lubridate) # date & time manipulation

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(ggplot2) # data visualization  
library(tidyr) # collection of statistical packages, packages loaded individually  
library(corrplot) # to visualize correlations

## Warning: package 'corrplot' was built under R version 4.3.3

## corrplot 0.92 loaded

library(leaps) # for subset selection

## Warning: package 'leaps' was built under R version 4.3.2

library(caret) # test for correlation

## Warning: package 'caret' was built under R version 4.3.2

## Loading required package: lattice

library(car) # for VIF

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

library(scales) # for visualizing plots in %

##   
## Attaching package: 'scales'

## The following object is masked from 'package:readr':  
##   
## col\_factor

library(forcats) # ordering data frames  
library(codebookr) # adding appendix to r code

## Warning: package 'codebookr' was built under R version 4.3.3

library(gtsummary) # creating tables

## Warning: package 'gtsummary' was built under R version 4.3.3

library(cardx) # to include statistic results

## Warning: package 'cardx' was built under R version 4.3.3

library(moments) # to calculate skewness and kurtosis  
library(tigris) # to access US geographic data

## Warning: package 'tigris' was built under R version 4.3.3

## To enable caching of data, set `options(tigris\_use\_cache = TRUE)`  
## in your R script or .Rprofile.

*Data load*

# load data  
all\_sdoh\_data <- read\_csv("data/sdoh\_data.csv")

## Warning: One or more parsing issues, call `problems()` on your data frame for details,  
## e.g.:  
## dat <- vroom(...)  
## problems(dat)

## Rows: 3229 Columns: 682  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (15): COUNTYFIPS, STATEFIPS, STATE, COUNTY, REGION, CAF\_ADJ\_COUNTY\_1, C...  
## dbl (664): YEAR, TERRITORY, ACS\_TOT\_POP\_WT, ACS\_TOT\_POP\_US\_ABOVE1, ACS\_TOT\_P...  
## lgl (3): CAF\_ADJ\_COUNTY\_12, CAF\_ADJ\_COUNTY\_13, CAF\_ADJ\_COUNTY\_14  
##   
## 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.

dim(all\_sdoh\_data)

## [1] 3229 682

all\_chr\_data <- read\_csv("data/chr\_data.csv", skip = 1)

## Rows: 3194 Columns: 720  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (2): state, county  
## dbl (572): statecode, countycode, fipscode, year, county\_ranked, v001\_rawval...  
## lgl (146): v002\_numerator, v002\_denominator, v036\_numerator, v036\_denominato...  
##   
## 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.

dim(all\_chr\_data)

## [1] 3194 720

# data sets are extremely large, it is determined to remove unwanted features before any analysis

*Initial data cleaning*

# remove unwanted features, create calculated feature, convert fips\_code to data type matching chr\_data  
  
sdoh\_data <- all\_sdoh\_data %>%   
 select("COUNTYFIPS",   
 "STATE",   
 "COUNTY",   
 "REGION",   
 "ACS\_TOT\_POP\_WT",   
 "ACS\_AVG\_HH\_SIZE",   
 "ACS\_PCT\_MALE",   
 "ACS\_PCT\_FEMALE",   
 "ACS\_PCT\_NON\_CITIZEN",   
 "ACS\_PCT\_CTZ\_NATURALIZED",   
 "ACS\_PCT\_CTZ\_ABOVE18",   
 "ACS\_PCT\_ENGL\_NOT\_ALL",   
 "ACS\_PCT\_AIAN",   
 "ACS\_PCT\_ASIAN",   
 "ACS\_PCT\_BLACK",   
 "ACS\_PCT\_HISPANIC",   
 "ACS\_PCT\_OTHER\_RACE",   
 "ACS\_PCT\_WHITE",   
 "ACS\_PCT\_CHILD\_1FAM",   
 "ACS\_PCT\_CHILDREN\_GRANDPARENT",   
 "ACS\_PCT\_GRANDP\_RESPS\_NO\_P",   
 "ACS\_PCT\_GRANDP\_RESPS\_P",   
 "ACS\_PCT\_HH\_NO\_COMP\_DEV",   
 "ACS\_PCT\_HH\_SMARTPHONE",   
 "ACS\_PCT\_HH\_TABLET",   
 "ACS\_PCT\_HH\_PC",   
 "ACS\_PCT\_HH\_OTHER\_COMP",   
 "ACS\_PCT\_HH\_INTERNET",   
 "ACS\_PCT\_HH\_BROADBAND\_ANY",   
 "ACS\_PCT\_HH\_CELLULAR",   
 "ACS\_PCT\_HH\_NO\_INTERNET",   
 "ACS\_PCT\_HH\_SAT\_INTERNET",   
 "ACS\_PCT\_HH\_DIAL\_INTERNET\_ONLY",   
 "ACS\_PCT\_ADMIN",   
 "ACS\_PCT\_ART",   
 "ACS\_PCT\_CONSTRUCT",   
 "ACS\_PCT\_EDUC",   
 "ACS\_PCT\_FINANCE",   
 "ACS\_PCT\_GOVT",   
 "ACS\_PCT\_INFORM",   
 "ACS\_PCT\_MANUFACT",   
 "ACS\_PCT\_NATURE",   
 "ACS\_PCT\_OTHER",   
 "ACS\_PCT\_PROFESS",   
 "ACS\_PCT\_PVT\_NONPROFIT",   
 "ACS\_PCT\_RETAIL",   
 "ACS\_PCT\_TRANSPORT",   
 "ACS\_PCT\_WHOLESALE",   
 "ACS\_PCT\_EMPLOYED",   
 "ACS\_PCT\_UNEMPLOY",   
 "ACS\_GINI\_INDEX",   
 "ACS\_PCT\_HH\_INC\_10000",   
 "ACS\_PCT\_HH\_INC\_100000",   
 "ACS\_PCT\_HH\_INC\_14999",   
 "ACS\_PCT\_HH\_INC\_24999",   
 "ACS\_PCT\_HH\_INC\_49999",   
 "ACS\_PCT\_HH\_INC\_99999",   
 "ACS\_PER\_CAPITA\_INC",   
 "ACS\_MEDIAN\_HOME\_VALUE",   
 "ACS\_MEDIAN\_RENT",   
 "ACS\_PCT\_VACANT\_HU",   
 "ACS\_PCT\_COMMT\_15MIN",   
 "ACS\_PCT\_COMMT\_29MIN",   
 "ACS\_PCT\_COMMT\_59MIN",   
 "ACS\_PCT\_COMMT\_60MINUP",   
 "ACS\_PCT\_PUBL\_TRANSIT",   
 "ACS\_PCT\_MEDICAID\_ANY",   
 "ACS\_PCT\_MEDICARE\_ONLY",   
 "AHRF\_ADV\_NURSES\_RATE",   
 "AHRF\_CLIN\_NURSE\_SPEC\_RATE",   
 "AHRF\_DENTISTS\_RATE",   
 "AHRF\_NURSE\_ANESTH\_RATE",   
 "AHRF\_NURSE\_MIDWIVES\_RATE",   
 "AHRF\_NURSE\_PRACT\_RATE",   
 "AHRF\_PHYSICIAN\_ASSIST\_RATE",   
 "AMFAR\_SSP\_RATE",   
 "AMFAR\_MEDSAFAC\_RATE",  
 "AMFAR\_MHFAC\_RATE",  
 "CEN\_AREALAND\_SQM\_COUNTY",  
 "CEN\_POPDENSITY\_COUNTY",  
 "NEPHTN\_HEATIND\_90",   
 "EPAA\_2NDMAX\_CO\_1HR",  
 "EPAA\_98PR\_NO2\_1HR",  
 "EPAA\_MAX\_PB\_3MON",  
 "EPAA\_98PR\_PM25\_DAILY",  
 "EPAA\_99PR\_SO2\_1HR",   
 "SAIPE\_MEDIAN\_HH\_INCOME",  
 "SAIPE\_PCT\_POV",  
 "LTC\_AVG\_OBS\_REHOSP\_RATE",  
 "LTC\_AVG\_OBS\_SUCCESSFUL\_DISC\_RATE",  
 "MGV\_PER\_CAPITA\_STD\_IP",  
 "MGV\_PER\_CAPITA\_STD\_OP",  
 "MGV\_PER\_CAPITA\_STD\_EM",  
 "MGV\_PER\_CAPITA\_STD\_PA",  
 "MGV\_PER\_CAPITA\_STD\_HC",   
 "POS\_MEDIAN\_DIST\_ED",   
 "POS\_MEDIAN\_DIST\_MEDSURG\_ICU",  
 "POS\_MEDIAN\_DIST\_TRAUMA",  
 "POS\_MEDIAN\_DIST\_PED\_ICU",   
 "POS\_MEDIAN\_DIST\_OBSTETRICS",  
 "POS\_MEDIAN\_DIST\_CLINIC",   
 "POS\_MEDIAN\_DIST\_ALC",   
 ) %>%   
 mutate(percent\_grandparents\_as\_guardians = ACS\_PCT\_CHILDREN\_GRANDPARENT \* ((ACS\_PCT\_GRANDP\_RESPS\_P + ACS\_PCT\_GRANDP\_RESPS\_NO\_P)/100)) %>%   
 select(-ACS\_PCT\_GRANDP\_RESPS\_P, -ACS\_PCT\_GRANDP\_RESPS\_NO\_P, -ACS\_PCT\_CHILDREN\_GRANDPARENT) %>%   
 rename("fips\_code" = "COUNTYFIPS",  
 "state" = "STATE",  
 "county" = "COUNTY",  
 "region" = "REGION",  
 "weighted\_population" = "ACS\_TOT\_POP\_WT",  
 "average\_hh\_size" = "ACS\_AVG\_HH\_SIZE",  
 "pct\_male" = "ACS\_PCT\_MALE",  
 "pct\_female" = "ACS\_PCT\_FEMALE",  
 "pct\_not\_citizens" = "ACS\_PCT\_NON\_CITIZEN",  
 "pct\_naturalized\_citizens" = "ACS\_PCT\_CTZ\_NATURALIZED",  
 "pct\_adult\_citizens" = "ACS\_PCT\_CTZ\_ABOVE18",  
 "pct\_no\_english\_spoken" = "ACS\_PCT\_ENGL\_NOT\_ALL",  
 "pct\_native\_american" = "ACS\_PCT\_AIAN",  
 "pct\_asian" = "ACS\_PCT\_ASIAN",  
 "pct\_black" = "ACS\_PCT\_BLACK",  
 "pct\_hispanic" = "ACS\_PCT\_HISPANIC",  
 "pct\_other\_race" = "ACS\_PCT\_OTHER\_RACE",  
 "pct\_white" = "ACS\_PCT\_WHITE",  
 "pct\_single\_parent" = "ACS\_PCT\_CHILD\_1FAM",  
 "pct\_hh\_no\_computing\_device" = "ACS\_PCT\_HH\_NO\_COMP\_DEV",  
 "pct\_hh\_smartphone" = "ACS\_PCT\_HH\_SMARTPHONE",  
 "pct\_hh\_tablet" = "ACS\_PCT\_HH\_TABLET",  
 "pct\_hh\_computer" = "ACS\_PCT\_HH\_PC",  
 "pct\_hh\_other\_computer" = "ACS\_PCT\_HH\_OTHER\_COMP",  
 "pct\_hh\_internet" = "ACS\_PCT\_HH\_INTERNET",  
 "pct\_hh\_broadband" = "ACS\_PCT\_HH\_BROADBAND\_ANY",  
 "pct\_hh\_cell\_data" = "ACS\_PCT\_HH\_CELLULAR",  
 "pct\_hh\_no\_internet" = "ACS\_PCT\_HH\_NO\_INTERNET",  
 "pct\_hh\_satellite" = "ACS\_PCT\_HH\_SAT\_INTERNET",  
 "pct\_hh\_dial\_up" = "ACS\_PCT\_HH\_DIAL\_INTERNET\_ONLY",  
 "pct\_employed\_admin" = "ACS\_PCT\_ADMIN",  
 "pct\_employed\_arts" = "ACS\_PCT\_ART",   
 "pct\_employed\_construction" = "ACS\_PCT\_CONSTRUCT",  
 "pct\_employed\_education" = "ACS\_PCT\_EDUC",  
 "pct\_employed\_finance" = "ACS\_PCT\_FINANCE",  
 "pct\_employed\_government" = "ACS\_PCT\_GOVT",  
 "pct\_employed\_information" = "ACS\_PCT\_INFORM",  
 "pct\_employed\_manufacturing" = "ACS\_PCT\_MANUFACT",  
 "pct\_employed\_nature" = "ACS\_PCT\_NATURE",  
 "pct\_employed\_other" = "ACS\_PCT\_OTHER",  
 "pct\_employed\_professional" = "ACS\_PCT\_PROFESS",  
 "pct\_employed\_nonprofit" = "ACS\_PCT\_PVT\_NONPROFIT",  
 "pct\_employed\_retail" = "ACS\_PCT\_RETAIL",  
 "pct\_employed\_transportation" = "ACS\_PCT\_TRANSPORT",  
 "pct\_employed\_wholesale" = "ACS\_PCT\_WHOLESALE",  
 "pct\_employed" = "ACS\_PCT\_EMPLOYED",  
 "pct\_unemployed" = "ACS\_PCT\_UNEMPLOY",  
 "gini\_index" = "ACS\_GINI\_INDEX",  
 "pct\_hh\_inc\_10,000" = "ACS\_PCT\_HH\_INC\_10000",  
 "pct\_hh\_inc\_100,000" = "ACS\_PCT\_HH\_INC\_100000",  
 "pct\_hh\_inc\_14,999" = "ACS\_PCT\_HH\_INC\_14999",  
 "pct\_hh\_inc\_24,999" = "ACS\_PCT\_HH\_INC\_24999",  
 "pct\_hh\_inc\_49,999" = "ACS\_PCT\_HH\_INC\_49999",  
 "pct\_hh\_inc\_99999" = "ACS\_PCT\_HH\_INC\_99999", # renamed by mg  
 "per\_capita\_income" = "ACS\_PER\_CAPITA\_INC",  
 "median\_home\_value" = "ACS\_MEDIAN\_HOME\_VALUE",  
 "median\_rent" = "ACS\_MEDIAN\_RENT",  
 "pct\_houses\_vacant" = "ACS\_PCT\_VACANT\_HU",  
 "pct\_15\_min\_commute" = "ACS\_PCT\_COMMT\_15MIN",  
 "pct\_29\_min\_commute" = "ACS\_PCT\_COMMT\_29MIN",  
 "pct\_59\_min\_commute" = "ACS\_PCT\_COMMT\_59MIN",  
 "pct\_60\_min\_plus\_commute" = "ACS\_PCT\_COMMT\_60MINUP",  
 "pct\_public\_transportatin" = "ACS\_PCT\_PUBL\_TRANSIT",  
 "pct\_w\_medicaid" = "ACS\_PCT\_MEDICAID\_ANY",  
 "pct\_w\_medicare" = "ACS\_PCT\_MEDICARE\_ONLY",  
 "adv\_practice\_nurse\_pt" = "AHRF\_ADV\_NURSES\_RATE",  
 "clinical\_nurse\_pt" = "AHRF\_CLIN\_NURSE\_SPEC\_RATE",  
 "dentist\_pt" = "AHRF\_DENTISTS\_RATE",  
 "anesthetist\_nurse\_pt" = "AHRF\_NURSE\_ANESTH\_RATE",  
 "midwife\_pt" = "AHRF\_NURSE\_MIDWIVES\_RATE",  
 "nurse\_practitioner\_pt" = "AHRF\_NURSE\_PRACT\_RATE",  
 "pa\_pt" = "AHRF\_PHYSICIAN\_ASSIST\_RATE",  
 "syringe\_exchange\_pt" = "AMFAR\_SSP\_RATE",  
 "substance\_abuse\_facility\_pt" = "AMFAR\_MEDSAFAC\_RATE",  
 "mental\_health\_faciliy\_pt" = "AMFAR\_MHFAC\_RATE",  
 "land\_area\_sqm" = "CEN\_AREALAND\_SQM\_COUNTY",  
 "population\_density" = "CEN\_POPDENSITY\_COUNTY",  
 "days\_over\_90\_f" = "NEPHTN\_HEATIND\_90",  
 "co\_measure" = "EPAA\_2NDMAX\_CO\_1HR",  
 "no2\_measure" = "EPAA\_98PR\_NO2\_1HR",  
 "pb\_measure" = "EPAA\_MAX\_PB\_3MON",  
 "pm\_2.5\_measure" = "EPAA\_98PR\_PM25\_DAILY",  
 "so2\_measure" = "EPAA\_99PR\_SO2\_1HR",  
 "median\_hh\_income" = "SAIPE\_MEDIAN\_HH\_INCOME",  
 "pct\_people\_in\_poverty" = "SAIPE\_PCT\_POV",  
 "rehospitalization\_rate" = "LTC\_AVG\_OBS\_REHOSP\_RATE",  
 "successful\_discharge\_rate" = "LTC\_AVG\_OBS\_SUCCESSFUL\_DISC\_RATE",  
 "medicare\_inpatient\_payment" = "MGV\_PER\_CAPITA\_STD\_IP" ,  
 "medicare\_outpatient\_payment" = "MGV\_PER\_CAPITA\_STD\_OP",  
 "medicare\_e&m\_payment" = "MGV\_PER\_CAPITA\_STD\_EM",  
 "medicare\_acute\_care\_payment" = "MGV\_PER\_CAPITA\_STD\_PA",  
 "medicare\_fqrc\_rhc\_payment" = "MGV\_PER\_CAPITA\_STD\_HC",  
 "median\_er\_dist" = "POS\_MEDIAN\_DIST\_ED",  
 "median\_surgery\_dist" = "POS\_MEDIAN\_DIST\_MEDSURG\_ICU",  
 "median\_trauma\_center\_dist" = "POS\_MEDIAN\_DIST\_TRAUMA",  
 "median\_pediatric\_icu\_dist" = "POS\_MEDIAN\_DIST\_PED\_ICU",  
 "median\_obstetrics\_dist" = "POS\_MEDIAN\_DIST\_OBSTETRICS",  
 "median\_health\_clinic\_dist" = "POS\_MEDIAN\_DIST\_CLINIC",  
 "median\_drug\_alcohol\_care\_dist" = "POS\_MEDIAN\_DIST\_ALC"  
   
