Team Beta Capstone Project

Mirande Gemme, Udumaga Onyeukwu, Steve Uzupis

2024-08-25

*Package Installation*

# install packages not on machine and needed for project  
installed\_packages <- installed.packages()  
  
list\_of\_packages <- c(  
 "readr",  
 "readxl",  
 "dplyr",  
 "lubridate",  
 "ggplot2",  
 "tidyr",  
 "corrplot",  
 "leaps",  
 "caret",  
 "car",  
 "scales",  
 "forcats",  
 "codebookr",  
 "gtsummary",  
 "tigris",  
 "cardx",  
 "moments",  
 "VIM",  
 "pROC",  
 "randomForest",  
 "glmnet",  
 "tidymodels",  
 "recipes",  
 "lmtest",  
 "xgboost",  
 "openxlsx",  
 "class",  
 "e1071",  
 "vip",  
 "pdp",  
 "gbm",  
 "gtsummary",  
 "gt"  
 )  
  
new\_packages <- list\_of\_packages[!(  
 list\_of\_packages %in% installed.packages()[,"Package"]  
 )]  
  
if(length(new\_packages)) install.packages(new\_packages)

rm(  
 installed\_packages,  
 list\_of\_packages,  
 new\_packages  
)

*Packages Load*

# load packages  
  
library(readr) # load csv files  
library(readxl) # load excel files  
library(dplyr) # data manipulation  
#library(lubridate) # date & time manipulation  
library(ggplot2) # data visualization  
library(tidyr) # collection of statistical packages  
#library(corrplot) # to visualize correlations  
#library(leaps) # for subset selection  
library(caret) # test for correlation  
#library(car) # for VIF  
#library(scales) # for visualizing plots in %  
#library(forcats) # ordering data frames  
library(codebookr) # adding appendix to r code  
library(gtsummary) # creating tables  
#library(tigris) # access to US geographic data  
#library(cardx) # to include statistic results  
#library(moments) # to calculate skewness and kurtosis  
#library(VIM) # to run K- Nearest Neighbor  
#library(pROC) # to analyse and display receiver operating characteristics(ROC) curves  
#library(randomForest) # to impalement Random Forest algorithm  
#library(glmnet) # for Regularized regression models.  
library(tidymodels) # for Machine learning workflows.  
#library(recipes) # for data preprocessing  
#library(lmtest) # to implement Hypothesis testing for linear models.  
#library(xgboost) # to build Gradient boosted models  
#library(openxlsx) # for Excel file manipulation   
#library(class) # for K-nearest neighbors classification  
library(e1071) # Support Vector Machines and other machine learning algorithms.  
#library(vip) # Variable importance plots for machine learning models.  
library(pdp) # partial dependence plots   
library(gbm) # evaluate GBM for feature   
library(gtsummary)  
#library(gt)

*Data load*

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

## [1] 3229 682

#cdc\_data <- read\_csv("data/cdc\_data.csv") found better response data set  
  
all\_chr\_data <- read\_csv("data/chr\_data.csv",  
 skip = 1)  
dim(all\_chr\_data)

## [1] 3194 720

*Initial data cleaning*

# remove unused features - not using  
  
#cdc\_data <- cdc\_data %>%   
# filter(Response == '65+') %>%   
# select('LocationAbbr', 'LocationDesc', 'Data\_Value', 'Number', 'WeightedNumber', 'StratificationCategory1', 'Stratification1', 'LocationID')

# remove unwanted features, create calculated feature  
  
sdoh\_data <- all\_sdoh\_data %>%   
 select(#"YEAR",  
 "COUNTYFIPS", #"STATEFIPS",  
 "STATE", "COUNTY", "REGION", #"TERRITORY",  
 "ACS\_TOT\_POP\_WT", #"ACS\_TOT\_POP\_US\_ABOVE1", "ACS\_TOT\_POP\_ABOVE5", "ACS\_TOT\_POP\_ABOVE15", "ACS\_TOT\_POP\_ABOVE16", "ACS\_TOT\_POP\_16\_19", "ACS\_TOT\_POP\_ABOVE25", "ACS\_TOT\_CIVIL\_POP\_ABOVE18", "ACS\_TOT\_CIVIL\_VET\_POP\_ABOVE25", "ACS\_TOT\_OWN\_CHILD\_BELOW17", "ACS\_TOT\_WORKER\_NWFH", "ACS\_TOT\_WORKER\_HH", "ACS\_TOT\_CIVILIAN\_LABOR", "ACS\_TOT\_CIVIL\_EMPLOY\_POP", "ACS\_TOT\_POP\_POV", "ACS\_TOT\_CIVIL\_NONINST\_POP\_POV", "ACS\_TOT\_CIVIL\_POP\_POV", "ACS\_TOT\_GRANDCHILDREN\_GP", "ACS\_TOT\_HU", "ACS\_TOT\_HH",  
 "ACS\_AVG\_HH\_SIZE",# "ACS\_TOT\_CIVIL\_NONINST\_POP", "ACS\_TOT\_CIVIL\_VET\_POP", "ACS\_PCT\_CHILD\_DISAB", "ACS\_PCT\_DISABLE", "ACS\_PCT\_NONVET\_DISABLE\_18\_64", "ACS\_PCT\_VET\_DISABLE\_18\_64",   
 "ACS\_PCT\_MALE", "ACS\_PCT\_FEMALE", #"ACS\_PCT\_CTZ\_US\_BORN", "ACS\_PCT\_CTZ\_NONUS\_BORN", "ACS\_PCT\_FOREIGN\_BORN",  
 "ACS\_PCT\_NON\_CITIZEN", "ACS\_PCT\_CTZ\_NATURALIZED", "ACS\_PCT\_CTZ\_ABOVE18", #"ACS\_PCT\_NONCTN\_1990", "ACS\_PCT\_NONCTN\_1999", "ACS\_PCT\_NONCTN\_2000", "ACS\_PCT\_NONCTN\_2010", "ACS\_PCT\_API\_LANG",   
 "ACS\_PCT\_ENGL\_NOT\_ALL", #"ACS\_PCT\_ENGL\_NOT\_WELL", "ACS\_PCT\_ENGL\_VERY\_WELL", "ACS\_PCT\_ENGL\_WELL", "ACS\_PCT\_ENGLISH", "ACS\_PCT\_HH\_LIMIT\_ENGLISH", "ACS\_PCT\_OTH\_EURP", "ACS\_PCT\_OTH\_LANG", "ACS\_PCT\_SPANISH", "ACS\_PCT\_VET", "ACS\_PCT\_GULFWAR\_1990\_2001", "ACS\_PCT\_GULFWAR\_2001", "ACS\_PCT\_GULFWAR\_VIETNAM", "ACS\_PCT\_VIETNAM", "ACS\_MEDIAN\_AGE", "ACS\_MEDIAN\_AGE\_MALE", "ACS\_MEDIAN\_AGE\_FEMALE", "ACS\_PCT\_AGE\_0\_4", "ACS\_PCT\_AGE\_5\_9", "ACS\_PCT\_AGE\_10\_14", "ACS\_PCT\_AGE\_15\_17", "ACS\_PCT\_AGE\_0\_17", "ACS\_PCT\_AGE\_18\_29", "ACS\_PCT\_AGE\_18\_44", "ACS\_PCT\_AGE\_30\_44", "ACS\_PCT\_AGE\_45\_64", "ACS\_PCT\_AGE\_50\_64", "ACS\_PCT\_AGE\_ABOVE65", "ACS\_PCT\_AGE\_ABOVE80",   
 "ACS\_PCT\_AIAN", #"ACS\_PCT\_AIAN\_FEMALE", "ACS\_PCT\_AIAN\_NONHISP",   
 "ACS\_PCT\_ASIAN", #"ACS\_PCT\_ASIAN\_FEMALE", "ACS\_PCT\_ASIAN\_MALE", "ACS\_PCT\_ASIAN\_NONHISP",   
 "ACS\_PCT\_BLACK", #"ACS\_PCT\_BLACK\_FEMALE", "ACS\_PCT\_BLACK\_MALE", "ACS\_PCT\_BLACK\_NONHISP", #"ACS\_PCT\_HISP\_FEMALE", "ACS\_PCT\_HISP\_MALE",  
 "ACS\_PCT\_HISPANIC", #"ACS\_PCT\_MULT\_RACE", "ACS\_PCT\_MULT\_RACE\_FEMALE", "ACS\_PCT\_MULT\_RACE\_MALE", "ACS\_PCT\_MULT\_RACE\_NONHISP", "ACS\_PCT\_NHPI", "ACS\_PCT\_NHPI\_FEMALE", "ACS\_PCT\_NHPI\_MALE", "ACS\_PCT\_NHPI\_NONHISP", "ACS\_PCT\_OTHER\_FEMALE", "ACS\_PCT\_OTHER\_MALE", "ACS\_PCT\_OTHER\_NONHISP",   
 "ACS\_PCT\_OTHER\_RACE", "ACS\_PCT\_WHITE", #"ACS\_PCT\_WHITE\_FEMALE", "ACS\_PCT\_WHITE\_MALE", "ACS\_PCT\_WHITE\_NONHISP", "ACS\_PCT\_HOUSEHOLDER\_WHITE", "ACS\_PCT\_HOUSEHOLDER\_BLACK", "ACS\_PCT\_HOUSEHOLDER\_AIAN", "ACS\_PCT\_HOUSEHOLDER\_ASIAN", "ACS\_PCT\_HOUSEHOLDER\_NHPI", "ACS\_PCT\_HOUSEHOLDER\_OTHER", "ACS\_PCT\_HOUSEHOLDER\_MULT", "ACS\_PCT\_AIAN\_COMB", "ACS\_PCT\_ASIAN\_COMB", "ACS\_PCT\_BLACK\_COMB", "ACS\_PCT\_NHPI\_COMB", "ACS\_PCT\_WHITE\_COMB",   
 "ACS\_PCT\_CHILD\_1FAM", "ACS\_PCT\_CHILDREN\_GRANDPARENT", "ACS\_PCT\_GRANDP\_RESPS\_NO\_P", "ACS\_PCT\_GRANDP\_RESPS\_P", #"ACS\_PCT\_GRANDP\_NO\_RESPS", "ACS\_PCT\_HH\_KID\_1PRNT",   
 "ACS\_PCT\_HH\_NO\_COMP\_DEV", "ACS\_PCT\_HH\_SMARTPHONE", #"ACS\_PCT\_HH\_SMARTPHONE\_ONLY",   
 "ACS\_PCT\_HH\_TABLET", #"ACS\_PCT\_HH\_TABLET\_ONLY",   
 "ACS\_PCT\_HH\_PC", #"ACS\_PCT\_HH\_PC\_ONLY",   
 "ACS\_PCT\_HH\_OTHER\_COMP", #"ACS\_PCT\_HH\_OTHER\_COMP\_ONLY",   
 "ACS\_PCT\_HH\_INTERNET", #"ACS\_PCT\_HH\_INTERNET\_NO\_SUBS", "ACS\_PCT\_HH\_BROADBAND", "ACS\_PCT\_HH\_BROADBAND\_ONLY",   
 "ACS\_PCT\_HH\_BROADBAND\_ANY", "ACS\_PCT\_HH\_CELLULAR", #"ACS\_PCT\_HH\_CELLULAR\_ONLY",   
 "ACS\_PCT\_HH\_NO\_INTERNET", "ACS\_PCT\_HH\_SAT\_INTERNET", "ACS\_PCT\_HH\_DIAL\_INTERNET\_ONLY", #"ACS\_PCT\_DIVORCED\_F", "ACS\_PCT\_DIVORCED\_M", "ACS\_PCT\_MARRIED\_SP\_AB\_F", "ACS\_PCT\_MARRIED\_SP\_AB\_M", "ACS\_PCT\_MARRIED\_SP\_PR\_F", "ACS\_PCT\_MARRIED\_SP\_PR\_M", "ACS\_PCT\_NVR\_MARRIED\_F", "ACS\_PCT\_NVR\_MARRIED\_M", "ACS\_PCT\_WIDOWED\_F", "ACS\_PCT\_WIDOWED\_M", "ACS\_PCT\_POP\_SAME\_SEX\_UNMRD\_P", "ACS\_PCT\_POP\_SAME\_SEX\_SPOUSE",   
 "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\_PVT\_PROFIT",   
 "ACS\_PCT\_RETAIL", "ACS\_PCT\_TRANSPORT", "ACS\_PCT\_WHOLESALE", #"ACS\_PCT\_WORK\_RES\_F", "ACS\_PCT\_WORK\_RES\_M",   
 "ACS\_PCT\_EMPLOYED", "ACS\_PCT\_UNEMPLOY", #"ACS\_PCT\_NOT\_LABOR", "ACS\_PCT\_VET\_UNEMPL\_18\_64", "ACS\_PCT\_VET\_LABOR\_FORCE\_18\_64", "ACS\_PCT\_ARMED\_FORCES",   
 "ACS\_GINI\_INDEX", #"ACS\_MDN\_GRNDPRNT\_NO\_PRNT\_INC", "ACS\_MDN\_GRNDPRNT\_INC", "ACS\_MEDIAN\_HH\_INC\_AIAN", "ACS\_MEDIAN\_HH\_INC\_ASIAN" , "ACS\_MEDIAN\_HH\_INC\_BLACK", "ACS\_MEDIAN\_HH\_INC\_HISP", "ACS\_MEDIAN\_HH\_INC\_MULTI", "ACS\_MEDIAN\_HH\_INC\_NHPI", "ACS\_MEDIAN\_HH\_INC\_OTHER", "ACS\_MEDIAN\_HH\_INC\_WHITE", "ACS\_MEDIAN\_HH\_INC", "ACS\_MEDIAN\_INC\_F", "ACS\_MEDIAN\_INC\_M", "ACS\_MEDIAN\_NONVET\_INC", "ACS\_MEDIAN\_VET\_INC", "ACS\_PCT\_INC50\_ABOVE65", "ACS\_PCT\_INC50\_BELOW17", "ACS\_PCT\_HEALTH\_INC\_BELOW137", "ACS\_PCT\_HEALTH\_INC\_138\_199", "ACS\_PCT\_HEALTH\_INC\_200\_399", "ACS\_PCT\_HEALTH\_INC\_ABOVE400",   
 "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\_PCT\_INC50", "ACS\_PCT\_NONVET\_POV\_18\_64", "ACS\_PCT\_VET\_POV\_18\_64", "ACS\_PCT\_PERSON\_INC\_100\_124", "ACS\_PCT\_PERSON\_INC\_125\_199", "ACS\_PCT\_PERSON\_INC\_ABOVE200", "ACS\_PCT\_PERSON\_INC\_BELOW99",   
 "ACS\_PER\_CAPITA\_INC", #"ACS\_PCT\_POV\_AIAN", "ACS\_PCT\_POV\_ASIAN", "ACS\_PCT\_POV\_BLACK", "ACS\_PCT\_POV\_HISPANIC", "ACS\_PCT\_POV\_MULTI", "ACS\_PCT\_POV\_NHPI", "ACS\_PCT\_POV\_OTHER", "ACS\_PCT\_POV\_WHITE", "ACS\_PCT\_HH\_1FAM\_FOOD\_STMP", "ACS\_PCT\_HH\_FOOD\_STMP", "ACS\_PCT\_HH\_PUB\_ASSIST", "ACS\_PCT\_HH\_FOOD\_STMP\_BLW\_POV", "ACS\_PCT\_HH\_NO\_FD\_STMP\_BLW\_POV", "ACS\_PCT\_COLLEGE\_ASSOCIATE\_DGR", "ACS\_PCT\_BACHELOR\_DGR", "ACS\_PCT\_NO\_WORK\_NO\_SCHL\_16\_19", "ACS\_PCT\_GRADUATE\_DGR", "ACS\_PCT\_HS\_GRADUATE", "ACS\_PCT\_LT\_HS", "ACS\_PCT\_POSTHS\_ED", "ACS\_PCT\_VET\_BACHELOR", "ACS\_PCT\_VET\_COLLEGE", "ACS\_PCT\_VET\_HS",   