 ) %>%   
 mutate(fips\_code = as.numeric(fips\_code))

# remove unwanted features  
# convert principal care providers from per 100,000 people to per 1,000 people to match other data  
  
chr\_data <- all\_chr\_data %>%  
 select("fipscode",  
 "v002\_rawvalue",  
 "v042\_rawvalue",  
 "v037\_rawvalue",  
 "v009\_rawvalue",  
 "v011\_rawvalue",  
 "v133\_rawvalue",  
 "v070\_rawvalue",  
 "v132\_rawvalue",   
 "v049\_rawvalue",   
 "v014\_rawvalue",   
 "v085\_rawvalue",   
 "v062\_rawvalue",  
 "v050\_rawvalue",  
 "v155\_rawvalue",   
 "v168\_rawvalue",   
 "v069\_rawvalue",   
 "v023\_rawvalue",   
 "v024\_rawvalue",   
 "v044\_rawvalue",   
 "v140\_rawvalue",  
 "v135\_rawvalue",  
 "v125\_rawvalue",  
 "v124\_rawvalue",  
 "v136\_other\_data\_1",  
 "v136\_other\_data\_2",  
 "v136\_other\_data\_3",  
 "v137\_rawvalue",  
 "v147\_rawvalue",  
 "v127\_rawvalue",  
 "v128\_rawvalue",  
 "v129\_rawvalue",  
 "v144\_rawvalue",  
 "v061\_rawvalue",  
 "v139\_rawvalue",  
 "v138\_rawvalue",  
 "v143\_rawvalue",   
 "v021\_rawvalue",  
 "v149\_rawvalue",   
 "v159\_rawvalue",   
 "v160\_rawvalue",  
 "v167\_rawvalue",   
 "v169\_rawvalue",   
 "v151\_rawvalue",   
 "v063\_rawvalue",  
 "v170\_rawvalue",  
 "v065\_rawvalue",  
 "v141\_rawvalue",  
 "v171\_rawvalue",   
 "v015\_rawvalue",   
 "v161\_rawvalue",   
 "v148\_rawvalue",   
 "v158\_rawvalue",   
 "v177\_rawvalue",   
 "v156\_rawvalue",  
 "v153\_numerator",   
 "v052\_rawvalue",   
 "v053\_rawvalue",   
 "v058\_rawvalue",   
 "v004\_rawvalue",   
 "v005\_rawvalue"  
 ) %>%   
 mutate(pcp\_pt = v004\_rawvalue/100) %>%   
 select(-v004\_rawvalue) %>%   
 rename("fips\_code" = "fipscode",  
 "pct\_poor\_to\_fair\_health" = "v002\_rawvalue",  
 "pct\_adult\_smokers" = "v009\_rawvalue",  
 "pct\_obese\_adults" = "v011\_rawvalue",  
 "pct\_no\_exercise" = "v070\_rawvalue",  
 "pct\_binge\_drinkers" = "v049\_rawvalue",  
 "pct\_under\_65\_no\_health\_insurance" = "v085\_rawvalue",  
 "pct\_highschool\_diploma" = "v168\_rawvalue",  
 "pct\_some\_college" = "v069\_rawvalue",  
 "pct\_adult\_poverty" = "v024\_rawvalue",  
 "inequality\_ratio" = "v044\_rawvalue",  
 "social\_clubs\_per\_10k" = "v140\_rawvalue",  
 "air\_polution\_metric" = "v125\_rawvalue",  
 "water\_quality" = "v124\_rawvalue", # renamed by mg  
 "pct\_high\_housing\_costs" = "v136\_other\_data\_1",  
 "pct\_overcrowded\_hh" = "v136\_other\_data\_2",  
 "pct\_no\_kitchen\_or\_plumbinmg" = "v136\_other\_data\_3",  
 "pct\_food\_insecurities" = "v139\_rawvalue",  
 "pct\_insufficient\_sleep" = "v143\_rawvalue",  
 "school\_funding\_gap" = "v169\_rawvalue",  
 "pct\_income\_to\_childcare" = "v171\_rawvalue",  
 "pct\_voters" = "v177\_rawvalue",  
 "pct\_home\_owner" = "v153\_numerator",  
 "pct\_0\_17\_age" = "v052\_rawvalue",  
 "pct\_65\_plus" = "v053\_rawvalue",  
 "pct\_rural\_population" = "v058\_rawvalue",  
 "poor\_mental\_health" = "v042\_rawvalue",  
 "pct\_low\_birthweight" = "v037\_rawvalue",  
 "food\_enviroment" = "v133\_rawvalue",  
 "pct\_access\_to\_exercise" = "v132\_rawvalue",  
 "teen\_births\_prk\_1k" = "v014\_rawvalue",  
 "mental\_health\_providers\_per\_100k" = "v062\_rawvalue",  
 "hospital\_stay\_per\_100k" = "v005\_rawvalue",  
 "pct\_elderly\_mmmograms" = "v050\_rawvalue",  
 "pct\_flu\_vaccines\_billed" = "v155\_rawvalue",  
 "pct\_unemployed" = "v023\_rawvalue",  
 "injury\_death\_rate\_per\_100k" = "v135\_rawvalue",  
 "life\_expectancy\_years" = "v147\_rawvalue",  
 "premature\_deaths\_per\_100k" = "v127\_rawvalue",  
 "underage\_deaths\_per\_100k" = "v128\_rawvalue",  
 "infant\_deaths\_per\_1k\_births" = "v129\_rawvalue",  
 "pct\_poor\_health" = "v144\_rawvalue",  
 "pct\_hiv" = "v061\_rawvalue",  
 "drug\_overdose\_per\_100k" = "v138\_rawvalue",  
 "pct\_insufficieficient\_sleep" = "v143\_rawvalue",  
 "pct\_on\_time\_hs\_graduation" = "v021\_rawvalue",  
 "pct\_disconnected\_youth" = "v149\_rawvalue",  
 "children\_reading\_score" = "v159\_rawvalue",  
 "children\_math\_score" = "v160\_rawvalue",  
 "school\_segregation" = "v167\_rawvalue",  
 "women\_to\_man\_pay\_ratio" = "v151\_rawvalue",  
 "median\_hh\_income" = "v063\_rawvalue",  
 "hourly\_living\_wage" = "v170\_rawvalue",  
 "children\_eligible\_for\_lunch" = "v065\_rawvalue",  
 "black\_white\_segregation" = "v141\_rawvalue",  
 "homicides\_per\_100k" = "v015\_rawvalue",  
 "suicides\_per\_100k" = "v161\_rawvalue",  
 "firearm\_fatalities\_per\_100k" = "v148\_rawvalue",  
 "juvenile\_arrests\_per\_1k" = "v158\_rawvalue",  
 "traffic\_per\_meter" = "v156\_rawvalue",  
 "pct\_30\_min\_plus\_commute" = "v137\_rawvalue")

rm(all\_chr\_data)  
rm(all\_sdoh\_data)

*Combine datasets*

# Create and clean the qol\_data dataset  
qol\_data <- sdoh\_data %>%  
 inner\_join(chr\_data, by = "fips\_code") %>%  
 mutate(response = ifelse(pct\_poor\_to\_fair\_health >= 0.12, "worse", "better")) %>%  
 mutate(response = as.factor(response)) %>%  
 #select(-pct\_poor\_to\_fair\_health) %>% # keep until analysis has been performed  
 mutate\_at(vars(state, county, region), as.factor) # convert characters to factors

*Features with NA’s*

# sum of NAs in each column  
na\_counts <- colSums(is.na(qol\_data))  
  
# combine column names and NA counts into a dataframe  
na\_counts\_df <- data.frame(variable\_name = names(na\_counts), na\_count = na\_counts)  
  
# Sort the dataframe by NA\_Count in descending order  
na\_counts\_df <- na\_counts\_df[order(-na\_counts\_df$na\_count), ]  
  
# Convert the dataframe to a tibble for better printing  
na\_counts\_tbl <- as\_tibble(na\_counts\_df)  
  
# Print the tibble  
print(na\_counts\_tbl)

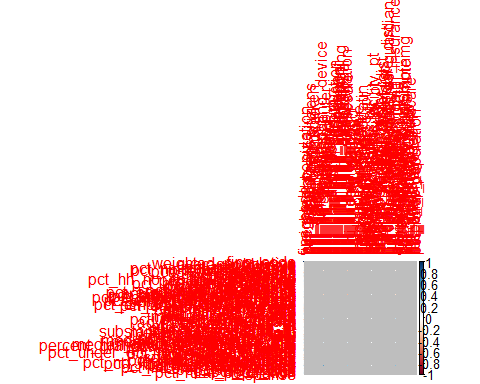
## # A tibble: 161 × 2  
## variable\_name na\_count  
## <chr> <dbl>  
## 1 hourly\_living\_wage 3142  
## 2 pb\_measure 3106  
## 3 co\_measure 2988  
## 4 no2\_measure 2915  
## 5 so2\_measure 2860  
## 6 pm\_2.5\_measure 2616  
## 7 pct\_disconnected\_youth 1938  
## 8 infant\_deaths\_per\_1k\_births 1925  
## 9 homicides\_per\_100k 1816  
## 10 drug\_overdose\_per\_100k 1345  
## # ℹ 151 more rows

# Due to large amount of NA values in these observations, the following will be removed as no clear value can be used to replace the NA's and there are reasonable alternatives to that predictor:  
# hourly\_living\_wage, pb\_measure, co\_measure, no2\_measure, so2\_measure, pm\_2.5\_measure, pct\_disconnected\_youth, infant\_deaths\_per\_1k\_births, homicides\_per\_100k, drug\_overdose\_per\_100k, underage\_deaths\_per\_100k, juvenile\_arrests\_per\_1k, black\_white\_segregation, firearm\_fatalities\_per\_100k, pct\_on\_time\_hs\_graduation, suicides\_per\_100k, children\_eligible\_for\_lunch, pct\_hiv, children\_math\_score, children\_reading\_score, successful\_discharge\_rate, rehospitalization\_rate, school\_segregation\_0:1\_\_low:high, mental\_health\_providers\_per\_100k, teen\_births\_prk\_1k, traffic\_per\_meter, pcp\_pt, pct\_low\_birthweight, injury\_death\_rate\_per\_100k, hospital\_stay\_per\_100k, premature\_deaths\_per\_100k  
  
# This still leaves a large number of predictors which can be further winnowed down due to duplication or near duplication between data-sets:  
# pct\_unemployed.y, median\_hh\_income.y, pct\_15\_min\_commute, pct\_29\_min\_commute, pct\_59\_min\_commute, pct\_60\_min\_plus\_commute, pct\_access\_to\_exercise, poor\_mental\_health, life\_expectancy\_years, food\_enviroment\_1:10\_bad:good, pct\_poor\_health  
  
# Others will be removed due to the data being obscure for our purposes:  
# medicare\_inpatient\_payment, medicare\_outpatient\_payment, medicare\_e&m\_payment, medicare\_acute\_care\_payment, medicare\_fqrc/rhc\_payment, pct\_elderly\_mmmograms, pct\_flu\_vaccines\_billed, pct\_insufficieficient\_sleep, women\_to\_man\_pay\_ratio

qol\_data <- qol\_data %>%   
 select(  
 -hourly\_living\_wage,   
 -pb\_measure,   
 -co\_measure,   
 -no2\_measure,   
 -so2\_measure,   
 -pm\_2.5\_measure,   
 -pct\_disconnected\_youth,   
 -infant\_deaths\_per\_1k\_births,   
 -homicides\_per\_100k,   
 -drug\_overdose\_per\_100k,   
 -underage\_deaths\_per\_100k,   
 -juvenile\_arrests\_per\_1k,   
 -black\_white\_segregation,   
 -firearm\_fatalities\_per\_100k,   
 -pct\_on\_time\_hs\_graduation,   
 -suicides\_per\_100k,   
 -children\_eligible\_for\_lunch,   
 -pct\_hiv,   
 -children\_math\_score,   
 -children\_reading\_score,   
 -successful\_discharge\_rate,   
 -rehospitalization\_rate,   
 -school\_segregation,   
 -mental\_health\_providers\_per\_100k,   
 -teen\_births\_prk\_1k,   
 -traffic\_per\_meter,   
 -pcp\_pt,   
 -pct\_low\_birthweight,   
 -injury\_death\_rate\_per\_100k,   
 -hospital\_stay\_per\_100k,   
 -premature\_deaths\_per\_100k,  
 -pct\_unemployed.y,   
 -median\_hh\_income.y,   
 -pct\_15\_min\_commute,   
 -pct\_29\_min\_commute,   
 -pct\_59\_min\_commute,   
 -pct\_60\_min\_plus\_commute,   
 -pct\_access\_to\_exercise,   
 -pct\_poor\_health,   
 # -life\_expectancy\_years, # keep for initial analysis   
 -food\_enviroment,   
 -poor\_mental\_health,  
 -medicare\_inpatient\_payment,   
 -medicare\_outpatient\_payment,   
 -matches("medicare\_e&m\_payment"),   
 -medicare\_acute\_care\_payment,   
 -medicare\_fqrc\_rhc\_payment,   
 -pct\_elderly\_mmmograms,   
 -pct\_flu\_vaccines\_billed,   
 -pct\_insufficieficient\_sleep,   
 -women\_to\_man\_pay\_ratio   
 ) %>%   
 na.omit()

*Correlation*

# find predictors with high correlation to shrink the model  
  
# subset qol\_data to include only numeric variables  
# identify values with variance inflation factors  
qol\_numeric <- qol\_data %>%  
 mutate(response = if\_else(response == "worse", 0, 1)) %>%   
 select(-state,  
 -county,  
 -region)  
  
# calculate the correlation matrix  
cor\_matrix <- cor(qol\_numeric)  
  
# plot correlations to visualize  
corrplot(cor\_matrix, method = "circle") # data too dense to visualize



# find the indices of correlations greater than 0.7  
high\_cor\_indices <- which(abs(cor\_matrix) > 0.7, arr.ind = TRUE) # 0.8 is more commonly accepted, also run regression andview variance inflation factors with logistic regression  
  
# extract the pairs of variables with correlation greater than 0.7  
high\_cor\_pairs <- data.frame(  
 var1 = rownames(cor\_matrix)[high\_cor\_indices[, 1]],  
 var2 = colnames(cor\_matrix)[high\_cor\_indices[, 2]],  
 correlation = cor\_matrix[high\_cor\_indices]  
)  
  
# filter out duplicates and self-correlations  
high\_cor\_pairs <- high\_cor\_pairs[high\_cor\_pairs$var1 != high\_cor\_pairs$var2, ]  
high\_cor\_pairs <- high\_cor\_pairs[!duplicated(t(apply(high\_cor\_pairs, 1, sort))), ]  
  
high\_cor\_pairs\_tbl <- as\_tibble(high\_cor\_pairs)  
  
# Print the tibble  
print(high\_cor\_pairs\_tbl)

## # A tibble: 120 × 3  
## var1 var2 correlation  
## <chr> <chr> <dbl>  
## 1 pct\_home\_owner weighted\_population 0.977  
## 2 pct\_female pct\_male -1.00   
## 3 pct\_naturalized\_citizens pct\_not\_citizens 0.746  
## 4 pct\_adult\_citizens pct\_not\_citizens -0.760  
## 5 pct\_no\_english\_spoken pct\_not\_citizens 0.804  
## 6 pct\_hispanic pct\_not\_citizens 0.711  
## 7 pct\_asian pct\_naturalized\_citizens 0.721  
## 8 pct\_0\_17\_age pct\_adult\_citizens -0.785  
## 9 pct\_hispanic pct\_no\_english\_spoken 0.703  
## 10 pct\_white pct\_black -0.831  
## # ℹ 110 more rows

# Remove predictors with large correlation value (|0.7|) or greater with multi-colinearity and low relevance.

qol\_data <- qol\_data %>%  
 select(-adv\_practice\_nurse\_pt,  
 -pct\_0\_17\_age,  
 -pct\_adult\_poverty,  
 -pct\_female,  
 -pct\_hh\_broadband,  
 -pct\_hh\_cell\_data,  
 -matches("pct\_hh\_inc\_10,000"),  
 -matches("pct\_hh\_inc\_100,000"),  
 -matches("pct\_hh\_inc\_14,999"),  
 -matches("pct\_hh\_inc\_24,999"),  
 -matches("pct\_hh\_inc\_49,999"),  
 -pct\_hh\_no\_computing\_device,  
 -pct\_hh\_no\_internet,  
 -pct\_hh\_smartphone,  
 -pct\_hh\_tablet,  
 -pct\_no\_exercise,  
 -pct\_not\_citizens,  
 -pct\_people\_in\_poverty,  
 -pct\_unemployed.x,  
 -per\_capita\_income,  
 -weighted\_population,  
 -pct\_food\_insecurities,  
 -pct\_hh\_computer,  
 -pct\_naturalized\_citizens,  
 -median\_rent,  
 -median\_home\_value,  
 -pct\_w\_medicaid  
 )

*Near zero variance*

# identify predictors with near-zero variance (high collinearity)  
nzv <- nearZeroVar(qol\_numeric, saveMetrics = TRUE)  
  
# select for predictors with near zero variance  
nzv <- nzv[nzv$nzv == TRUE, ]  
print(nzv)

## freqRatio percentUnique zeroVar nzv  
## syringe\_exchange\_pt 447.3333 5.922166 FALSE TRUE

# remove features with near zero variance  
qol\_data <- qol\_data %>%  
 select(-syringe\_exchange\_pt  
 )

*Features with low interest*

# per group discussion, remove low interest features from model  
qol\_data <- qol\_data %>%  
 select(-pct\_adult\_citizens,  
 -pct\_no\_english\_spoken,  
 -pct\_hh\_satellite,  
 -pct\_hh\_dial\_up,  
 -pct\_employed\_admin,  
 -pct\_employed\_arts,  
 -pct\_employed\_construction,  
 -pct\_employed\_education,  
 -pct\_employed\_finance,  
 -pct\_employed\_government,  
 -pct\_employed\_information,  
 -pct\_employed\_manufacturing,  
 -pct\_employed\_nature,  
 -pct\_employed\_other,  
 -pct\_employed\_professional,  
 -pct\_employed\_nonprofit,  
 -pct\_employed\_retail,  
 -pct\_employed\_transportation,  
 -pct\_employed\_wholesale,  
 -gini\_index,  
 -pct\_houses\_vacant,  
 -pct\_public\_transportatin,  
 -anesthetist\_nurse\_pt,  
 -midwife\_pt,  
 -nurse\_practitioner\_pt,  
 -substance\_abuse\_facility\_pt,  
 -land\_area\_sqm,  
 -median\_surgery\_dist,  
 -median\_obstetrics\_dist,  
 -pct\_some\_college,  
 -pct\_no\_kitchen\_or\_plumbinmg,  
 -pct\_income\_to\_childcare,  
 )

*Variance inflation features*

# run a pca (DO this with 0, 1) or lda (prefered) on numeric data

qol\_numeric <- qol\_data %>%   
 select(-state,  
 -county,  
 -region,  
 -life\_expectancy\_years,  
 -fips\_code)  
  
  
  
vif\_values <- vif(lm(response ~ ., data = qol\_numeric))

## Warning in model.response(mf, "numeric"): using type = "numeric" with a factor  
## response will be ignored

## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors

## Warning in Ops.factor(r, 2): '^' not meaningful for factors

## Warning in cov2cor(v): diag(.) had 0 or NA entries; non-finite result is  
## doubtful

high\_vif <- names(vif\_values[vif\_values > 5])  
high\_vif

## [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  
## [26] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA

# no further feature removal needed at this point

*Subset regression viability check*

# identify best model for response with regsubset, tests for linear models   
#qol\_regfit\_full <- regsubsets(response ~ ., qol\_numeric,  
# really.big = TRUE,  
# nvmax = 45)

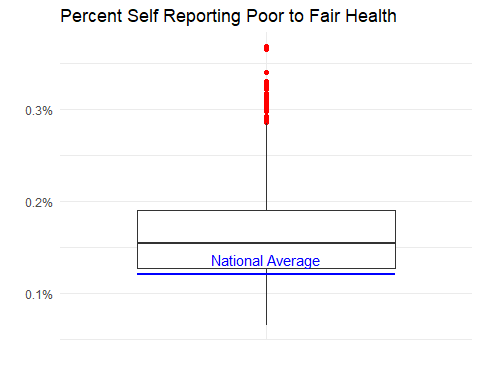
#reg\_fit\_summary <- summary(qol\_regfit\_full)  
  
# identifying ideal number of variables for each metric  
#which.min(reg\_fit\_summary$rss) # always selects for model with all predictors, over-fits to training data  
#which.max(reg\_fit\_summary$adjr2) # increases with additional predictors, susceptible to training error  
#which.min(reg\_fit\_summary$cp) # penalizes models with more predictors with unbiased measure of MSE  
#which.min(reg\_fit\_summary$bic) # like Cp, but includes penalty term log(n) in error so more error introduced with more predictors

# create plots of each metric to visualize   
  
#par(mfrow = c(2,2))  
  
#plot(reg\_fit\_summary$rss,  
# xlab="Number of Variables",  
# ylab="RSS",  
# type="l")  
#points(45, reg\_fit\_summary$rss[45], col="red",cex=1.5,pch =20)  
  
#plot(reg\_fit\_summary$adjr2,  
# xlab = "Number of Variables",  
# ylab = "Adjusted R-squared",  
# type = "l")  
#points(30, reg\_fit\_summary$adjr2[30], col="red",cex=1.5,pch =20)  
  
#plot(reg\_fit\_summary$cp,  
# xlab = "Number of Variables",  
# ylab = "Cp",  
# type = "l")  
#points(27, reg\_fit\_summary$cp[27], col="red",cex=1.5,pch =20)  
  