 "ACS\_MEDIAN\_HOME\_VALUE", "ACS\_MEDIAN\_RENT", #"ACS\_PCT\_1UP\_RENT\_1ROOM", "ACS\_PCT\_1UP\_OWNER\_1ROOM", "ACS\_PCT\_1UP\_PERS\_1ROOM", "ACS\_PCT\_HH\_1PERS", "ACS\_PCT\_10UNITS", "ACS\_PCT\_GRP\_QRT", "ACS\_PCT\_HU\_MOBILE\_HOME", "ACS\_PCT\_OWNER\_HU", "ACS\_PCT\_OWNER\_HU\_CHILD", "ACS\_PCT\_RENTER\_HU", "ACS\_PCT\_RENTER\_HU\_ABOVE65", "ACS\_PCT\_RENTER\_HU\_CHILD", "ACS\_PCT\_RENTER\_HU\_COST\_30PCT", "ACS\_PCT\_RENTER\_HU\_COST\_50PCT",   
 "ACS\_PCT\_VACANT\_HU", #"ACS\_PCT\_HU\_NO\_FUEL", "ACS\_PCT\_HU\_UTILITY\_GAS", "ACS\_PCT\_HU\_BOT\_TANK\_LP\_GAS", "ACS\_PCT\_HU\_OIL", "ACS\_PCT\_HU\_WOOD", "ACS\_PCT\_HU\_COAL", "ACS\_PCT\_HU\_OTHER", "ACS\_PCT\_HU\_ELEC", "ACS\_PCT\_HU\_SOLAR", "ACS\_MDN\_OWNER\_COST\_MORTGAGE", "ACS\_MDN\_OWNER\_COST\_NO\_MORTG", #"ACS\_PCT\_OWNER\_HU\_COST\_30PCT", "ACS\_PCT\_OWNER\_HU\_COST\_50PCT", "ACS\_MEDIAN\_YEAR\_BUILT", "ACS\_PCT\_HU\_BUILT\_1979", "ACS\_PCT\_HU\_KITCHEN", "ACS\_PCT\_HU\_PLUMBING", "ACS\_PCT\_IN\_STATE\_MOVE", "ACS\_PCT\_IN\_COUNTY\_MOVE", "ACS\_PCT\_DIF\_STATE", "ACS\_PCT\_HH\_ABOVE65", "ACS\_PCT\_HH\_ALONE\_ABOVE65",   
 "ACS\_PCT\_COMMT\_15MIN", "ACS\_PCT\_COMMT\_29MIN", "ACS\_PCT\_COMMT\_59MIN", "ACS\_PCT\_COMMT\_60MINUP", #"ACS\_PCT\_HU\_NO\_VEH", "ACS\_PCT\_WORK\_NO\_CAR", "ACS\_PCT\_DRIVE\_2WORK",   
 "ACS\_PCT\_PUBL\_TRANSIT", #"ACS\_PCT\_PUB\_COMMT\_15MIN", "ACS\_PCT\_PUB\_COMMT\_29MIN", "ACS\_PCT\_PUB\_COMMT\_59MIN", "ACS\_PCT\_PUB\_COMMT\_60MINUP","ACS\_PCT\_TAXICAB\_2WORK", "ACS\_PCT\_WALK\_2WORK",   
 "ACS\_PCT\_MEDICAID\_ANY", #"ACS\_PCT\_MEDICAID\_ANY\_BELOW64",  
 "ACS\_PCT\_MEDICARE\_ONLY", #"ACS\_PCT\_OTHER\_INS", "ACS\_PCT\_PVT\_EMPL\_DRCT", "ACS\_PCT\_PVT\_EMPL\_DRCT\_BELOW64", "ACS\_PCT\_PRIVATE\_ANY", "ACS\_PCT\_PRIVATE\_ANY\_BELOW64", "ACS\_PCT\_PRIVATE\_EMPL", "ACS\_PCT\_PRIVATE\_EMPL\_BELOW64", "ACS\_PCT\_PRIVATE\_MDCR", "ACS\_PCT\_PRIVATE\_MDCR\_35\_64", "ACS\_PCT\_PRIVATE\_OTHER", "ACS\_PCT\_PRIVATE\_OTHER\_BELOW64", "ACS\_PCT\_PRIVATE\_SELF", "ACS\_PCT\_PRIVATE\_SELF\_BELOW64", "ACS\_PCT\_PUBLIC\_ONLY", "ACS\_PCT\_PUBLIC\_OTHER", "ACS\_PCT\_PUBLIC\_OTHER\_BELOW64", "ACS\_PCT\_SELF\_MDCR\_ABOVE35", "ACS\_PCT\_TRICARE\_VA", "ACS\_PCT\_TRICARE\_VA\_BELOW64", "ACS\_PCT\_UNINSURED", "ACS\_PCT\_UNINSURED\_BELOW64", "AHRF\_USDA\_RUCC\_2013", "AHRF\_VET", "AHRF\_VET\_MALE", "AHRF\_VET\_FEMALE", "AHRF\_UNEMPLOYED\_RATE", "AHRF\_DAYS\_AIR\_QLT", "AHRF\_PCT\_GOOD\_AQ", "AHRF\_TXC\_SITE\_NO\_DATA", "AHRF\_TXC\_SITE\_CNTRL", "AHRF\_TXC\_SITE\_NO\_CNTRL", "AHRF\_HPSA\_DENTIST", "AHRF\_HPSA\_MENTAL", "AHRF\_HPSA\_PRIM", "AHRF\_MARKET\_ENROL", "AHRF\_MARKET\_ENROL\_NEW", "AHRF\_MARKET\_ENROL\_ACTIVE", "AHRF\_MARKET\_ENROL\_AUTO", "AHRF\_MARKET\_ENROL\_150", "AHRF\_MARKET\_ENROL\_200", "AHRF\_MARKET\_ENROL\_250", "AHRF\_MARKET\_ENROL\_300", "AHRF\_MARKET\_ENROL\_400", "AHRF\_MARKET\_ENROL\_OTHER", "AHRF\_MARKET\_ENROL\_NO\_ASST", "AHRF\_PRESC\_ENROLLMENT", "AHRF\_PCT\_PRESC\_PEN", "AHRF\_TOT\_ADV\_NURSES",   
 "AHRF\_ADV\_NURSES\_RATE", #"AHRF\_TOT\_CLIN\_NURSE\_SPEC",   
 "AHRF\_CLIN\_NURSE\_SPEC\_RATE", #"AHRF\_TOT\_DENTISTS",   
 "AHRF\_DENTISTS\_RATE", #"AHRF\_TOT\_NURSE\_ANESTH",   
 "AHRF\_NURSE\_ANESTH\_RATE", #"AHRF\_TOT\_NURSE\_MIDWIVES",   
 "AHRF\_NURSE\_MIDWIVES\_RATE", #"AHRF\_TOT\_NURSE\_PRACT",   
 "AHRF\_NURSE\_PRACT\_RATE", #"AHRF\_TOT\_PHYSICIAN\_ASSIST",   
 "AHRF\_PHYSICIAN\_ASSIST\_RATE", #"AMFAR\_SSP",   
 "AMFAR\_SSP\_RATE", #"AMFAR\_TOT\_MEDSAFAC",  
 "AMFAR\_MEDSAFAC\_RATE", #"AMFAR\_TOT\_AMATFAC", "AMFAR\_AMATFAC\_RATE", "AMFAR\_TOT\_HCVTFAC", "AMFAR\_HCVTFAC\_RATE", "AMFAR\_TOT\_HIVHCVTFAC", "AMFAR\_HIVHCVTFAC\_RATE", "AMFAR\_TOT\_HIVTFAC", "AMFAR\_HIVTFAC\_RATE", "AMFAR\_TOT\_MEDAMATFAC", "AMFAR\_MEDAMATFAC\_RATE", "AMFAR\_TOT\_MEDHCVTFAC", "AMFAR\_MEDHCVTFAC\_RATE", "AMFAR\_TOT\_MEDHIVHCVTFAC", "AMFAR\_MEDHIVHCVTFAC\_RATE", "AMFAR\_TOT\_MEDHIVTFAC", "AMFAR\_MEDHIVTFAC\_RATE", "AMFAR\_TOT\_MEDMHFAC", "AMFAR\_MEDMHFAC\_RATE", "AMFAR\_TOT\_MHFAC",   
 "AMFAR\_MHFAC\_RATE", #"CAF\_ADJ\_COUNTY\_1", "CAF\_ADJ\_COUNTY\_2", "CAF\_ADJ\_COUNTY\_3", "CAF\_ADJ\_COUNTY\_4", "CAF\_ADJ\_COUNTY\_5", "CAF\_ADJ\_COUNTY\_6", "CAF\_ADJ\_COUNTY\_7", "CAF\_ADJ\_COUNTY\_8", "CAF\_ADJ\_COUNTY\_9", "CAF\_ADJ\_COUNTY\_10", "CAF\_ADJ\_COUNTY\_11", "CAF\_ADJ\_COUNTY\_12", "CAF\_ADJ\_COUNTY\_13", "CAF\_ADJ\_COUNTY\_14","CDCW\_TOT\_POPULATION", "CDCW\_INJURY\_DTH\_RATE", "CDCW\_TRANSPORT\_DTH\_RATE", "CDCW\_SELFHARM\_DTH\_RATE", "CDCW\_ASSAULT\_DTH\_RATE", "CDCW\_MATERNAL\_DTH\_RATE", "CDCW\_OPIOID\_DTH\_RATE", "CDCW\_DRUG\_DTH\_RATE",   
 "CEN\_AREALAND\_SQM\_COUNTY", "CEN\_POPDENSITY\_COUNTY", #"CHR\_TOT\_MENTAL\_PROV", "CHR\_MENTAL\_PROV\_RATE", "CHR\_TOT\_POPULATION", "MP\_MEDICARE\_ELIGIBLES", "MP\_MEDICARE\_ADVTG\_ENROLLED", "MP\_PCT\_ADVTG\_PEN", "NCHS\_URCS\_2006", "NCHS\_URCS\_2013", "NEPHTN\_ARSENIC\_MEAN\_CSW", "NEPHTN\_ARSENIC\_MEAN\_POP", "NEPHTN\_PCT\_ARSENIC\_MCL\_NOTDETECT", "NEPHTN\_PCT\_ARSENIC\_MCL\_LESS10", "NEPHTN\_PCT\_ARSENIC\_MCL\_GREATER10",   
 "NEPHTN\_HEATIND\_90", #"NEPHTN\_HEATIND\_95", "NEPHTN\_HEATIND\_100", "NEPHTN\_HEATIND\_105", "NEPHTN\_MAXDROUGHT", "NEPHTN\_NUMDROUGHT", "NEPHTN\_TEMPERATURE\_90", "NEPHTN\_TEMPERATURE\_95", "NEPHTN\_TEMPERATURE\_100", "NEPHTN\_TEMPERATURE\_105", "NHC\_AVG\_LIC\_STAFF", "NHC\_AVG\_REP\_NURSE\_STAFF", "NHC\_AVG\_ADJ\_NURSE\_STAFF", "NHC\_TOT\_FACS", "NHC\_FACS\_RATE",   
 "EPAA\_2NDMAX\_CO\_1HR", #"EPAA\_2NDMAX\_CO\_8HR",   
 "EPAA\_98PR\_NO2\_1HR", #"EPAA\_MEAN\_NO2\_1HR", "EPAA\_2NDMAX\_O3\_1HR", "EPAA\_4THMAX\_O3\_8HR",   
 "EPAA\_MAX\_PB\_3MON", #"EPAA\_2NDMAX\_PM10\_24HR", "EPAA\_MEAN\_WTD\_PM10", "EPAA\_MEAN\_WTD\_PM25",   