#plot(reg\_fit\_summary$bic,  
# xlab = "Number of Variables",  
# ylab = "BIC",  
# type = "l")  
#points (14, reg\_fit\_summary$bic[14], col = "red", cex = 1.5, pch = 20)

*Updating response*

ggplot(qol\_data,  
 mapping = aes(x = '', y = pct\_poor\_to\_fair\_health)) +  
 geom\_boxplot(outlier.colour = 'red') +  
 theme\_minimal() +  
 labs(x = "",  
 y = "",  
 title = "Percent Self Reporting Poor to Fair Health") +  
 scale\_y\_continuous(labels = scales::percent\_format(scale = 1)) +  
 annotate("segment",   
 x = 0.625, xend = 1.375,   
 y = 0.12, yend = 0.12,   
 color = "blue", size = 1) +  
 annotate("text",   
 x = 1, y = 0.137,   
 label = "National Average",   
 color = "blue",   
 size = 4,   
 hjust = 0.5)

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



# with the presence of outliers having high percent self reporting and a lower limit that is not able to be less tha, will use the

# identify new median  
median\_result <- median(qol\_data$pct\_poor\_to\_fair\_health)  
  
qol\_data <- qol\_data %>%  
 mutate(response = ifelse(pct\_poor\_to\_fair\_health >= median\_result, "worse", "better")) %>%  
 mutate(response = as.factor(response))  
  
# United States self reported poor to fair health is 12% per chr data. However, the median value of this result is 15.4% for included observations (not weighted)

*Data overview*

# inspect dataset structure and summary  
str(qol\_data)

## tibble [2,955 × 51] (S3: tbl\_df/tbl/data.frame)  
## $ fips\_code : num [1:2955] 1001 1003 1005 1007 1009 ...  
## $ state : Factor w/ 51 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ county : Factor w/ 1875 levels "Abbeville County",..: 84 91 102 151 166 227 237 250 298 320 ...  
## $ region : Factor w/ 4 levels "Midwest","Northeast",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ average\_hh\_size : num [1:2955] 2.55 2.56 2.37 2.85 2.7 2.84 2.92 2.5 2.42 2.38 ...  
## $ pct\_male : num [1:2955] 48.6 48.5 52.6 53.7 49.6 ...  
## $ pct\_native\_american : num [1:2955] 0.28 0.69 0.35 0.05 0.1 0 0.33 0.31 0.24 0.58 ...  
## $ pct\_asian : num [1:2955] 1.17 0.93 0.49 0.25 0.41 1.35 1.32 0.81 1.1 0.1 ...  
## $ pct\_black : num [1:2955] 19.53 8.77 47.67 22.55 1.4 ...  
## $ pct\_hispanic : num [1:2955] 2.88 4.56 4.44 2.68 9.28 8.1 1.47 3.85 2.52 1.66 ...  
## $ pct\_other\_race : num [1:2955] 0.67 1.56 3.1 0.04 1.8 3.13 0.5 1.64 0.79 0.81 ...  
## $ pct\_white : num [1:2955] 75.8 85.4 46.3 76.6 94 ...  
## $ pct\_single\_parent : num [1:2955] 27.4 18.1 52.8 32.2 25.8 ...  
## $ pct\_hh\_other\_computer : num [1:2955] 1.05 1.75 2.15 0.23 2.16 1.84 5.07 2.43 1.63 0.85 ...  
## $ pct\_hh\_internet : num [1:2955] 82.8 85.5 65 76.2 80 ...  
## $ pct\_employed : num [1:2955] 97.1 96.1 93.1 92.6 94.8 ...  
## $ pct\_hh\_inc\_99999 : num [1:2955] 30.4 30.5 23.6 34.9 28.6 ...  
## $ pct\_w\_medicare : num [1:2955] 6.87 7.27 7.51 7.74 8.81 5.87 7.49 5.18 7.36 9.26 ...  
## $ clinical\_nurse\_pt : num [1:2955] 0.02 0.02 0 0 0 0 0 0.01 0 0.04 ...  
## $ dentist\_pt : num [1:2955] 0.34 0.49 0.37 0.27 0.19 0.2 0.36 0.68 0.21 0.15 ...  
## $ pa\_pt : num [1:2955] 0.04 0.17 0.04 0.23 0 0.1 0 0.1 0 0 ...  
## $ mental\_health\_faciliy\_pt : num [1:2955] 0.0178 0.0174 0.0813 0 0.0173 ...  
## $ population\_density : num [1:2955] 93.6 137.3 28.3 35.9 89.6 ...  
## $ days\_over\_90\_f : num [1:2955] 104 97 104 97 80 103 103 84 87 85 ...  
## $ median\_hh\_income.x : num [1:2955] 67565 71135 38866 50907 55203 ...  
## $ median\_er\_dist : num [1:2955] 2.25 6.01 5.58 8.44 10.56 ...  
## $ median\_trauma\_center\_dist : num [1:2955] 12.1 25.4 41.4 26.4 26.2 ...  
## $ median\_pediatric\_icu\_dist : num [1:2955] 55.6 22.6 63.8 22 52.2 ...  
## $ median\_health\_clinic\_dist : num [1:2955] 9.2 8.57 1.17 3.85 3.91 1.63 2.83 4.08 4.42 6.57 ...  
## $ median\_drug\_alcohol\_care\_dist : num [1:2955] 12.2 13.2 35.7 26.5 25.2 ...  
## $ percent\_grandparents\_as\_guardians: num [1:2955] 4.9 5.96 11.17 12.54 6.23 ...  
## $ pct\_poor\_to\_fair\_health : num [1:2955] 0.169 0.149 0.275 0.216 0.184 0.297 0.227 0.196 0.215 0.193 ...  
## $ pct\_adult\_smokers : num [1:2955] 0.183 0.169 0.259 0.228 0.218 0.255 0.223 0.21 0.217 0.227 ...  
## $ pct\_obese\_adults : num [1:2955] 0.373 0.326 0.464 0.384 0.327 0.483 0.454 0.374 0.412 0.413 ...  
## $ pct\_binge\_drinkers : num [1:2955] 0.167 0.19 0.134 0.159 0.163 ...  
## $ pct\_under\_65\_no\_health\_insurance : num [1:2955] 0.106 0.109 0.144 0.13 0.133 ...  
## $ pct\_highschool\_diploma : num [1:2955] 0.896 0.91 0.757 0.805 0.836 ...  
## $ inequality\_ratio : num [1:2955] 4.79 4.3 5.18 5.03 4.8 ...  
## $ social\_clubs\_per\_10k : num [1:2955] 12.65 9.59 9.35 9.04 6.74 ...  
## $ air\_polution\_metric : num [1:2955] 10 7.6 9.4 9.8 9.6 9.3 9.1 9.7 9.8 9.5 ...  
## $ water\_quality : num [1:2955] 0 1 1 0 0 0 0 0 0 0 ...  
## $ pct\_high\_housing\_costs : num [1:2955] 0.1264 0.1056 0.1346 0.0799 0.0738 ...  
## $ pct\_overcrowded\_hh : num [1:2955] 0.0112 0.0129 0.0385 0.0116 0.018 ...  
## $ pct\_30\_min\_plus\_commute : num [1:2955] 0.416 0.376 0.365 0.551 0.595 0.494 0.347 0.3 0.281 0.456 ...  
## $ life\_expectancy\_years : num [1:2955] 76.6 77.7 72.9 73.6 74.2 ...  
## $ school\_funding\_gap : num [1:2955] -2077 343 -13560 -2660 -889 ...  
## $ pct\_voters : num [1:2955] 0.662 0.653 0.54 0.546 0.642 ...  
## $ pct\_home\_owner : num [1:2955] 16227 67242 5654 5580 16865 ...  
## $ pct\_65\_plus : num [1:2955] 0.16 0.215 0.2 0.167 0.187 ...  
## $ pct\_rural\_population : num [1:2955] 0.42 0.423 0.678 0.684 0.9 ...  
## $ response : Factor w/ 2 levels "better","worse": 2 1 2 2 2 2 2 2 2 2 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:187] 68 69 70 71 72 73 74 75 76 77 ...  
## ..- attr(\*, "names")= chr [1:187] "68" "69" "70" "71" ...

summary(qol\_data)

## fips\_code state county region   
## Min. : 1001 Texas : 234 Washington County: 30 Midwest :1018   
## 1st Qu.:18182 Georgia : 158 Jefferson County : 25 Northeast: 199   
## Median :29099 Kentucky: 120 Franklin County : 23 South :1354   
## Mean :30172 Missouri: 114 Lincoln County : 23 West : 384   
## 3rd Qu.:44002 Illinois: 102 Jackson County : 22   
## Max. :56045 Kansas : 102 Madison County : 19   
## (Other) :2125 (Other) :2813   
## average\_hh\_size pct\_male pct\_native\_american pct\_asian   
## Min. :1.840 Min. :41.99 Min. : 0.000 Min. : 0.000   
## 1st Qu.:2.330 1st Qu.:48.90 1st Qu.: 0.150 1st Qu.: 0.310   
## Median :2.460 Median :49.62 Median : 0.330 Median : 0.640   
## Mean :2.492 Mean :50.06 Mean : 1.632 Mean : 1.273   
## 3rd Qu.:2.610 3rd Qu.:50.57 3rd Qu.: 0.810 3rd Qu.: 1.290   
## Max. :4.350 Max. :69.54 Max. :87.310 Max. :31.250   
##   
## pct\_black pct\_hispanic pct\_other\_race pct\_white   
## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 9.35   
## 1st Qu.: 0.700 1st Qu.: 2.320 1st Qu.: 0.410 1st Qu.:75.23   
## Median : 2.390 Median : 4.350 Median : 1.050 Median :88.22   
## Mean : 9.076 Mean : 9.572 Mean : 2.211 Mean :82.30   
## 3rd Qu.:10.295 3rd Qu.: 9.875 3rd Qu.: 2.515 3rd Qu.:94.06   
## Max. :87.790 Max. :98.900 Max. :54.700 Max. :99.90   
##   
## pct\_single\_parent pct\_hh\_other\_computer pct\_hh\_internet pct\_employed   
## Min. : 1.25 Min. : 0.000 Min. :41.38 Min. : 69.61   
## 1st Qu.:23.19 1st Qu.: 1.295 1st Qu.:74.60 1st Qu.: 93.67   
## Median :29.01 Median : 1.830 Median :80.13 Median : 95.10   
## Mean :29.99 Mean : 2.110 Mean :78.98 Mean : 94.79   
## 3rd Qu.:35.22 3rd Qu.: 2.470 3rd Qu.:84.43 3rd Qu.: 96.32   
## Max. :93.95 Max. :17.850 Max. :96.81 Max. :100.00   
##   
## pct\_hh\_inc\_99999 pct\_w\_medicare clinical\_nurse\_pt dentist\_pt   
## Min. :13.50 Min. : 1.750 Min. :0.00000 Min. :0.0000   
## 1st Qu.:28.61 1st Qu.: 5.020 1st Qu.:0.00000 1st Qu.:0.2600   
## Median :31.49 Median : 6.190 Median :0.00000 Median :0.4200   
## Mean :31.13 Mean : 6.494 Mean :0.01304 Mean :0.4561   
## 3rd Qu.:33.94 3rd Qu.: 7.555 3rd Qu.:0.02000 3rd Qu.:0.6100   
## Max. :47.18 Max. :19.440 Max. :0.41000 Max. :7.4900   
##   
## pa\_pt mental\_health\_faciliy\_pt population\_density days\_over\_90\_f   
## Min. : 0.0000 Min. :0.00000 Min. : 0.44 Min. : 0.00   
## 1st Qu.: 0.0800 1st Qu.:0.00000 1st Qu.: 18.75 1st Qu.: 25.00   
## Median : 0.2300 Median :0.03680 Median : 45.70 Median : 55.00   
## Mean : 0.3117 Mean :0.05371 Mean : 213.44 Mean : 55.98   
## 3rd Qu.: 0.4300 3rd Qu.:0.07135 3rd Qu.: 115.72 3rd Qu.: 87.00   
## Max. :13.1400 Max. :1.01270 Max. :71895.54 Max. :144.00   
##   
## median\_hh\_income.x median\_er\_dist median\_trauma\_center\_dist  
## Min. : 25997 Min. : 0.130 Min. : 0.13   
## 1st Qu.: 47731 1st Qu.: 2.930 1st Qu.: 6.59   
## Median : 55010 Median : 4.970 Median : 18.13   
## Mean : 57130 Mean : 6.739 Mean : 20.56   
## 3rd Qu.: 63763 3rd Qu.: 8.235 3rd Qu.: 29.30   
## Max. :132509 Max. :61.450 Max. :136.88   
##   
## median\_pediatric\_icu\_dist median\_health\_clinic\_dist  
## Min. : 0.77 Min. : 0.130   
## 1st Qu.: 20.32 1st Qu.: 2.060   
## Median : 33.60 Median : 3.400   
## Mean : 37.81 Mean : 4.772   
## 3rd Qu.: 50.54 3rd Qu.: 5.470   
## Max. :172.30 Max. :57.560   
##   
## median\_drug\_alcohol\_care\_dist percent\_grandparents\_as\_guardians  
## Min. : 0.61 Min. : 0.000   
## 1st Qu.: 8.00 1st Qu.: 2.546   
## Median : 19.84 Median : 4.255   
## Mean : 23.08 Mean : 5.156   
## 3rd Qu.: 31.84 3rd Qu.: 6.841   
## Max. :148.80 Max. :28.906   
##   
## pct\_poor\_to\_fair\_health pct\_adult\_smokers pct\_obese\_adults pct\_binge\_drinkers  
## Min. :0.0650 Min. :0.0670 Min. :0.1760 Min. :0.08195   
## 1st Qu.:0.1270 1st Qu.:0.1760 1st Qu.:0.3380 1st Qu.:0.16755   
## Median :0.1540 Median :0.2000 Median :0.3680 Median :0.18814   
## Mean :0.1613 Mean :0.2016 Mean :0.3637 Mean :0.19032   
## 3rd Qu.:0.1900 3rd Qu.:0.2270 3rd Qu.:0.3930 3rd Qu.:0.21273   
## Max. :0.3680 Max. :0.4110 Max. :0.5320 Max. :0.28928   
##   
## pct\_under\_65\_no\_health\_insurance pct\_highschool\_diploma inequality\_ratio  
## Min. :0.02278 Min. :0.4967 Min. : 2.773   
## 1st Qu.:0.07989 1st Qu.:0.8462 1st Qu.: 4.002   
## Median :0.10691 Median :0.8914 Median : 4.409   
## Mean :0.11827 Mean :0.8785 Mean : 4.537   
## 3rd Qu.:0.14812 3rd Qu.:0.9215 3rd Qu.: 4.937   
## Max. :0.37345 Max. :0.9862 Max. :11.128   
##   
## social\_clubs\_per\_10k air\_polution\_metric water\_quality   
## Min. : 0.000 Min. : 0.900 Min. :0.0000   
## 1st Qu.: 8.087 1st Qu.: 6.600 1st Qu.:0.0000   
## Median :10.777 Median : 7.800 Median :0.0000   
## Mean :11.390 Mean : 7.615 Mean :0.3425   
## 3rd Qu.:14.000 3rd Qu.: 8.900 3rd Qu.:1.0000   
## Max. :48.860 Max. :15.600 Max. :1.0000   
##   
## pct\_high\_housing\_costs pct\_overcrowded\_hh pct\_30\_min\_plus\_commute  
## Min. :0.02103 Min. :0.00000 Min. :0.0000   
## 1st Qu.:0.07982 1st Qu.:0.01248 1st Qu.:0.2360   
## Median :0.09764 Median :0.01878 Median :0.3250   
## Mean :0.10150 Mean :0.02312 Mean :0.3311   
## 3rd Qu.:0.11909 3rd Qu.:0.02824 3rd Qu.:0.4185   
## Max. :0.24733 Max. :0.15427 Max. :0.7840   
##   
## life\_expectancy\_years school\_funding\_gap pct\_voters pct\_home\_owner   
## Min. : 64.11 Min. :-18852.8 Min. :0.1942 Min. : 371   
## 1st Qu.: 74.91 1st Qu.: -2420.8 1st Qu.:0.5883 1st Qu.: 3411   
## Median : 76.89 Median : 441.6 Median :0.6554 Median : 7577   
## Mean : 76.84 Mean : 247.0 Mean :0.6530 Mean : 26068   
## 3rd Qu.: 78.82 3rd Qu.: 2866.0 3rd Qu.:0.7215 3rd Qu.: 19303   
## Max. :103.31 Max. : 27719.2 Max. :1.0000 Max. :1545929   
##   
## pct\_65\_plus pct\_rural\_population response   
## Min. :0.05073 Min. :0.0000 better:1457   
## 1st Qu.:0.17014 1st Qu.:0.3392 worse :1498   
## Median :0.19606 Median :0.5891   
## Mean :0.19983 Mean :0.5831   
## 3rd Qu.:0.22408 3rd Qu.:0.8382   
## Max. :0.58171 Max. :1.0000   
##