 "EPAA\_98PR\_PM25\_DAILY", "EPAA\_99PR\_SO2\_1HR", #"EPAA\_2NDMAX\_SO2\_24HR", "EPAA\_MEAN\_SO2\_1HR", "NOAAC\_AVG\_TEMP\_APR", "NOAAC\_AVG\_TEMP\_AUG", "NOAAC\_AVG\_TEMP\_DEC", "NOAAC\_AVG\_TEMP\_FEB", "NOAAC\_AVG\_TEMP\_JAN", "NOAAC\_AVG\_TEMP\_JUL", "NOAAC\_AVG\_TEMP\_JUN", "NOAAC\_AVG\_TEMP\_MAR", "NOAAC\_AVG\_TEMP\_MAY", "NOAAC\_AVG\_TEMP\_NOV", "NOAAC\_AVG\_TEMP\_OCT", "NOAAC\_AVG\_TEMP\_SEP", "NOAAC\_MAX\_TEMP\_APR", "NOAAC\_MAX\_TEMP\_AUG", "NOAAC\_MAX\_TEMP\_DEC", "NOAAC\_MAX\_TEMP\_FEB", "NOAAC\_MAX\_TEMP\_JAN", "NOAAC\_MAX\_TEMP\_JUL", "NOAAC\_MAX\_TEMP\_JUN", "NOAAC\_MAX\_TEMP\_MAR", "NOAAC\_MAX\_TEMP\_MAY", "NOAAC\_MAX\_TEMP\_NOV", "NOAAC\_MAX\_TEMP\_OCT", "NOAAC\_MAX\_TEMP\_SEP", "NOAAC\_MIN\_TEMP\_APR", "NOAAC\_MIN\_TEMP\_AUG", "NOAAC\_MIN\_TEMP\_DEC", "NOAAC\_MIN\_TEMP\_FEB", "NOAAC\_MIN\_TEMP\_JAN", "NOAAC\_MIN\_TEMP\_JUL", "NOAAC\_MIN\_TEMP\_JUN", "NOAAC\_MIN\_TEMP\_MAR", "NOAAC\_MIN\_TEMP\_MAY", "NOAAC\_MIN\_TEMP\_NOV", "NOAAC\_MIN\_TEMP\_OCT", "NOAAC\_MIN\_TEMP\_SEP", "NOAAC\_PRECIPITATION\_APR", "NOAAC\_PRECIPITATION\_AUG", "NOAAC\_PRECIPITATION\_DEC", "NOAAC\_PRECIPITATION\_FEB", "NOAAC\_PRECIPITATION\_JAN", "NOAAC\_PRECIPITATION\_JUL", "NOAAC\_PRECIPITATION\_JUN", "NOAAC\_PRECIPITATION\_MAR", "NOAAC\_PRECIPITATION\_MAY", "NOAAC\_PRECIPITATION\_NOV", "NOAAC\_PRECIPITATION\_OCT", "NOAAC\_PRECIPITATION\_SEP", "NOAAS\_PROPERTY\_DAMAGE", "NOAAS\_TOT\_DEATHS\_DIRECT", "NOAAS\_TOT\_DEATHS\_INDIRECT", "NOAAS\_TOT\_INJURIES\_DIRECT", "NOAAS\_TOT\_INJURIES\_INDIRECT", "NOAAS\_TOT\_STORMEVENT", "NOAAS\_TOT\_TORNADO", "NOAAS\_TOT\_WIND", "NOAAS\_TOT\_HAIL", "NOAAS\_TOT\_HURRICANE\_STORM", "NOAAS\_TOT\_FLOOD", "NOAAS\_TOT\_WILDFIRE", "NOAAS\_TOT\_HEAT\_EVENTS", "NOAAS\_TOT\_DROUGHT",   
 "SAIPE\_MEDIAN\_HH\_INCOME", "SAIPE\_PCT\_POV", #"SAIPE\_PCT\_POV\_0\_17", "SAIPE\_PCT\_POV\_5\_17", "SAIPE\_TOT\_POV", "SAIPE\_TOT\_POV\_0\_17", "SAIPE\_TOT\_POV\_5\_17", "WUSTL\_AVG\_PM25", "HHC\_PCT\_HHA\_NURSING", "HHC\_PCT\_HHA\_PHYS\_THERAPY", "HHC\_PCT\_HHA\_OCC\_THERAPY", "HHC\_PCT\_HHA\_SPEECH", "HHC\_PCT\_HHA\_MEDICAL", "HHC\_PCT\_HHA\_AIDE", "HIFLD\_MEDIAN\_DIST\_UC", "HIFLD\_MEAN\_DIST\_UC", "HIFLD\_MIN\_DIST\_UC", "HIFLD\_MAX\_DIST\_UC", "HIFLD\_UC", "HIFLD\_UC\_RATE", "IHS\_HEALTH\_FACILITY", "IHS\_HEALTH\_FACILITY\_RATE", "LTC\_AVG\_PCT\_PRESSURE\_ULCER", "LTC\_PCT\_RESD\_BLACK", "LTC\_PCT\_RESD\_HISPANIC", "LTC\_PCT\_RESD\_WHITE", "LTC\_AVG\_AGE", "LTC\_PCT\_MULTI\_FAC", "LTC\_PCT\_FOR\_PROFIT", "LTC\_AVG\_ACUITY\_INDEX", "LTC\_TOT\_BEDS", "LTC\_TOT\_BEDS\_RATE",  
 "LTC\_AVG\_OBS\_REHOSP\_RATE", "LTC\_AVG\_OBS\_SUCCESSFUL\_DISC\_RATE", #"LTC\_AVG\_PCT\_MEDICAID", "LTC\_AVG\_PCT\_MEDICARE", "LTC\_OCCUPANCY\_RATE", "MGV\_PCT\_MEDICAID", "MGV\_TOT\_BEN\_PART\_A\_B", "MGV\_TOT\_BEN\_FFS", "MGV\_PCT\_BEN\_FFS\_WHITE", "MGV\_PCT\_BEN\_FFS\_BLACK", "MGV\_PCT\_BEN\_FFS\_HISPANIC", "MGV\_PER\_CAPITA\_ACTUAL\_IP",   
 "MGV\_PER\_CAPITA\_STD\_IP", #"MGV\_PER\_CAPITA\_ACTUAL\_OP",   
 "MGV\_PER\_CAPITA\_STD\_OP", #"MGV\_PER\_CAPITA\_ACTUAL\_EM",   
 "MGV\_PER\_CAPITA\_STD\_EM", #"MGV\_PER\_CAPITA\_ACTUAL\_PA",   
 "MGV\_PER\_CAPITA\_STD\_PA", #"MGV\_PER\_CAPITA\_ACTUAL\_HC",   
 "MGV\_PER\_CAPITA\_STD\_HC", #"MMD\_OVERALL\_PQI\_M\_RATE", "MMD\_OVERALL\_PQI\_F\_RATE", "MMD\_OVERALL\_PQI\_WHITE\_RATE", "MMD\_OVERALL\_PQI\_BLACK\_RATE", "MMD\_OVERALL\_PQI\_OTHER\_RATE", "MMD\_OVERALL\_PQI\_ASIAN\_RATE", "MMD\_OVERALL\_PQI\_HISP\_RATE", "MMD\_OVERALL\_PQI\_AIAN\_RATE", "MMD\_ACUTE\_PQI\_M\_RATE", "MMD\_ACUTE\_PQI\_F\_RATE", "MMD\_ACUTE\_PQI\_WHITE\_RATE", "MMD\_ACUTE\_PQI\_BLACK\_RATE", "MMD\_ACUTE\_PQI\_OTHER\_RATE", "MMD\_ACUTE\_PQI\_ASIAN\_RATE", "MMD\_ACUTE\_PQI\_HISP\_RATE", "MMD\_ACUTE\_PQI\_AIAN\_RATE", "MMD\_CHRONIC\_PQI\_M\_RATE", "MMD\_CHRONIC\_PQI\_F\_RATE", "MMD\_CHRONIC\_PQI\_WHITE\_RATE", "MMD\_CHRONIC\_PQI\_BLACK\_RATE", "MMD\_CHRONIC\_PQI\_OTHER\_RATE", "MMD\_CHRONIC\_PQI\_ASIAN\_RATE", "MMD\_CHRONIC\_PQI\_HISP\_RATE", "MMD\_CHRONIC\_PQI\_AIAN\_RATE", "MMD\_ED\_VISITS\_M\_RATE", "MMD\_ED\_VISITS\_F\_RATE", "MMD\_ED\_VISITS\_WHITE\_RATE", "MMD\_ED\_VISITS\_BLACK\_RATE", "MMD\_ED\_VISITS\_OTHER\_RATE", "MMD\_ED\_VISITS\_ASIAN\_RATE", "MMD\_ED\_VISITS\_HISP\_RATE", "MMD\_ED\_VISITS\_AIAN\_RATE", "MMD\_ED\_VISITS\_MED\_RATE", "MMD\_ED\_VISITS\_DUAL\_RATE", "MMD\_READM\_M\_RATE", "MMD\_READM\_F\_RATE", "MMD\_READM\_WHITE\_RATE", "MMD\_READM\_BLACK\_RATE", "MMD\_READM\_OTHER\_RATE", "MMD\_READM\_ASIAN\_RATE", "MMD\_READM\_HISP\_RATE", "MMD\_READM\_AIAN\_RATE", "MMD\_ANXIETY\_DISD", "MMD\_BIPOLAR\_DISD", "MMD\_DEPR\_DISD", "MMD\_PERSONALITY\_DISD", "MMD\_OUD\_IND", "PC\_PCT\_MEDICARE\_APPRVD\_FULL\_AMT", "PC\_PCT\_MCARE\_MAY\_ACPT\_APPRVD\_AMT",   
 "POS\_MEDIAN\_DIST\_ED", #"POS\_MEAN\_DIST\_ED", "POS\_MIN\_DIST\_ED", "POS\_MAX\_DIST\_ED",   
 "POS\_MEDIAN\_DIST\_MEDSURG\_ICU", #"POS\_MEAN\_DIST\_MEDSURG\_ICU", "POS\_MIN\_DIST\_MEDSURG\_ICU", "POS\_MAX\_DIST\_MEDSURG\_ICU",   
 "POS\_MEDIAN\_DIST\_TRAUMA", #"POS\_MEAN\_DIST\_TRAUMA", "POS\_MIN\_DIST\_TRAUMA", "POS\_MAX\_DIST\_TRAUMA",   
 "POS\_MEDIAN\_DIST\_PED\_ICU", #"POS\_MEAN\_DIST\_PED\_ICU", "POS\_MIN\_DIST\_PED\_ICU", "POS\_MAX\_DIST\_PED\_ICU",   
 "POS\_MEDIAN\_DIST\_OBSTETRICS", #"POS\_MEAN\_DIST\_OBSTETRICS", "POS\_MIN\_DIST\_OBSTETRICS", "POS\_MAX\_DIST\_OBSTETRICS",   
 "POS\_MEDIAN\_DIST\_CLINIC", #"POS\_MEAN\_DIST\_CLINIC", "POS\_MIN\_DIST\_CLINIC", "POS\_MAX\_DIST\_CLINIC",   
 "POS\_MEDIAN\_DIST\_ALC", #"POS\_MEAN\_DIST\_ALC", "POS\_MIN\_DIST\_ALC", "POS\_MAX\_DIST\_ALC", "POS\_TOT\_FQHC", "POS\_FQHC\_RATE", "POS\_TOT\_CMHC", "POS\_CMHC\_RATE", "POS\_TOT\_RHC", "POS\_RHC\_RATE", "POS\_TOT\_HHA", "POS\_HHA\_RATE", "POS\_TOT\_HOSPICE", "POS\_HOSPICE\_RATE", "POS\_TOT\_ASC", "POS\_ASC\_RATE", "POS\_TOT\_NF", "POS\_NF\_RATE", "POS\_TOT\_NF\_BEDS", "POS\_NF\_BEDS\_RATE", "POS\_TOT\_SNF", "POS\_SNF\_RATE", "POS\_TOT\_SNF\_BEDS", "POS\_SNF\_BEDS\_RATE", "POS\_TOT\_HOSP\_OBSTETRIC", "POS\_HOSP\_OBSTETRIC\_RATE", "POS\_TOT\_HOSP\_PED\_ICU", "POS\_HOSP\_PED\_ICU\_RATE", "POS\_TOT\_HOSP\_BURN", "POS\_HOSP\_BURN\_RATE", "POS\_TOT\_HOSP\_MEDSURG\_ICU", "POS\_HOSP\_MEDSURG\_ICU\_RATE", "POS\_TOT\_HOSP\_REHAB", "POS\_HOSP\_REHAB\_RATE", "POS\_TOT\_HOSP\_ALC", "POS\_HOSP\_ALC\_RATE", "POS\_TOT\_HOSP\_PSYCH", "POS\_HOSP\_PSYCH\_RATE", "POS\_TOT\_HOSP\_AMBULANCE", "POS\_HOSP\_AMBULANCE\_RATE", "POS\_TOT\_HOSP\_CHEMO", "POS\_HOSP\_CHEMO\_RATE", "POS\_TOT\_HOSP\_ED", "POS\_HOSP\_ED\_RATE", "POS\_PCT\_HOSP\_FOR\_PROFIT", "POS\_PCT\_HOSP\_NON\_PROFIT", "POS\_PCT\_HOSP\_GOV", "CEN\_AIAN\_NH\_IND"  
 ) %>%   
 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)

sdoh\_data <- sdoh\_data %>%   
 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",  
 "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"   
 )

# 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(  
 #"statecode","countycode",  
 "fipscode", #"state","county", "year", "county\_ranked", "v001\_rawvalue", "v001\_numerator", "v001\_denominator", "v001\_cilow", "v001\_cihigh", "v001\_flag", "v001\_race\_aian", "v001\_race\_aian\_cilow", "v001\_race\_aian\_cihigh", "v001\_race\_aian\_flag", "v001\_race\_asian", "v001\_race\_asian\_cilow", "v001\_race\_black", "v001\_race\_black\_cilow", "v001\_race\_hispanic", "v001\_race\_hispanic\_cilow", "v001\_race\_hispanic\_cihigh", "v001\_race\_hispanic\_flag", "v001\_race\_white", "v001\_race\_white\_cilow", "v001\_race\_white\_cihigh", "v001\_race\_white\_flag",   
 "v002\_rawvalue", #"v002\_numerator", "v002\_denominator", "v002\_cilow", "v002\_cihigh", "v036\_rawvalue", "v036\_numerator", "v036\_denominator", "v036\_cilow", "v036\_cihigh",  
 "v042\_rawvalue", #"v042\_numerator", "v042\_denominator", "v042\_cilow", "v042\_cihigh",  
 "v037\_rawvalue", #"v037\_numerator", "v037\_denominator", "v037\_cilow", "v037\_cihigh", "v037\_flag", "v037\_race\_aian", "v037\_race\_aian\_cilow", "v037\_race\_aian\_cihigh", "v037\_race\_asian", "v037\_race\_asian\_cilow", "v037\_race\_asian\_cihigh", "v037\_race\_black", "v037\_race\_black\_cilow", "v037\_race\_black\_cihigh", "v037\_race\_hispanic", "v037\_race\_hispanic\_cilow", "v037\_race\_hispanic\_cihigh", "v037\_race\_white", "v037\_race\_white\_cilow", "v037\_race\_white\_cihigh",  
 "v009\_rawvalue", #"v009\_numerator", "v009\_denominator", "v009\_cilow", "v009\_cihigh",  
 "v011\_rawvalue", #"v011\_numerator", "v011\_denominator", "v011\_cilow", "v011\_cihigh",   
 "v133\_rawvalue", #"v133\_numerator", "v133\_denominator", "v133\_cilow", "v133\_cihigh",   
 "v070\_rawvalue", #"v070\_numerator", "v070\_denominator", "v070\_cilow", "v070\_cihigh",   
 "v132\_rawvalue", #"v132\_numerator", "v132\_denominator", "v132\_cilow", "v132\_cihigh",   
 "v049\_rawvalue", #"v049\_numerator", "v049\_denominator", "v049\_cilow", "v049\_cihigh", "v134\_rawvalue", "v134\_numerator", "v134\_denominator", "v134\_cilow", "v134\_cihigh", "v045\_rawvalue", "v045\_numerator", "v045\_denominator", "v045\_cilow", "v045\_cihigh",   
 "v014\_rawvalue", #"v014\_numerator", "v014\_denominator", "v014\_cilow", "v014\_cihigh", "v014\_race\_aian", "v014\_race\_aian\_cihigh", "v014\_race\_asian", "v014\_race\_asian\_cilow", "v014\_race\_asian\_cihigh", "v014\_race\_black", "v014\_race\_black\_cilow", "v014\_race\_black\_cihigh", "v014\_race\_hispanic", "v014\_race\_hispanic\_cilow", "v014\_race\_hispanic\_cihigh", "v014\_race\_white", "v014\_race\_white\_cilow", "v014\_race\_white\_cihigh",  
 "v085\_rawvalue", #"v085\_numerator", "v085\_denominator", "v085\_cilow", "v085\_cihigh",   
 "v004\_rawvalue", #"v004\_numerator", "v004\_denominator", "v004\_cilow", "v004\_cihigh","v004\_other\_data\_1", "v088\_rawvalue", "v088\_numerator", "v088\_denominator", "v088\_cilow", "v088\_cihigh", "v088\_other\_data\_1",  
 "v062\_rawvalue", #"v062\_numerator", "v062\_denominator", "v062\_cilow", "v062\_cihigh", "v062\_other\_data\_1",   
 "v005\_rawvalue", #"v005\_numerator", "v005\_denominator", "v005\_cilow", "v005\_cihigh", "v005\_race\_aian", "v005\_race\_asian", "v005\_race\_black", "v005\_race\_hispanic", "v005\_race\_white",   
 "v050\_rawvalue", #"v050\_numerator", "v050\_denominator", "v050\_cilow", "v050\_cihigh", "v050\_race\_aian", "v050\_race\_asian", "v050\_race\_black", "v050\_race\_hispanic", "v050\_race\_white",   
 "v155\_rawvalue", #"v155\_numerator", "v155\_denominator", "v155\_cilow", "v155\_cihigh", "v155\_race\_aian", "v155\_race\_asian", "v155\_race\_black", "v155\_race\_hispanic", "v155\_race\_white",   
 "v168\_rawvalue", #"v168\_numerator", "v168\_denominator", "v168\_cilow", "v168\_cihigh",   
 "v069\_rawvalue", #"v069\_numerator", "v069\_denominator", "v069\_cilow", "v069\_cihigh",   
 "v023\_rawvalue", #"v023\_numerator", "v023\_denominator", "v023\_cilow", "v023\_cihigh",  
 "v024\_rawvalue", #"v024\_numerator", "v024\_denominator", "v024\_cilow", "v024\_cihigh", "v024\_race\_aian", "v024\_race\_aian\_cilow", "v024\_race\_aian\_cihigh", "v024\_race\_asian", "v024\_race\_asian\_cilow", "v024\_race\_asian\_cihigh", "v024\_race\_black", "v024\_race\_black\_cilow", "v024\_race\_black\_cihigh", "v024\_race\_hispanic", "v024\_race\_hispanic\_cilow", "v024\_race\_hispanic\_cihigh", "v024\_race\_white", "v024\_race\_white\_cilow", "v024\_race\_white\_cihigh",   
 "v044\_rawvalue", #"v044\_numerator", "v044\_denominator", "v044\_cilow", "v044\_cihigh", "v082\_rawvalue", "v082\_numerator", "v082\_denominator", "v082\_cilow", "v082\_cihigh",   
 "v140\_rawvalue", #"v140\_numerator", "v140\_denominator", "v140\_cilow", "v140\_cihigh",   
 "v135\_rawvalue", #"v135\_numerator", "v135\_denominator", "v135\_cilow", "v135\_cihigh", "v135\_race\_aian", "v135\_race\_aian\_cilow", "v135\_race\_aian\_cihigh", "v135\_race\_asian", "v135\_race\_asian\_cilow", "v135\_race\_asian\_cihigh", "v135\_race\_black", "v135\_race\_black\_cilow", "v135\_race\_black\_cihigh", "v135\_race\_hispanic" , "v135\_race\_hispanic\_cilow", "v135\_race\_hispanic\_cihigh", "v135\_race\_white", "v135\_race\_white\_cilow", "v135\_race\_white\_cihigh",   
 "v125\_rawvalue", #"v125\_numerator", "v125\_denominator", "v125\_cilow", "v125\_cihigh",   
 "v124\_rawvalue", #"v124\_numerator", "v124\_denominator", "v124\_cilow", "v124\_cihigh", "v136\_rawvalue", "v136\_numerator", "v136\_denominator", "v136\_cilow", "v136\_cihigh",   
 "v136\_other\_data\_1", #"v136\_other\_data\_1\_cilow", "v136\_other\_data\_1\_cihigh",  
 "v136\_other\_data\_2", #"v136\_other\_data\_2\_cilow", "v136\_other\_data\_2\_cihigh",   
 "v136\_other\_data\_3", #"v136\_other\_data\_3\_cilow", "v136\_other\_data\_3\_cihigh", "v067\_rawvalue", "v067\_numerator", "v067\_denominator", "v067\_cilow", "v067\_cihigh", "v067\_race\_aian", "v067\_race\_aian\_cilow", "v067\_race\_aian\_cihigh", "v067\_race\_asian", "v067\_race\_asian\_cilow", "v067\_race\_asian\_cihigh", "v067\_race\_black", "v067\_race\_black\_cilow", "v067\_race\_black\_cihigh", "v067\_race\_hispanic", "v067\_race\_hispanic\_cilow", "v067\_race\_hispanic\_cihigh", "v067\_race\_white", "v067\_race\_white\_cilow", "v067\_race\_white\_cihigh",   
 "v137\_rawvalue", #"v137\_numerator", "v137\_denominator", "v137\_cilow", "v137\_cihigh",   
 "v147\_rawvalue", #"v147\_numerator", "v147\_denominator", "v147\_cilow", "v147\_cihigh", "v147\_race\_aian", "v147\_race\_aian\_cilow", "v147\_race\_aian\_cihigh", "v147\_race\_asian", "v147\_race\_asian\_cilow", "v147\_race\_asian\_cihigh", "v147\_race\_black", "v147\_race\_black\_cilow", "v147\_race\_black\_cihigh", "v147\_race\_hispanic", "v147\_race\_hispanic\_cilow", "v147\_race\_hispanic\_cihigh", "v147\_race\_white", "v147\_race\_white\_cilow", "v147\_race\_white\_cihigh",   
 "v127\_rawvalue", #"v127\_numerator", "v127\_denominator", "v127\_cilow", "v127\_cihigh", "v127\_race\_aian", "v127\_race\_aian\_cilow", "v127\_race\_aian\_cihigh", "v127\_race\_asian", "v127\_race\_asian\_cilow", "v127\_race\_asian\_cihigh", "v127\_race\_black", "v127\_race\_black\_cilow", "v127\_race\_black\_cihigh", "v127\_race\_hispanic", "v127\_race\_hispanic\_cilow", "v127\_race\_hispanic\_cihigh", "v127\_race\_white", "v127\_race\_white\_cilow", "v127\_race\_white\_cihigh",   
 "v128\_rawvalue", #"v128\_numerator", "v128\_denominator", "v128\_cilow", "v128\_cihigh", "v128\_race\_aian", "v128\_race\_aian\_cilow", "v128\_race\_aian\_cihigh", "v128\_race\_asian", "v128\_race\_asian\_cilow", "v128\_race\_asian\_cihigh", "v128\_race\_black", "v128\_race\_black\_cilow", "v128\_race\_black\_cihigh", "v128\_race\_hispanic", "v128\_race\_hispanic\_cilow", "v128\_race\_hispanic\_cihigh", "v128\_race\_white", "v128\_race\_white\_cilow", "v128\_race\_white\_cihigh",   