*Create groups of interest*

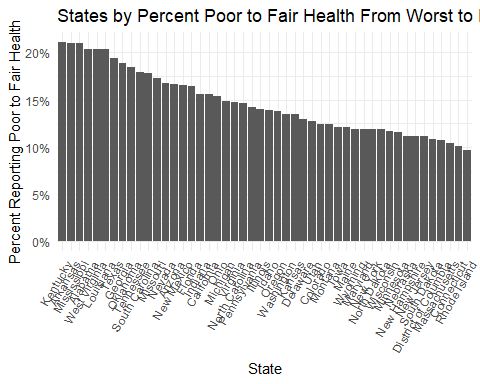
# qol dataset grouped by median values  
qol\_state\_median <- qol\_data %>%  
 group\_by(state) %>%  
 summarize(across(where(is.numeric), median, na.rm = TRUE))

## Warning: There was 1 warning in `summarize()`.  
## ℹ In argument: `across(where(is.numeric), median, na.rm = TRUE)`.  
## ℹ In group 1: `state = Alabama`.  
## Caused by warning:  
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.  
## Supply arguments directly to `.fns` through an anonymous function instead.  
##   
## # Previously  
## across(a:b, mean, na.rm = TRUE)  
##   
## # Now  
## across(a:b, \(x) mean(x, na.rm = TRUE))

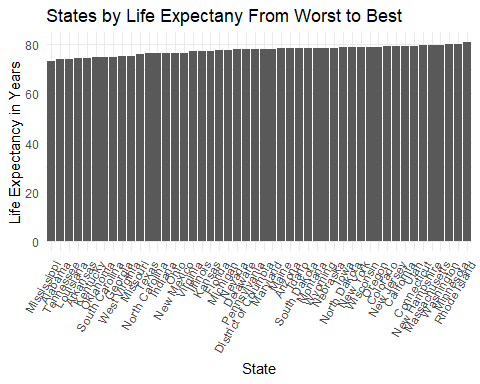
# qol dataset grouped by mean values  
qol\_state\_mean <- qol\_data %>%  
 group\_by(state) %>%  
 summarize(across(where(is.numeric), mean, na.rm = TRUE))

*Explore state distributions*

# reorder states by median pct\_poor\_to\_fair\_health  
qol\_state\_median <- qol\_state\_median %>%  
 mutate(state = fct\_reorder(state, pct\_poor\_to\_fair\_health, .desc = TRUE))  
  
# create the bar plot with pct\_poor\_to\_fair\_health vs state  
ggplot(data = qol\_state\_median,  
 aes(x = state, y = pct\_poor\_to\_fair\_health)) +  
 geom\_bar(stat = "identity") +  
 scale\_y\_continuous(labels = percent\_format(scale = 100)) +  
 labs(title = "States by Percent Poor to Fair Health From Worst to Best",  
 x = "State",   
 y = "Percent Reporting Poor to Fair Health") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 60, hjust = 1))

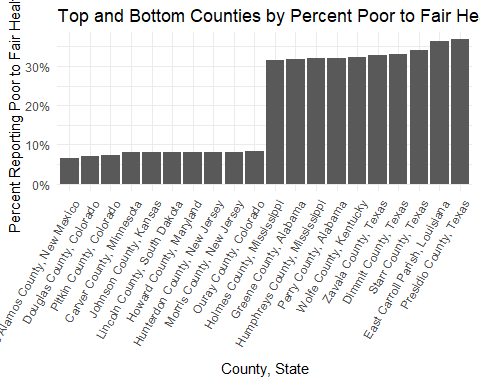


# reorder states by median life expectancy  
qol\_state\_median <- qol\_state\_median %>%  
 mutate(state = fct\_reorder(state, life\_expectancy\_years, .desc = FALSE))  
  
# create the bar plot with life expectancy vs state  
ggplot(data = qol\_state\_median,  
 aes(x = state, y = life\_expectancy\_years)) +  
 geom\_bar(stat = "identity") +  
 labs(title = "States by Life Expectany From Worst to Best",  
 x = "State",   
 y = "Life Expectancy in Years") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 60, hjust = 1))

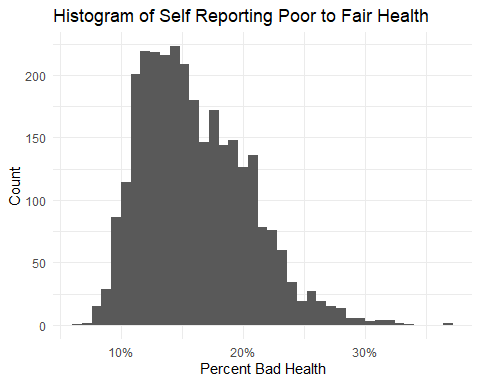


*County response distribution*

# Remove rows with missing values in pct\_poor\_to\_fair\_health  
#qol\_data\_clean <- qol\_data %>%  
# filter(!is.na(pct\_poor\_to\_fair\_health))  
  
# Arrange qol\_data by pct\_poor\_to\_fair\_health  
qol\_data\_sorted <- qol\_data %>%  
 arrange(pct\_poor\_to\_fair\_health)  
  
# Select top 10 and bottom 10 counties  
top\_bottom\_counties <- qol\_data\_sorted %>%  
 slice(c(1:10, (n() - 9):n())) # Select first 10 and last 10 rows  
  
# Plotting the bar chart with formatted labels  
ggplot(top\_bottom\_counties, aes(x = reorder(paste(county, state, sep = ", "), pct\_poor\_to\_fair\_health), y = pct\_poor\_to\_fair\_health)) +  
 geom\_bar(stat = "identity") +  
 scale\_y\_continuous(labels = scales::percent\_format(scale = 100)) +  
 labs(title = "Top and Bottom Counties by Percent Poor to Fair Health",  
 x = "County, State",  
 y = "Percent Reporting Poor to Fair Health") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 60, hjust = 1))



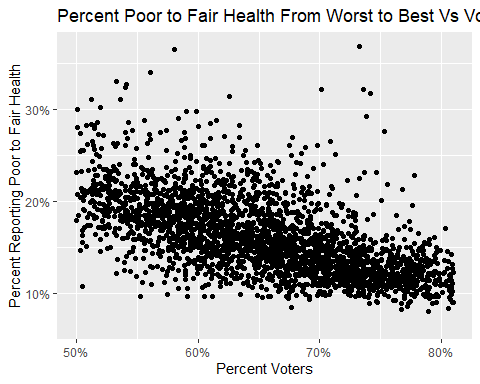
ggplot(data = qol\_data,  
 mapping = aes(x = pct\_poor\_to\_fair\_health)) +  
 geom\_histogram(binwidth = 0.008) +  
 scale\_x\_continuous(labels = percent\_format(scale = 100)) +  
 labs(x = "Percent Bad Health",  
 y = "Count",  
 title = "Histogram of Self Reporting Poor to Fair Health") +  
 theme\_minimal()



*Exploring points of interest*

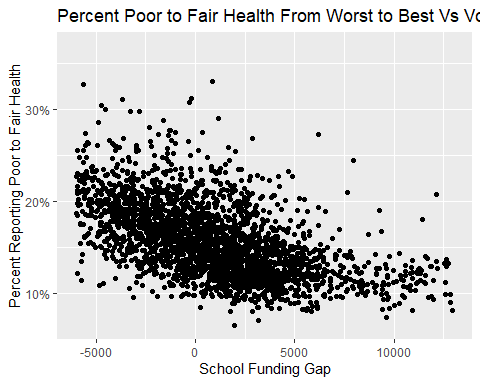
# Create the scatter plot with x-axis limit, percentage formatting, and state labels  
ggplot(data = qol\_data,  
 aes(x = pct\_voters, y = pct\_poor\_to\_fair\_health)) +  
 geom\_point() +  
 #geom\_text(aes(label = state), hjust = 1.2, vjust = 0.5, size = 2) + # Add state labels  
 scale\_y\_continuous(labels = percent\_format(scale = 100)) +  
 scale\_x\_continuous(labels = percent\_format(scale = 100), limits = c(0.5, 0.81)) +  
 labs(title = "Percent Poor to Fair Health From Worst to Best Vs Voters",  
 x = "Percent Voters",   
 y = "Percent Reporting Poor to Fair Health")

## Warning: Removed 300 rows containing missing values (`geom\_point()`).

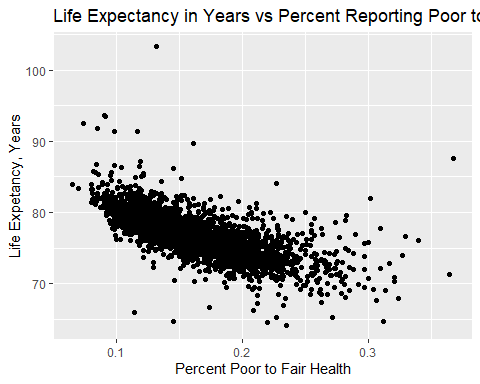


# Create the scatter plot with x-axis limit, percentage formatting, and state labels  
ggplot(data = qol\_data,  
 aes(x = school\_funding\_gap, y = pct\_poor\_to\_fair\_health)) +  
 geom\_point() +  
 # geom\_text(aes(label = state), hjust = 1.2, vjust = 0.5, size = 2) + # Add state labels  
 scale\_y\_continuous(labels = percent\_format(scale = 100)) +  
 scale\_x\_continuous(limits = c(-6000, 13000)) +  
 labs(title = "Percent Poor to Fair Health From Worst to Best Vs Voters",  
 x = "School Funding Gap",   
 y = "Percent Reporting Poor to Fair Health")

## Warning: Removed 273 rows containing missing values (`geom\_point()`).



# Create the scatter plot with x-axis limit, percentage formatting, and state labels  
ggplot(data = qol\_data,  
 aes(x = pct\_poor\_to\_fair\_health, y = life\_expectancy\_years)) +  
 geom\_point() +  
 # scale\_y\_continuous() +  
 # scale\_x\_continuous(labels = percent\_format(scale = 100), limits = c(0.42,0.80)) +  
 labs(title = "Life Expectancy in Years vs Percent Reporting Poor to Fair Health",  
 x = "Percent Poor to Fair Health",   
 y = "Life Expetancy, Years")