 "v129\_rawvalue", #"v129\_numerator", "v129\_denominator", "v129\_cilow", "v129\_cihigh", "v129\_race\_aian", "v129\_race\_aian\_cilow", "v129\_race\_aian\_cihigh", "v129\_race\_asian", "v129\_race\_asian\_cilow","v129\_race\_asian\_cihigh", "v129\_race\_black", "v129\_race\_black\_cilow", "v129\_race\_black\_cihigh", "v129\_race\_hispanic", "v129\_race\_hispanic\_cilow", "v129\_race\_hispanic\_cihigh", "v129\_race\_white", "v129\_race\_white\_cilow", "v129\_race\_white\_cihigh",   
 "v144\_rawvalue", #"v144\_numerator", "v144\_denominator", "v144\_cilow", "v144\_cihigh", "v145\_rawvalue", "v145\_numerator", "v145\_denominator", "v145\_cilow", "v145\_cihigh", "v060\_rawvalue", "v060\_numerator", "v060\_denominator", "v060\_cilow", "v060\_cihigh",   
 "v061\_rawvalue", #"v061\_numerator", "v061\_denominator", "v061\_cilow", "v061\_cihigh",   
 "v139\_rawvalue", #"v139\_numerator", "v139\_denominator", "v139\_cilow", "v139\_cihigh", "v083\_rawvalue", "v083\_numerator", "v083\_denominator", "v083\_cilow", "v083\_cihigh",   
 "v138\_rawvalue", #"v138\_numerator", "v138\_denominator", "v138\_cilow", "v138\_cihigh", "v138\_race\_aian", "v138\_race\_aian\_cilow", "v138\_race\_aian\_cihigh", "v138\_race\_asian", "v138\_race\_asian\_cilow", "v138\_race\_asian\_cihigh", "v138\_race\_black", "v138\_race\_black\_cilow", "v138\_race\_black\_cihigh", "v138\_race\_hispanic", "v138\_race\_hispanic\_cilow", "v138\_race\_hispanic\_cihigh", "v138\_race\_white", "v138\_race\_white\_cilow", "v138\_race\_white\_cihigh",   
 "v143\_rawvalue", #"v143\_numerator", "v143\_denominator", "v143\_cilow", "v143\_cihigh", "v003\_rawvalue", "v003\_numerator", "v003\_denominator", "v003\_cilow", "v003\_cihigh", "v122\_rawvalue", "v122\_numerator", "v122\_denominator", "v122\_cilow", "v122\_cihigh", "v131\_rawvalue", "v131\_numerator, "v131\_denominator", "v131\_cilow", "v131\_cihigh", "v131\_other\_data\_1",   
 "v021\_rawvalue", #"v021\_numerator", "v021\_denominator", "v021\_cilow", "v021\_cihigh",   
 "v149\_rawvalue", #"v149\_numerator", "v149\_denominator", "v149\_cilow", "v149\_cihigh",   
 "v159\_rawvalue", #"v159\_numerator", "v159\_denominator", "v159\_cilow", "v159\_cihigh", "v159\_race\_aian", "v159\_race\_asian", "v159\_race\_black", "v159\_race\_hispanic", "v159\_race\_white",   
 "v160\_rawvalue", #"v160\_numerator", "v160\_denominator", "v160\_cilow", "v160\_cihigh", "v160\_race\_aian", "v160\_race\_asian", "v160\_race\_black", "v160\_race\_hispanic", "v160\_race\_white",   
 "v167\_rawvalue", #"v167\_numerator", "v167\_denominator", "v167\_cilow", "v167\_cihigh",   
 "v169\_rawvalue", #"v169\_numerator", "v169\_denominator", "v169\_cilow", "v169\_cihigh",   
 "v151\_rawvalue", #"v151\_numerator", "v151\_denominator", "v151\_cilow", "v151\_cihigh",   
 "v063\_rawvalue", #"v063\_numerator", "v063\_denominator", "v063\_cilow", "v063\_cihigh", "v063\_race\_aian", "v063\_race\_aian\_cilow", "v063\_race\_aian\_cihigh", "v063\_race\_asian", "v063\_race\_asian\_cilow", "v063\_race\_asian\_cihigh", "v063\_race\_black", "v063\_race\_black\_cilow", "v063\_race\_black\_cihigh", "v063\_race\_hispanic", "v063\_race\_hispanic\_cilow", "v063\_race\_hispanic\_cihigh", "v063\_race\_white", "v063\_race\_white\_cilow", "v063\_race\_white\_cihigh",   
 "v170\_rawvalue", #"v170\_numerator", "v170\_denominator", "v170\_cilow", "v170\_cihigh",   
 "v065\_rawvalue", #"v065\_numerator", "v065\_denominator", "v065\_cilow", "v065\_cihigh",   
 "v141\_rawvalue", #"v141\_numerator", "v141\_denominator", "v141\_cilow", "v141\_cihigh",   
 "v171\_rawvalue", #"v171\_numerator", "v171\_denominator", "v171\_cilow", "v171\_cihigh", "v172\_rawvalue", "v172\_numerator", "v172\_denominator", "v172\_cilow", "v172\_cihigh",   
 "v015\_rawvalue", #"v015\_numerator", "v015\_denominator", "v015\_cilow", "v015\_cihigh", "v015\_race\_aian", "v015\_race\_aian\_cilow", "v015\_race\_aian\_cihigh", "v015\_race\_asian", "v015\_race\_asian\_cilow", "v015\_race\_asian\_cihigh", "v015\_race\_black", "v015\_race\_black\_cilow", "v015\_race\_black\_cihigh", "v015\_race\_hispanic", "v015\_race\_hispanic\_cilow", "v015\_race\_hispanic\_cihigh", "v015\_race\_white", "v015\_race\_white\_cilow", "v015\_race\_white\_cihigh",   
 "v161\_rawvalue", #"v161\_numerator", "v161\_denominator", "v161\_cilow", "v161\_cihigh", "v161\_other\_data\_1", "v161\_race\_aian", "v161\_race\_aian\_cilow", "v161\_race\_aian\_cihigh", "v161\_race\_asian", "v161\_race\_asian\_cilow", "v161\_race\_asian\_cihigh", "v161\_race\_black", "v161\_race\_black\_cilow", "v161\_race\_black\_cihigh", "v161\_race\_hispanic", "v161\_race\_hispanic\_cilow", "v161\_race\_hispanic\_cihigh", "v161\_race\_white", "v161\_race\_white\_cilow", "v161\_race\_white\_cihigh",   
 "v148\_rawvalue", #"v148\_numerator", "v148\_denominator", "v148\_cilow", "v148\_cihigh", "v148\_race\_aian", "v148\_race\_aian\_cilow", "v148\_race\_aian\_cihigh", "v148\_race\_asian", "v148\_race\_asian\_cilow", "v148\_race\_asian\_cihigh", "v148\_race\_black", "v148\_race\_black\_cilow", "v148\_race\_black\_cihigh", "v148\_race\_hispanic", "v148\_race\_hispanic\_cilow", "v148\_race\_hispanic\_cihigh", "v148\_race\_white", "v148\_race\_white\_cilow", "v148\_race\_white\_cihigh", "v039\_rawvalue", "v039\_numerator", "v039\_denominator", "v039\_cilow", "v039\_cihigh", "v039\_race\_aian", "v039\_race\_aian\_cilow", "v039\_race\_aian\_cihigh", "v039\_race\_asian", "v039\_race\_asian\_cilow", "v039\_race\_asian\_cihigh", "v039\_race\_black", "v039\_race\_black\_cilow", "v039\_race\_black\_cihigh", "v039\_race\_hispanic", "v039\_race\_hispanic\_cilow", "v039\_race\_hispanic\_cihigh", "v039\_race\_white", "v039\_race\_white\_cilow", "v039\_race\_white\_cihigh",   
 "v158\_rawvalue", #"v158\_numerator", "v158\_denominator", "v158\_cilow", "v158\_cihigh", "v158\_other\_data\_1", "v158\_other\_data\_2",   
 "v177\_rawvalue", #"v177\_numerator", "v177\_denominator", "v177\_cilow", "v177\_cihigh", "v178\_rawvalue", "v178\_numerator", "v178\_denominator", "v178\_cilow", "v178\_cihigh",   
 "v156\_rawvalue", #"v156\_numerator", "v156\_denominator", "v156\_cilow", "v156\_cihigh", "v153\_rawvalue",   
 "v153\_numerator", #"v153\_denominator", "v153\_cilow", "v153\_cihigh", "v154\_rawvalue", "v154\_numerator", "v154\_denominator", "v154\_cilow", "v154\_cihigh", "v166\_rawvalue", "v166\_numerator", "v166\_denominator", "v166\_cilow", "v166\_cihigh", "v051\_rawvalue", "v051\_numerator", "v051\_denominator", "v051\_cilow", "v051\_cihigh",   
 "v052\_rawvalue", #"v052\_numerator", "v052\_denominator", "v052\_cilow", "v052\_cihigh",   
 "v053\_rawvalue", #"v053\_numerator", "v053\_denominator", "v053\_cilow", "v053\_cihigh", "v054\_rawvalue", "v054\_numerator", "v054\_denominator", "v054\_cilow", "v054\_cihigh", "v055\_rawvalue", "v055\_numerator", "v055\_denominator", "v055\_cilow", "v055\_cihigh", "v081\_rawvalue", "v081\_numerator", "v081\_denominator", "v081\_cilow", "v081\_cihigh", "v080\_rawvalue", "v080\_numerator" , "v080\_denominator", "v080\_cilow", "v080\_cihigh", "v056\_rawvalue", "v056\_numerator", "v056\_denominator", "v056\_cilow", "v056\_cihigh", "v126\_rawvalue", "v126\_numerator", "v126\_denominator", "v126\_cilow", "v126\_cihigh", "v059\_rawvalue", "v059\_numerator", "v059\_denominator", "v059\_cilow", "v059\_cihigh", "v057\_rawvalue", "v057\_numerator", "v057\_denominator", "v057\_cilow", "v057\_cihigh",   
 "v058\_rawvalue", #"v058\_numerator", "v058\_denominator", "v058\_cilow", "v058\_cihigh"  
 ) %>%   
 mutate(pcp\_pt = v004\_rawvalue/100) %>%   
 select(-v004\_rawvalue)

chr\_data <- chr\_data %>%  
 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",  
 "pct\_high\_housing\_costs" = "v136\_other\_data\_1",  
 "pct\_overcrowded\_hh" = "v136\_other\_data\_2",  
 "pct\_no\_kitchen\_or\_plumbing" = "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"  
 )

# convert fips code to numeric to match sdoh value  
sdoh\_data <- sdoh\_data %>%   
 mutate(fips\_code = as.numeric(fips\_code))  
# check for errors  
print(sum(is.na(sdoh\_data$fips\_code)))

## [1] 0

*Combined data*

# join sdoh and chr datasets   
qol\_data <- sdoh\_data %>%  
 inner\_join(chr\_data, "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

# clean enviroment, remove large data frames: mg  
rm(  
 all\_chr\_data,  
 all\_sdoh\_data  
 )

sdoh\_data\_2 <- sdoh\_data %>%  
 select(  
 "fips\_code",  
 "weighted\_population",  
 "average\_hh\_size",  
 "pct\_male",  
 "pct\_native\_american",  
 "pct\_asian",  
 "pct\_black",  
 "pct\_hispanic",  
 "pct\_other\_race",  
 "pct\_white",  
 "pct\_single\_parent",  
 "pct\_hh\_other\_computer",  
 "pct\_hh\_internet",  
 "pct\_employed",  
 "pct\_hh\_inc\_99999",  
 "pct\_w\_medicare",  
 "clinical\_nurse\_pt",  
 "dentist\_pt",  
 "pa\_pt",  
 "mental\_health\_faciliy\_pt",  
 "population\_density",  
 "days\_over\_90\_f",  
 "median\_er\_dist",  
 "median\_trauma\_center\_dist",  
 "median\_pediatric\_icu\_dist",  
 "median\_health\_clinic\_dist",  
 "median\_drug\_alcohol\_care\_dist",  
 "percent\_grandparents\_as\_guardians"  
 )

chr\_2 <- chr\_data %>%  
 select(  
 "fips\_code",  
 "median\_hh\_income",  
 "pct\_poor\_to\_fair\_health",  
 "pct\_adult\_smokers",  
 "pct\_binge\_drinkers",  
 "pct\_under\_65\_no\_health\_insurance",  
 "pct\_highschool\_diploma",  
 "inequality\_ratio",  
 "social\_clubs\_per\_10k",  
 "air\_polution\_metric",  
 "water\_quality",  
 "pct\_high\_housing\_costs",  
 "pct\_overcrowded\_hh",  
 "pct\_30\_min\_plus\_commute",  
 "school\_funding\_gap",  
 "pct\_voters",  
 "pct\_home\_owner",  
 "pct\_65\_plus",  
 "pct\_rural\_population",  
 "pct\_obese\_adults",  
 "pct\_food\_insecurities"  
 )

beta\_data <- merge(  
 sdoh\_data\_2,  
 chr\_2,  
 by = "fips\_code")  
  
beta\_data <- na.omit(beta\_data)  
  
beta\_data <- beta\_data %>%  
 mutate(response = ifelse(pct\_poor\_to\_fair\_health >= 0.154, "worse", "better")) %>%  
 mutate(response = as.factor(response))

rm(chr\_2)  
rm(sdoh\_data\_2)

beta\_codebook <- beta\_data %>%  
 cb\_add\_col\_attributes(  
 fips\_code,  
 description = "State-county FIPS Code (5-digit)",  
 source = "Both SDOH and CHR") %>%  
 cb\_add\_col\_attributes(  
 weighted\_population,  
 description = "Total weighted population",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 average\_hh\_size,  
 description = "Average household size",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_male,  
 description = "Percentage of population that is male",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_native\_american,  
 description = "Percentage of population reporting American Indian and Alaska Native race alone",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_asian,  
 description = "Percentage of population reporting Asian race alone",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_black,  
 description = "Percentage of population reporting Black or African American ethnicity alone",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_hispanic,  
 description = "Percentage of population reporting Hispanic ethnicity",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_other\_race,  
 description = "Percentage of population reporting some other ethnicity alone",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_white,  
 description = "Percentage of population reporting White race alone",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_single\_parent,  
 description = "Percentage of families with children that are single-parent families",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_hh\_other\_computer,  
 description = "Percentage of households with other type of computer",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_hh\_internet,  
 description = "Percentage of households with internet access",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_employed,  
 description = "Percentage of civilian labor force that is employed (ages 16 and over)",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_hh\_inc\_99999,  
 description = "Percentage of population with household income between $50,000 and $99,999",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_w\_medicare,  
 description = "Percentage of population with Medicare",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 clinical\_nurse\_pt,  
 description = "Total number of clinical nurse specialists with NPI per 1,000 population",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 dentist\_pt,  
 description = "Total number of dentists with NPI per 1,000 population",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pa\_pt,  
 description = "Total number of physician assistants with NPI per 1,000 population",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 mental\_health\_faciliy\_pt,  
 description = "Total number of facilities that provide mental health services per 1,000 population",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 population\_density,  
 description = "Population density (County)",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 days\_over\_90\_f,  
 description = "Number of days over 90 degrees Fahrenheit",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 median\_hh\_income,  
 description = "Estimated median household income",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 median\_er\_dist,  
 description = "Median distance in miles to the nearest emergency department, calculated using population weighted tract centroids in the county",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 median\_trauma\_center\_dist,  
 description = "Median distance in miles to the nearest designated trauma center, calculated using population weighted tract centroids in the county",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 median\_pediatric\_icu\_dist,  
 description = "Median distance in miles to the nearest pediatric ICU, calculated using population weighted tract centroids in the county",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 median\_health\_clinic\_dist,  
 description = "Median distance in miles to the nearest health clinic (FQHC, RHC), calculated using population weighted tract centroids in the county",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 median\_drug\_alcohol\_care\_dist,  
 description = "Median distance in miles to the nearest hospital with alcohol and drug abuse inpatient care, calculated using population weighted tract centroids in the county",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 percent\_grandparents\_as\_guardians,  
 description = "Percentage of children living with grandparent householder whose grandparent is responsible for them: parent may or may not be present",  
 source = "SDOH") %>%  
 cb\_add\_col\_attributes(  
 pct\_poor\_to\_fair\_health,  
 description = "Percentage of adults reporting fair or poor health",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 pct\_adult\_smokers,  
 description = "Percentage of adults who are current smokers",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 pct\_binge\_drinkers,  
 description = "Percentage of adults reporting binge or heavy drinking",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 pct\_under\_65\_no\_health\_insurance,  
 description = "Percentage of population with no health insurance",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 pct\_highschool\_diploma,  
 description = "Percentage of adults ages 25 and over with a high school diploma or equivalent",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 inequality\_ratio,  
 description = "Ratio of household income at the 80th percentile to income at the 20th percentile",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 social\_clubs\_per\_10k,  
 description = "Number of membership associations per 10,000 population",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 air\_polution\_metric,  
 description = "Average daily density of fine particulate matter in micrograms per cubic meter",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 water\_quality,  
 description = "Indicator of the presence of health-related drinking water violations. 0=No, 1=Yes",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 pct\_high\_housing\_costs,  
 description = "Percentage of households with severe cost burden – monthly housing costs (including utilities) exceed 50% of monthly income",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 pct\_overcrowded\_hh,  
 description = "Percentage of households with overcrowding – more than 1 person per room",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 pct\_30\_min\_plus\_commute,  
 description = "Percentage of workers who commute in their car alone, the percentage that commute more than 30 minutes",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 school\_funding\_gap,  
 description = "The average gap in dollars between actual and required spending per pupil among public school districts. Required spending is an estimate of dollars needed to achieve U.S. average test scores in each district",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 pct\_voters,  
 description = "Percentage for voter turnout",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 pct\_home\_owner,  
 description = "Percentage of owner-occupied housing units",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 pct\_65\_plus,  
 description = "Percentage of 65 and Older raw value",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 pct\_rural\_population,  
 description = "Percentage Rural raw value",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 pct\_obese\_adults,  
 description = "Percentage of the adult population that is obese (ages 20 and over)",  
 source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 pct\_food\_insecurities,  
 description = "Percentage of population who lack  
adequate access to food",  
source = "CHR") %>%  
 cb\_add\_col\_attributes(  
 response,  
 description = "Response variable, percentage of adults reporting fair or poor health",  
 source = "Calculated as Better or Worse")

## The following attribute(s) are being added to a variable in the data frame for the first time: description, source. If you believe this/these attribute(s) were previously added, then check for a typo in the attribute name. If you are adding this/these attribute(s) for the first time, you can probably safely ignore this message.

print(x = codebook(beta\_codebook,   
 title = "Social Determinants of Health Impact on Quality of Life in US Counties Codebook",  
 subtitle = "By: Steven Uzupis, Udumaga Onyeukwu, and Miranda Gemme",  
 description = "Team Beta capstone project for Merrimack College, Masters in Data Science Program, focusing on using Social Determinants of Health to predict quality of life in US Counties.  
  