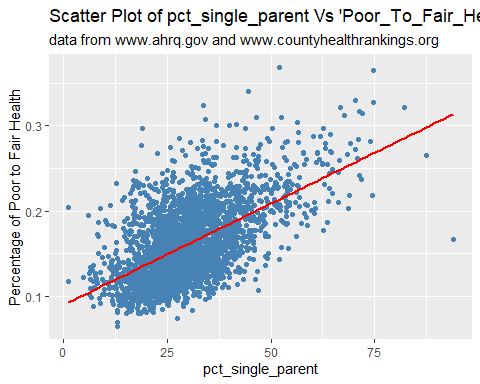


*Exploring community features*

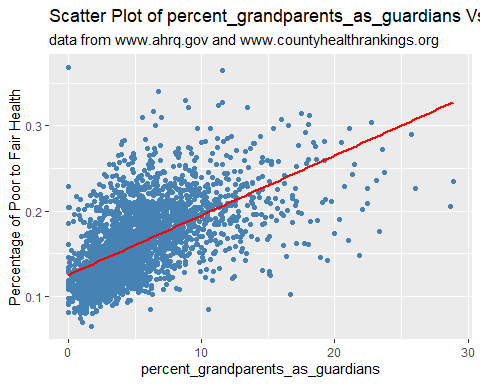
# Comparing important features from the "community aspect domain" of SDOH with the response variable "Percentage of adults reporting poor to fair health per county  
  
# Create a list of features  
features <- c("pct\_single\_parent", "percent\_grandparents\_as\_guardians",   
 "pct\_adult\_smokers", "pct\_obese\_adults", "pct\_binge\_drinkers",   
 "social\_clubs\_per\_10k", "pct\_overcrowded\_hh", "average\_hh\_size" )  
  
# Create a plot for each feature  
for (feature in features) {  
 print(  
 ggplot(qol\_data, aes\_string(x = feature, y = "pct\_poor\_to\_fair\_health")) +  
 geom\_point(color = "steelblue") +  
 geom\_smooth(method = "lm", se = FALSE, color = "red") +  
 labs(title = paste("Scatter Plot of", feature, "Vs 'Poor\_To\_Fair\_Health' Reported Per County"),  
 subtitle = "data from www.ahrq.gov and www.countyhealthrankings.org ",  
 x = feature,  
 y = "Percentage of Poor to Fair Health")  
 )  
}

## Warning: `aes\_string()` was deprecated in ggplot2 3.0.0.  
## ℹ Please use tidy evaluation idioms with `aes()`.  
## ℹ See also `vignette("ggplot2-in-packages")` for more information.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

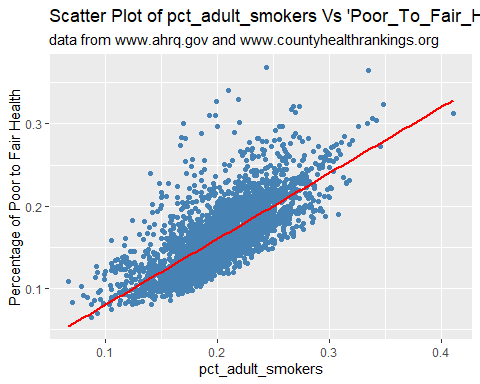
## `geom\_smooth()` using formula = 'y ~ x'



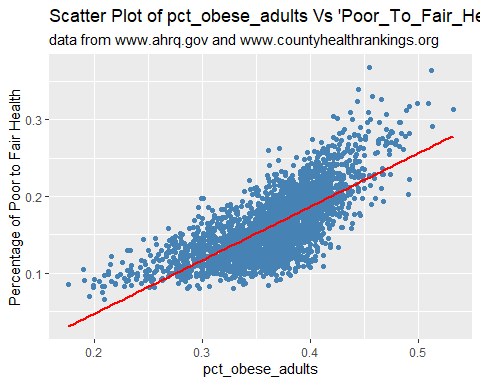
## `geom\_smooth()` using formula = 'y ~ x'



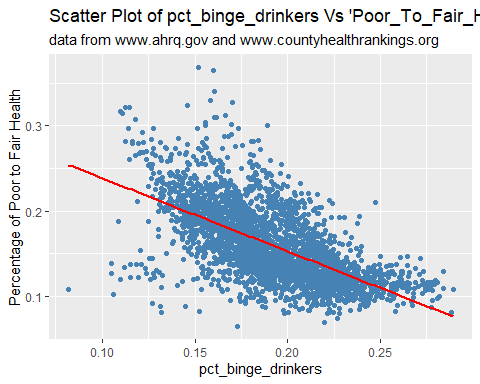
## `geom\_smooth()` using formula = 'y ~ x'



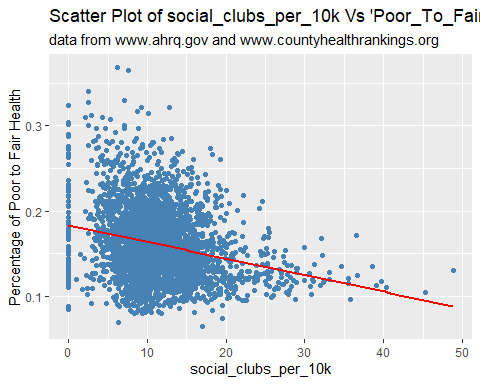
## `geom\_smooth()` using formula = 'y ~ x'



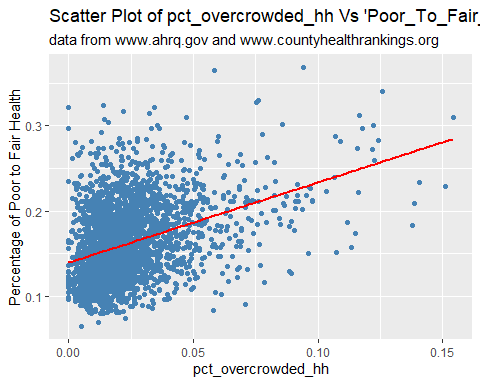
## `geom\_smooth()` using formula = 'y ~ x'



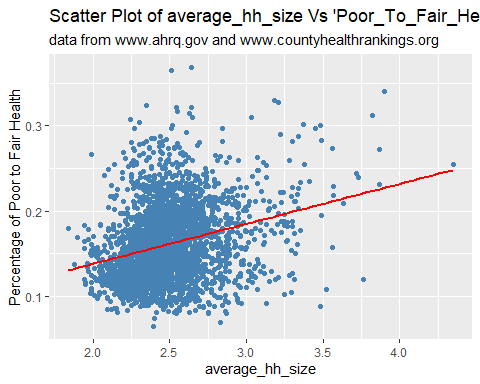
## `geom\_smooth()` using formula = 'y ~ x'



## `geom\_smooth()` using formula = 'y ~ x'



## `geom\_smooth()` using formula = 'y ~ x'



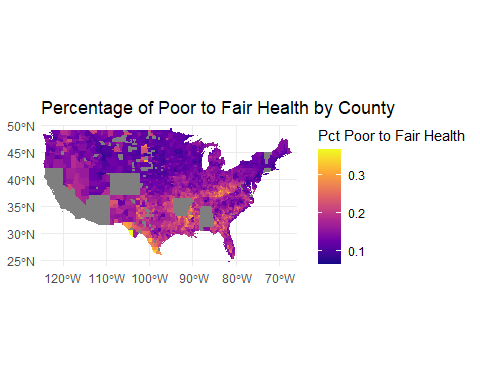
Exploring the relationship between median household income and health status in US counties along geographical lines by creating heat maps. UO

# Load US counties shapefile  
counties <- counties(cb = TRUE)

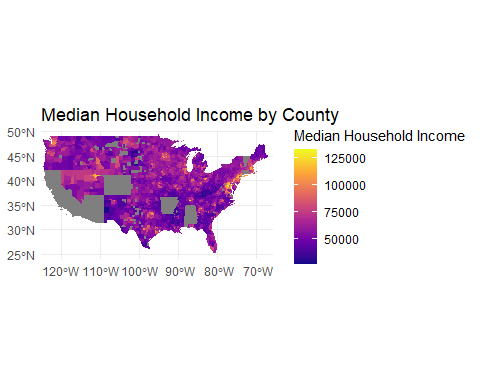
## Retrieving data for the year 2022

## | | | 0% | | | 1% | |= | 1% | |= | 2% | |== | 2% | |== | 3% | |=== | 4% | |==== | 5% | |==== | 6% | |===== | 7% | |====== | 8% | |====== | 9% | |======= | 10% | |======= | 11% | |======== | 11% | |======== | 12% | |========= | 12% | |========= | 13% | |========== | 14% | |========== | 15% | |=========== | 15% | |=========== | 16% | |============ | 17% | |============= | 18% | |============= | 19% | |============== | 19% | |============== | 20% | |============== | 21% | |=============== | 21% | |================ | 22% | |================= | 24% | |================= | 25% | |================== | 25% | |================== | 26% | |=================== | 27% | |=================== | 28% | |==================== | 29% | |===================== | 30% | |======================= | 33% | |======================= | 34% | |======================== | 34% | |======================== | 35% | |========================= | 36% | |========================== | 36% | |========================== | 37% | |========================== | 38% | |=========================== | 39% | |============================ | 40% | |============================ | 41% | |============================= | 41% | |============================= | 42% | |============================== | 43% | |=============================== | 44% | |=============================== | 45% | |================================= | 47% | |================================= | 48% | |================================== | 48% | |=================================== | 49% | |=================================== | 50% | |=================================== | 51% | |==================================== | 51% | |==================================== | 52% | |===================================== | 52% | |===================================== | 53% | |====================================== | 54% | |======================================= | 55% | |======================================= | 56% | |======================================== | 57% | |========================================== | 60% | |=========================================== | 61% | |=========================================== | 62% | |============================================ | 62% | |============================================ | 63% | |============================================= | 64% | |============================================== | 66% | |=============================================== | 67% | |=============================================== | 68% | |================================================ | 68% | |================================================ | 69% | |================================================= | 70% | |================================================== | 71% | |================================================== | 72% | |=================================================== | 72% | |=================================================== | 73% | |==================================================== | 74% | |===================================================== | 75% | |===================================================== | 76% | |====================================================== | 77% | |======================================================= | 78% | |======================================================= | 79% | |======================================================== | 79% | |======================================================== | 80% | |======================================================== | 81% | |========================================================= | 81% | |========================================================= | 82% | |========================================================== | 83% | |========================================================== | 84% | |=========================================================== | 85% | |============================================================ | 85% | |============================================================ | 86% | |============================================================= | 87% | |============================================================== | 89% | |=============================================================== | 90% | |================================================================ | 92% | |================================================================= | 92% | |================================================================= | 93% | |================================================================== | 94% | |================================================================== | 95% | |=================================================================== | 95% | |=================================================================== | 96% | |==================================================================== | 97% | |==================================================================== | 98% | |===================================================================== | 99% | |======================================================================| 100%

# Ensure fips\_code is a character  
qol\_data$fips\_code <- as.character(qol\_data$fips\_code)  
  
# Merge shapefile with qol\_data  
counties <- counties %>%  
 left\_join(qol\_data, by = c("GEOID" = "fips\_code"))  
  
# Set plot size  
options(repr.plot.width = 10, repr.plot.height = 8)  
  
# Create map for Percentage of Poor to Fair Health  
ggplot(data = counties) +  
 geom\_sf(aes(fill = pct\_poor\_to\_fair\_health), color = NA) +  
 scale\_fill\_viridis\_c(option = "plasma") +  
 labs(title = "Percentage of Poor to Fair Health by County",  
 fill = "Pct Poor to Fair Health") +  
 coord\_sf(xlim = c(-125, -66), ylim = c(24, 50), expand = FALSE) +  
 theme\_minimal()

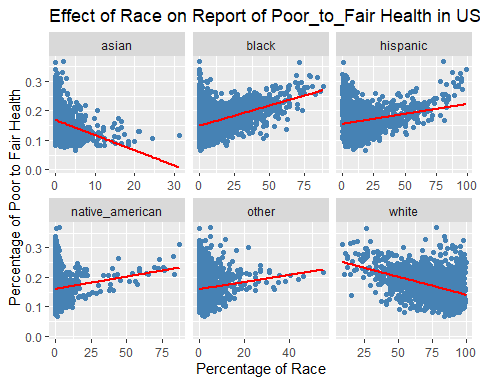


# Create map for Median Household income  
ggplot(data = counties) +  
 geom\_sf(aes(fill = median\_hh\_income.x), color = NA) +  
 scale\_fill\_viridis\_c(option = "plasma") +  
 labs(title = "Median Household Income by County",  
 fill = "Median Household Income") +  
 coord\_sf(xlim = c(-125, -66), ylim = c(24, 50), expand = FALSE) +  
 theme\_minimal()



# Create a list of race variables  
race\_vars <- c("pct\_native\_american", "pct\_asian", "pct\_black", "pct\_hispanic", "pct\_other\_race", "pct\_white")  
  
qol\_data\_long <- qol\_data %>%  
 pivot\_longer(  
 cols = all\_of(race\_vars),  
 names\_to = "race",  
 values\_to = "percentage"  
 ) %>%  
 mutate(race = case\_when(  
 race == "pct\_native\_american" ~ "native\_american",  
 race == "pct\_asian" ~ "asian",  
 race == "pct\_black" ~ "black",  
 race == "pct\_hispanic" ~ "hispanic",  
 race == "pct\_other\_race" ~ "other",  
 race == "pct\_white" ~ "white",  
 TRUE ~ race  
 ))  
  
  
# Create faceted plot  
ggplot(qol\_data\_long, aes(x = percentage, y = pct\_poor\_to\_fair\_health)) +  
 geom\_point(color = "steelblue") +  
 geom\_smooth(method = "lm", se = FALSE, color = "red") +  
 facet\_wrap(~race, scales = "free\_x") +  
 labs(title = "Effect of Race on Report of Poor\_to\_Fair Health in US Counties",  
 x = "Percentage of Race",  
 y = "Percentage of Poor to Fair Health")

## `geom\_smooth()` using formula = 'y ~ x'



*Exploring education, economic, and local environment features*

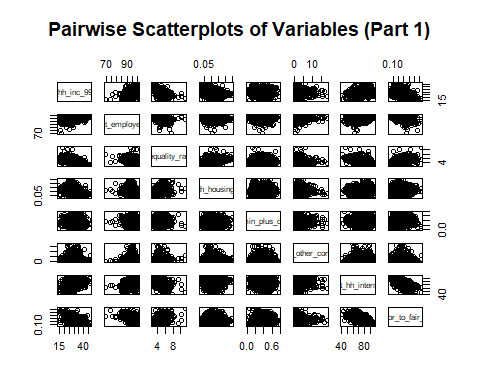
data\_mg <- qol\_data %>%  
 select("pct\_hh\_inc\_99999", "pct\_employed", "inequality\_ratio", "pct\_high\_housing\_costs", "pct\_30\_min\_plus\_commute", "pct\_hh\_other\_computer", "pct\_hh\_internet", "median\_hh\_income.x", "pct\_highschool\_diploma", "school\_funding\_gap", "population\_density", "days\_over\_90\_f", "air\_polution\_metric", "water\_quality", "pct\_poor\_to\_fair\_health")  
  
data\_mg <- na.omit(data\_mg)  
  
summary(data\_mg)

## pct\_hh\_inc\_99999 pct\_employed inequality\_ratio pct\_high\_housing\_costs  
## Min. :13.50 Min. : 69.61 Min. : 2.773 Min. :0.02103   
## 1st Qu.:28.61 1st Qu.: 93.67 1st Qu.: 4.002 1st Qu.:0.07982   
## Median :31.49 Median : 95.10 Median : 4.409 Median :0.09764   
## Mean :31.13 Mean : 94.79 Mean : 4.537 Mean :0.10150   
## 3rd Qu.:33.94 3rd Qu.: 96.32 3rd Qu.: 4.937 3rd Qu.:0.11909   
## Max. :47.18 Max. :100.00 Max. :11.128 Max. :0.24733   
## pct\_30\_min\_plus\_commute pct\_hh\_other\_computer pct\_hh\_internet  
## Min. :0.0000 Min. : 0.000 Min. :41.38   
## 1st Qu.:0.2360 1st Qu.: 1.295 1st Qu.:74.60   
## Median :0.3250 Median : 1.830 Median :80.13   
## Mean :0.3311 Mean : 2.110 Mean :78.98   
## 3rd Qu.:0.4185 3rd Qu.: 2.470 3rd Qu.:84.43   
## Max. :0.7840 Max. :17.850 Max. :96.81   
## median\_hh\_income.x pct\_highschool\_diploma school\_funding\_gap  
## Min. : 25997 Min. :0.4967 Min. :-18852.8   
## 1st Qu.: 47731 1st Qu.:0.8462 1st Qu.: -2420.8   
## Median : 55010 Median :0.8914 Median : 441.6   
## Mean : 57130 Mean :0.8785 Mean : 247.0   
## 3rd Qu.: 63763 3rd Qu.:0.9215 3rd Qu.: 2866.0   
## Max. :132509 Max. :0.9862 Max. : 27719.2   
## population\_density days\_over\_90\_f air\_polution\_metric water\_quality   
## Min. : 0.44 Min. : 0.00 Min. : 0.900 Min. :0.0000   
## 1st Qu.: 18.75 1st Qu.: 25.00 1st Qu.: 6.600 1st Qu.:0.0000   
## Median : 45.70 Median : 55.00 Median : 7.800 Median :0.0000   
## Mean : 213.44 Mean : 55.98 Mean : 7.615 Mean :0.3425   
## 3rd Qu.: 115.72 3rd Qu.: 87.00 3rd Qu.: 8.900 3rd Qu.:1.0000   
## Max. :71895.54 Max. :144.00 Max. :15.600 Max. :1.0000   
## pct\_poor\_to\_fair\_health  
## Min. :0.0650   
## 1st Qu.:0.1270   
## Median :0.1540   
## Mean :0.1613   
## 3rd Qu.:0.1900   
## Max. :0.3680

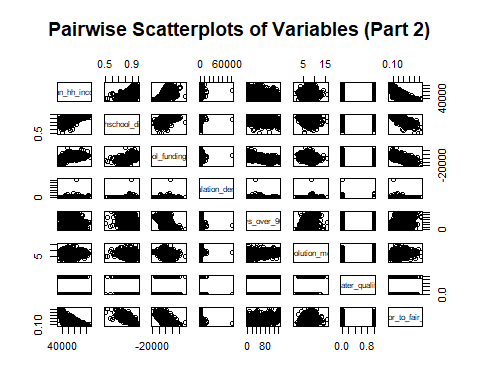
variables <- c(  
 "pct\_hh\_inc\_99999",  
 "pct\_employed",  
 "inequality\_ratio",  
 "pct\_high\_housing\_costs",  
 "pct\_30\_min\_plus\_commute",  
 "pct\_hh\_other\_computer",  
 "pct\_hh\_internet",  
 "median\_hh\_income.x",  
 "pct\_highschool\_diploma",  
 "school\_funding\_gap",  
 "population\_density",  
 "days\_over\_90\_f",  
 "air\_polution\_metric",  
 "water\_quality",  
 "pct\_poor\_to\_fair\_health"  
)  
  
mg\_sum\_stats <- function(data, vars) {  
 summary\_stats <- data %>%  
 select(all\_of(vars)) %>%  
 summarise(across(everything(), list(  
 mean = ~mean(. , na.rm = TRUE),  
 median = ~median(. , na.rm = TRUE),  
 sd = ~sd(. , na.rm = TRUE),  
 min = ~min(. , na.rm = TRUE),  
 max = ~max(. , na.rm = TRUE)  
 )))  
 return(summary\_stats)  
}  
  
summary\_stats <- mg\_sum\_stats(data\_mg, variables)  
  
summary\_matrix <- matrix(ncol = length(variables), nrow = 5)  
colnames(summary\_matrix) <- variables  
rownames(summary\_matrix) <- c("Mean", "Median", "sd", "Min", "Max")  
  
for (i in 1:length(variables)) {  
 summary\_matrix[1, i] <- summary\_stats[[paste0(variables[i], "\_mean")]]  
 summary\_matrix[2, i] <- summary\_stats[[paste0(variables[i], "\_median")]]  
 summary\_matrix[3, i] <- summary\_stats[[paste0(variables[i], "\_sd")]]  
 summary\_matrix[4, i] <- summary\_stats[[paste0(variables[i], "\_min")]]  
 summary\_matrix[5, i] <- summary\_stats[[paste0(variables[i], "\_max")]]  
}  
  
mg\_summary\_matrix <- round(summary\_matrix, 3)  
print(mg\_summary\_matrix)

## pct\_hh\_inc\_99999 pct\_employed inequality\_ratio pct\_high\_housing\_costs  
## Mean 31.129 94.787 4.537 0.101  
## Median 31.490 95.100 4.409 0.098  
## sd 4.245 2.453 0.787 0.032  
## Min 13.500 69.610 2.773 0.021  
## Max 47.180 100.000 11.128 0.247  
## pct\_30\_min\_plus\_commute pct\_hh\_other\_computer pct\_hh\_internet  
## Mean 0.331 2.11 78.976  
## Median 0.325 1.83 80.130  
## sd 0.125 1.57 8.033  
## Min 0.000 0.00 41.380  
## Max 0.784 17.85 96.810  
## median\_hh\_income.x pct\_highschool\_diploma school\_funding\_gap  
## Mean 57130.43 0.878 247.028  
## Median 55010.00 0.891 441.583  
## sd 13978.53 0.058 4726.185  
## Min 25997.00 0.497 -18852.820  
## Max 132509.00 0.986 27719.240  
## population\_density days\_over\_90\_f air\_polution\_metric water\_quality  
## Mean 213.439 55.982 7.615 0.342  
## Median 45.700 55.000 7.800 0.000  
## sd 1480.657 38.088 1.672 0.475  
## Min 0.440 0.000 0.900 0.000  
## Max 71895.540 144.000 15.600 1.000  
## pct\_poor\_to\_fair\_health  
## Mean 0.161  
## Median 0.154  
## sd 0.044  
## Min 0.065  
## Max 0.368

data\_mg\_part1 <- data\_mg %>%  
 select("pct\_hh\_inc\_99999", "pct\_employed", "inequality\_ratio", "pct\_high\_housing\_costs", "pct\_30\_min\_plus\_commute", "pct\_hh\_other\_computer", "pct\_hh\_internet", "pct\_poor\_to\_fair\_health")  
  
data\_mg\_part2 <- data\_mg %>%  
 select("median\_hh\_income.x", "pct\_highschool\_diploma", "school\_funding\_gap", "population\_density", "days\_over\_90\_f", "air\_polution\_metric", "water\_quality", "pct\_poor\_to\_fair\_health")  
  
pairs(data\_mg\_part1, main = "Pairwise Scatterplots of Variables (Part 1)")

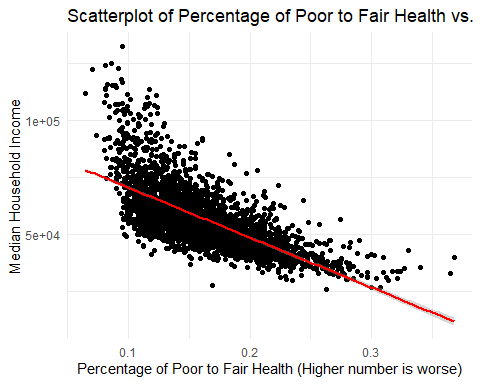


pairs(data\_mg\_part2, main = "Pairwise Scatterplots of Variables (Part 2)")



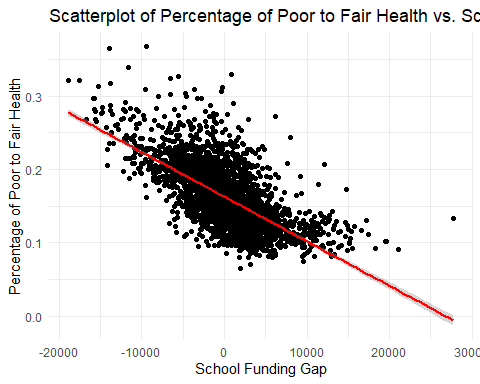
ggplot(data\_mg, aes(x = pct\_poor\_to\_fair\_health, y = median\_hh\_income.x)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", col = "red") +  
 labs(title = "Scatterplot of Percentage of Poor to Fair Health vs. Median Household Income",  
 x = "Percentage of Poor to Fair Health (Higher number is worse)",  
 y = "Median Household Income") +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'



ggplot(data\_mg, aes(x = school\_funding\_gap, y = pct\_poor\_to\_fair\_health)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", col = "red") +  
 labs(title = "Scatterplot of Percentage of Poor to Fair Health vs. School Funding Gap",  
 x = "School Funding Gap",  
 y = "Percentage of Poor to Fair Health") +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'



*Table creation with gtsummary*

# create longer table with race as a column  
qol\_data\_long <- qol\_data %>%  
 pivot\_longer(  
 cols = all\_of(race\_vars),  
 names\_to = "race",  
 values\_to = "percentage"  
 ) %>%  
 mutate(race = case\_when(  
 race == "pct\_native\_american" ~ "native\_american",  
 race == "pct\_asian" ~ "asian",  
 race == "pct\_black" ~ "black",  
 race == "pct\_hispanic" ~ "hispanic",  
 race == "pct\_other\_race" ~ "other",  
 race == "pct\_white" ~ "white",  
 TRUE ~ race  
 ))  
  
# new data frames by filtering for individual races  
qol\_data\_na <- qol\_data\_long %>%  
 filter(race == "native\_american")  
  
qol\_data\_asian <- qol\_data\_long %>%  
 filter(race == "asian")  
  
qol\_data\_black <- qol\_data\_long %>%  
 filter(race == "black")  
  
qol\_data\_hispanic <- qol\_data\_long %>%  
 filter(race == "hispanic")  
  
qol\_data\_white <- qol\_data\_long %>%  
 filter(race == "white")  
  
qol\_data\_other <- qol\_data\_long %>%  
 filter(race == "other")

df\_race\_list <- c("asian", "black", "hispanic", "native\_american", "other", "white")  
  
feature\_list <- c("pct\_poor\_to\_fair\_health", "life\_expectancy\_years", "pct\_voters")  
  
race\_stats\_calc <- function(data, race\_list, feature\_list) {  
  
 results <- list()  
   
 for (race in race\_list) {  
 filtered\_data <- data %>%  
 filter(race == !!race)  
   
 stats <- filtered\_data %>%  
 summarise(across(all\_of(feature\_list), list(  
 Mean = ~mean(.),  
 Median = ~median(.),  
 SD = ~sd(.),  
 Kurtosis = ~kurtosis(.),  
 Skewness = ~skewness(.)  
 ), .names = "{fn}\_{col}"))  
   
 results[[race]] <- stats  
 }  
   
 combined\_results <- bind\_rows(results, .id = "race")  
   
 return(combined\_results)  
}

st <- race\_stats\_calc(data = qol\_data\_long,  
 race = df\_race\_list,  
 feature = feature\_list)  
  
st

## # A tibble: 6 × 16  
## race Mean\_pct\_poor\_to\_fai…¹ Median\_pct\_poor\_to\_f…² SD\_pct\_poor\_to\_fair\_…³  
## <chr> <dbl> <dbl> <dbl>  
## 1 asian 0.161 0.154 0.0440  
## 2 black 0.161 0.154 0.0440  
## 3 hispanic 0.161 0.154 0.0440  
## 4 native\_a… 0.161 0.154 0.0440  
## 5 other 0.161 0.154 0.0440  
## 6 white 0.161 0.154 0.0440  
## # ℹ abbreviated names: ¹​Mean\_pct\_poor\_to\_fair\_health,  
## # ²​Median\_pct\_poor\_to\_fair\_health, ³​SD\_pct\_poor\_to\_fair\_health  
## # ℹ 12 more variables: Kurtosis\_pct\_poor\_to\_fair\_health <dbl>,  
## # Skewness\_pct\_poor\_to\_fair\_health <dbl>, Mean\_life\_expectancy\_years <dbl>,  
## # Median\_life\_expectancy\_years <dbl>, SD\_life\_expectancy\_years <dbl>,  
## # Kurtosis\_life\_expectancy\_years <dbl>, Skewness\_life\_expectancy\_years <dbl>,  
## # Mean\_pct\_voters <dbl>, Median\_pct\_voters <dbl>, SD\_pct\_voters <dbl>, …

# create function  
race\_stats\_calc <- function(data, race\_list, feature\_list) {  
 summary\_stats <- data %>%  
 filter(race %in% race\_list) %>%  
 group\_by(race) %>%  
 summarise(across(all\_of(feature\_list), list(  
 Mean = ~mean(.),  
 Median = ~median(.),  
 SD = ~sd(.),  
 Kurtosis = ~kurtosis(.),  
 Skewness = ~skewness(.)  
 ), .names = "{fn}\_{col}"))  
   
 return(summary\_stats)  
}

st <- race\_stats\_calc(data = qol\_data\_long,  
 race = df\_race\_list,  
 feature = feature\_list)  
  
st

## # A tibble: 6 × 16  
## race Mean\_pct\_poor\_to\_fai…¹ Median\_pct\_poor\_to\_f…² SD\_pct\_poor\_to\_fair\_…³  
## <chr> <dbl> <dbl> <dbl>  
## 1 asian 0.161 0.154 0.0440  
## 2 black 0.161 0.154 0.0440  
## 3 hispanic 0.161 0.154 0.0440  
## 4 native\_a… 0.161 0.154 0.0440  
## 5 other 0.161 0.154 0.0440  
## 6 white 0.161 0.154 0.0440  
## # ℹ abbreviated names: ¹​Mean\_pct\_poor\_to\_fair\_health,  
## # ²​Median\_pct\_poor\_to\_fair\_health, ³​SD\_pct\_poor\_to\_fair\_health  
## # ℹ 12 more variables: Kurtosis\_pct\_poor\_to\_fair\_health <dbl>,  
## # Skewness\_pct\_poor\_to\_fair\_health <dbl>, Mean\_life\_expectancy\_years <dbl>,  
## # Median\_life\_expectancy\_years <dbl>, SD\_life\_expectancy\_years <dbl>,  
## # Kurtosis\_life\_expectancy\_years <dbl>, Skewness\_life\_expectancy\_years <dbl>,  
## # Mean\_pct\_voters <dbl>, Median\_pct\_voters <dbl>, SD\_pct\_voters <dbl>, …

qol\_data\_na <- qol\_data\_na %>%   
 mutate(pct\_poor\_to\_fair\_health = pct\_poor\_to\_fair\_health \* percentage)  
  
  
native\_american <- summary(qol\_data\_na$pct\_voters)  
asian <- summary(qol\_data\_asian$pct\_voters)  
black <- summary(qol\_data\_black$pct\_voters)  
hispanic <- summary(qol\_data\_hispanic$pct\_voters)  
white <- summary(qol\_data\_white$pct\_voters)  
other <- summary(qol\_data\_other$pct\_voters)  
  
  
  
print(native\_american)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.1942 0.5883 0.6554 0.6530 0.7215 1.0000

print(black)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.1942 0.5883 0.6554 0.6530 0.7215 1.0000

print(hispanic)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.1942 0.5883 0.6554 0.6530 0.7215 1.0000

print(asian)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.1942 0.5883 0.6554 0.6530 0.7215 1.0000

print(white)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.1942 0.5883 0.6554 0.6530 0.7215 1.0000

print(other)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.1942 0.5883 0.6554 0.6530 0.7215 1.0000

table\_1 <- tbl\_summary(qol\_data,   
 include = c(pct\_hh\_inc\_99999,  
 pct\_employed,  
 inequality\_ratio,  
 pct\_high\_housing\_costs,  
 pct\_30\_min\_plus\_commute,  
 pct\_hh\_other\_computer,  
 pct\_hh\_internet,  
 median\_hh\_income.x,  
 pct\_highschool\_diploma,  
 school\_funding\_gap,  
 population\_density,  
 days\_over\_90\_f,  
 air\_polution\_metric,  
 water\_quality,  
 pct\_poor\_to\_fair\_health))  
  
table\_1

| **Characteristic** | **N = 2,955***1* |
| --- | --- |
| pct\_hh\_inc\_99999 | 31.5 (28.6, 33.9) |
| pct\_employed | 95.10 (93.67, 96.32) |
| inequality\_ratio | 4.41 (4.00, 4.94) |
| pct\_high\_housing\_costs | 0.10 (0.08, 0.12) |
| pct\_30\_min\_plus\_commute | 0.33 (0.24, 0.42) |
| pct\_hh\_other\_computer | 1.83 (1.29, 2.47) |
| pct\_hh\_internet | 80 (75, 84) |
| median\_hh\_income.x | 55,010 (47,722, 63,785) |
| pct\_highschool\_diploma | 0.89 (0.85, 0.92) |
| school\_funding\_gap | 442 (-2,426, 2,870) |
| population\_density | 46 (19, 116) |
| days\_over\_90\_f | 55 (25, 87) |
| air\_polution\_metric | 7.80 (6.60, 8.90) |
| water\_quality | 1,012 (34%) |
| pct\_poor\_to\_fair\_health | 0.15 (0.13, 0.19) |
| *1*Median (Q1, Q3); n (%) | |

# Ensure the trial dataset is loaded and available  
# trial <- your\_data\_loading\_function()  
  
# Create the summary table  
test\_table <- tbl\_summary(  
 data = qol\_data\_long,  
 include = c(pct\_poor\_to\_fair\_health, life\_expectancy\_years, pct\_voters),  
 by = race,  
 statistic = list(  
 all\_continuous() ~ "{median} ({sd})",  
 all\_dichotomous() ~ "{p}%"  
 ),  
 missing = "no"  
)  
  
# Print the summary table  
test\_table

| **Characteristic** | **asian** N = 2,955*1* | **black** N = 2,955*1* | **hispanic** N = 2,955*1* | **native\_american** N = 2,955*1* | **other** N = 2,955*1* | **white** N = 2,955*1* |
| --- | --- | --- | --- | --- | --- | --- |
| pct\_poor\_to\_fair\_health | 0.15 (0.04) | 0.15 (0.04) | 0.15 (0.04) | 0.15 (0.04) | 0.15 (0.04) | 0.15 (0.04) |
| life\_expectancy\_years | 76.89 (3.08) | 76.89 (3.08) | 76.89 (3.08) | 76.89 (3.08) | 76.89 (3.08) | 76.89 (3.08) |
| pct\_voters | 0.66 (0.10) | 0.66 (0.10) | 0.66 (0.10) | 0.66 (0.10) | 0.66 (0.10) | 0.66 (0.10) |
| *1*Median (SD) | | | | | | |

unique(qol\_data\_long$race)

## [1] "native\_american" "asian" "black" "hispanic"   
## [5] "other" "white"

# Assuming qol\_data\_long is now the correctly reshaped dataset  
test\_table <- tbl\_summary(  
 data = qol\_data\_long,  
 include = c(pct\_poor\_to\_fair\_health, life\_expectancy\_years, pct\_voters),  
 by = race,  
 statistic = list(  
 all\_continuous() ~ "{mean} ({sd}, {kurtosis}, {skewness})",  
 all\_dichotomous() ~ "{p}%"  
 ),  
 missing = "no"  
)  
  
# Print the summary table  
test\_table

| **Characteristic** | **asian** N = 2,955*1* | **black** N = 2,955*1* | **hispanic** N = 2,955*1* | **native\_american** N = 2,955*1* | **other** N = 2,955*1* | **white** N = 2,955*1* |
| --- | --- | --- | --- | --- | --- | --- |
| pct\_poor\_to\_fair\_health | 0.16 (0.04, 3.49, 0.73) | 0.16 (0.04, 3.49, 0.73) | 0.16 (0.04, 3.49, 0.73) | 0.16 (0.04, 3.49, 0.73) | 0.16 (0.04, 3.49, 0.73) | 0.16 (0.04, 3.49, 0.73) |
| life\_expectancy\_years | 76.84 (3.08, 6.40, 0.30) | 76.84 (3.08, 6.40, 0.30) | 76.84 (3.08, 6.40, 0.30) | 76.84 (3.08, 6.40, 0.30) | 76.84 (3.08, 6.40, 0.30) | 76.84 (3.08, 6.40, 0.30) |
| pct\_voters | 0.65 (0.10, 3.13, -0.08) | 0.65 (0.10, 3.13, -0.08) | 0.65 (0.10, 3.13, -0.08) | 0.65 (0.10, 3.13, -0.08) | 0.65 (0.10, 3.13, -0.08) | 0.65 (0.10, 3.13, -0.08) |
| *1*Mean (SD, kurtosis, skewness) | | | | | | |

# Assuming qol\_data\_long is your reshaped dataset  
  
# Filter the dataset for only observations where race == "white"  
  
  
test\_table2 <- tbl\_summary(  
 data = qol\_data\_white,  
 include = c(pct\_poor\_to\_fair\_health, life\_expectancy\_years, pct\_voters),  
 by = race,  
 statistic = list(  
 all\_continuous() ~ "{mean} ({sd})",  
 all\_dichotomous() ~ "{p}%"  
 ),  
 missing = "no"  
)  
  
# Print the summary table  
test\_table2

| **Characteristic** | **white** N = 2,955*1* |
| --- | --- |
| pct\_poor\_to\_fair\_health | 0.16 (0.04) |
| life\_expectancy\_years | 76.84 (3.08) |
| pct\_voters | 0.65 (0.10) |
| *1*Mean (SD) | |

test\_table2 <- tbl\_summary(  
 data = qol\_data\_black,  
 include = c(pct\_poor\_to\_fair\_health, life\_expectancy\_years, pct\_voters),  
 by = race,  
 statistic = list(  
 all\_continuous() ~ "{mean} ({sd})",  
 all\_dichotomous() ~ "{p}%"  
 ),  
 missing = "no"  
)  
  
# Print the summary table  
test\_table2

| **Characteristic** | **black** N = 2,955*1* |
| --- | --- |
| pct\_poor\_to\_fair\_health | 0.16 (0.04) |
| life\_expectancy\_years | 76.84 (3.08) |
| pct\_voters | 0.65 (0.10) |
| *1*Mean (SD) | |

# Assuming qol\_data\_long is now the correctly reshaped dataset  
test\_table <- tbl\_summary(  
 data = qol\_data\_long,  
 include = c(pct\_poor\_to\_fair\_health, life\_expectancy\_years, pct\_voters),  
 by = region,  
 statistic = list(  
 all\_continuous() ~ "{mean} ({sd}, {kurtosis}, {skewness})",  
 all\_dichotomous() ~ "{p}%"  
 ),  
 missing = "no"  
)  
  
# Print the summary table  
test\_table

| **Characteristic** | **Midwest** N = 6,108*1* | **Northeast** N = 1,194*1* | **South** N = 8,124*1* | **West** N = 2,304*1* |
| --- | --- | --- | --- | --- |
| pct\_poor\_to\_fair\_health | 0.14 (0.03, 4.90, 0.96) | 0.12 (0.02, 2.66, 0.15) | 0.19 (0.04, 3.68, 0.41) | 0.14 (0.03, 3.62, 0.60) |
| life\_expectancy\_years | 77.67 (2.60, 7.66, -0.55) | 78.75 (1.75, 3.03, 0.19) | 75.37 (2.67, 3.78, 0.30) | 78.88 (3.58, 10.91, 1.09) |
| pct\_voters | 0.68 (0.08, 3.66, -0.23) | 0.68 (0.08, 3.66, 0.25) | 0.61 (0.09, 3.24, 0.09) | 0.71 (0.10, 3.01, -0.25) |
| *1*Mean (SD, kurtosis, skewness) | | | | |