Research Question: How do social determinants of life affect the quality of life in different localities?  
  
Hypothesis: Social determinants of health, such as economic stability, social connectedness, access to healthcare, and neighborhood environment, significantly predict self-reported health status in US counties.  
  
Prediction: US counties with higher economic security, stronger social support infrastructure, better access to healthcare services, and safer, more accessible neighborhoods will report better overall health status than counties lacking these social determinants.  
  
The predictor variables are obtained from the Social Determinants of Health (SDOH) Database provided by the Agency for Healthcare Research and Quality (AHRQ). The response variable represents 'the percentage of adults reporting fair or poor health.'  
  
To test the research question or hypothesis, the statistical software R is being used."  
), target = "beta\_codebook\_v2.docx")

rm(beta\_codebook)  
rm(beta\_data)

*NA evaluation*

# 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), ]  
  
# View the sorted dataframe  
print(na\_counts\_df)

## variable\_name na\_count  
## hourly\_living\_wage hourly\_living\_wage 3142  
## pb\_measure pb\_measure 3106  
## co\_measure co\_measure 2988  
## no2\_measure no2\_measure 2915  
## so2\_measure so2\_measure 2860  
## pm\_2.5\_measure pm\_2.5\_measure 2616  
## pct\_disconnected\_youth pct\_disconnected\_youth 1938  
## infant\_deaths\_per\_1k\_births infant\_deaths\_per\_1k\_births 1925  
## homicides\_per\_100k homicides\_per\_100k 1816  
## drug\_overdose\_per\_100k drug\_overdose\_per\_100k 1345  
## underage\_deaths\_per\_100k underage\_deaths\_per\_100k 1258  
## juvenile\_arrests\_per\_1k juvenile\_arrests\_per\_1k 1178  
## black\_white\_segregation black\_white\_segregation 1059  
## firearm\_fatalities\_per\_100k firearm\_fatalities\_per\_100k 871  
## pct\_on\_time\_hs\_graduation pct\_on\_time\_hs\_graduation 830  
## suicides\_per\_100k suicides\_per\_100k 709  
## children\_eligible\_for\_lunch children\_eligible\_for\_lunch 577  
## pct\_hiv pct\_hiv 459  
## children\_math\_score children\_math\_score 448  
## children\_reading\_score children\_reading\_score 361  
## successful\_discharge\_rate successful\_discharge\_rate 345  
## rehospitalization\_rate rehospitalization\_rate 323  
## school\_segregation school\_segregation 232  
## mental\_health\_providers\_per\_100k mental\_health\_providers\_per\_100k 201  
## teen\_births\_prk\_1k teen\_births\_prk\_1k 189  
## traffic\_per\_meter traffic\_per\_meter 153  
## pcp\_pt pcp\_pt 147  
## pct\_low\_birthweight pct\_low\_birthweight 106  
## injury\_death\_rate\_per\_100k injury\_death\_rate\_per\_100k 105  
## hospital\_stay\_per\_100k hospital\_stay\_per\_100k 71  
## life\_expectancy\_years life\_expectancy\_years 70  
## pct\_access\_to\_exercise pct\_access\_to\_exercise 62  
## premature\_deaths\_per\_100k premature\_deaths\_per\_100k 60  
## school\_funding\_gap school\_funding\_gap 59  
## water\_quality water\_quality 43  
## percent\_grandparents\_as\_guardians percent\_grandparents\_as\_guardians 40  
## days\_over\_90\_f days\_over\_90\_f 34  
## food\_enviroment food\_enviroment 33  
## pct\_voters pct\_voters 30  
## air\_polution\_metric air\_polution\_metric 27  
## medicare\_fqrc\_rhc\_payment medicare\_fqrc\_rhc\_payment 25  
## pct\_elderly\_mmmograms pct\_elderly\_mmmograms 21  
## pct\_flu\_vaccines\_billed pct\_flu\_vaccines\_billed 18  
## median\_pediatric\_icu\_dist median\_pediatric\_icu\_dist 13  
## median\_rent median\_rent 11  
## medicare\_inpatient\_payment medicare\_inpatient\_payment 8  
## medicare\_acute\_care\_payment medicare\_acute\_care\_payment 7  
## inequality\_ratio inequality\_ratio 7  
## women\_to\_man\_pay\_ratio women\_to\_man\_pay\_ratio 7  
## pct\_rural\_population pct\_rural\_population 7  
## median\_home\_value median\_home\_value 6  
## medicare\_outpatient\_payment medicare\_outpatient\_payment 3  
## medicare\_e&m\_payment medicare\_e&m\_payment 3  
## pct\_single\_parent pct\_single\_parent 2  
## median\_hh\_income.x median\_hh\_income.x 2  
## pct\_people\_in\_poverty pct\_people\_in\_poverty 2  
## median\_surgery\_dist median\_surgery\_dist 2  
## median\_trauma\_center\_dist median\_trauma\_center\_dist 2  
## pct\_poor\_to\_fair\_health pct\_poor\_to\_fair\_health 2  
## poor\_mental\_health poor\_mental\_health 2  
## pct\_adult\_smokers pct\_adult\_smokers 2  
## pct\_obese\_adults pct\_obese\_adults 2  
## pct\_no\_exercise pct\_no\_exercise 2  
## pct\_binge\_drinkers pct\_binge\_drinkers 2  
## pct\_poor\_health pct\_poor\_health 2  
## pct\_insufficieficient\_sleep pct\_insufficieficient\_sleep 2  
## median\_hh\_income.y median\_hh\_income.y 2  
## pct\_income\_to\_childcare pct\_income\_to\_childcare 2  
## response response 2  
## weighted\_population weighted\_population 1  
## average\_hh\_size average\_hh\_size 1  
## pct\_male pct\_male 1  
## pct\_female pct\_female 1  
## pct\_not\_citizens pct\_not\_citizens 1  
## pct\_naturalized\_citizens pct\_naturalized\_citizens 1  
## pct\_adult\_citizens pct\_adult\_citizens 1  
## pct\_no\_english\_spoken pct\_no\_english\_spoken 1  
## pct\_native\_american pct\_native\_american 1  
## pct\_asian pct\_asian 1  
## pct\_black pct\_black 1  
## pct\_hispanic pct\_hispanic 1  
## pct\_other\_race pct\_other\_race 1  
## pct\_white pct\_white 1  
## pct\_hh\_no\_computing\_device pct\_hh\_no\_computing\_device 1  
## pct\_hh\_smartphone pct\_hh\_smartphone 1  
## pct\_hh\_tablet pct\_hh\_tablet 1  
## pct\_hh\_computer pct\_hh\_computer 1  
## pct\_hh\_other\_computer pct\_hh\_other\_computer 1  
## pct\_hh\_internet pct\_hh\_internet 1  
## pct\_hh\_broadband pct\_hh\_broadband 1  
## pct\_hh\_cell\_data pct\_hh\_cell\_data 1  
## pct\_hh\_no\_internet pct\_hh\_no\_internet 1  
## pct\_hh\_satellite pct\_hh\_satellite 1  
## pct\_hh\_dial\_up pct\_hh\_dial\_up 1  
## pct\_employed\_admin pct\_employed\_admin 1  
## pct\_employed\_arts pct\_employed\_arts 1  
## pct\_employed\_construction pct\_employed\_construction 1  
## pct\_employed\_education pct\_employed\_education 1  
## pct\_employed\_finance pct\_employed\_finance 1  
## pct\_employed\_government pct\_employed\_government 1  
## pct\_employed\_information pct\_employed\_information 1  
## pct\_employed\_manufacturing pct\_employed\_manufacturing 1  
## pct\_employed\_nature pct\_employed\_nature 1  
## pct\_employed\_other pct\_employed\_other 1  
## pct\_employed\_professional pct\_employed\_professional 1  
## pct\_employed\_nonprofit pct\_employed\_nonprofit 1  
## pct\_employed\_retail pct\_employed\_retail 1  
## pct\_employed\_transportation pct\_employed\_transportation 1  
## pct\_employed\_wholesale pct\_employed\_wholesale 1  
## pct\_employed pct\_employed 1  
## pct\_unemployed.x pct\_unemployed.x 1  
## gini\_index gini\_index 1  
## pct\_hh\_inc\_10,000 pct\_hh\_inc\_10,000 1  
## pct\_hh\_inc\_100,000 pct\_hh\_inc\_100,000 1  
## pct\_hh\_inc\_14,999 pct\_hh\_inc\_14,999 1  
## pct\_hh\_inc\_24,999 pct\_hh\_inc\_24,999 1  
## pct\_hh\_inc\_49,999 pct\_hh\_inc\_49,999 1  
## pct\_hh\_inc\_99999 pct\_hh\_inc\_99999 1  
## per\_capita\_income per\_capita\_income 1  
## pct\_houses\_vacant pct\_houses\_vacant 1  
## pct\_15\_min\_commute pct\_15\_min\_commute 1  
## pct\_29\_min\_commute pct\_29\_min\_commute 1  
## pct\_59\_min\_commute pct\_59\_min\_commute 1  
## pct\_60\_min\_plus\_commute pct\_60\_min\_plus\_commute 1  
## pct\_public\_transportatin pct\_public\_transportatin 1  
## pct\_w\_medicaid pct\_w\_medicaid 1  
## pct\_w\_medicare pct\_w\_medicare 1  
## land\_area\_sqm land\_area\_sqm 1  
## population\_density population\_density 1  
## median\_er\_dist median\_er\_dist 1  
## median\_obstetrics\_dist median\_obstetrics\_dist 1  
## median\_health\_clinic\_dist median\_health\_clinic\_dist 1  
## median\_drug\_alcohol\_care\_dist median\_drug\_alcohol\_care\_dist 1  
## pct\_under\_65\_no\_health\_insurance pct\_under\_65\_no\_health\_insurance 1  
## pct\_unemployed.y pct\_unemployed.y 1  
## pct\_adult\_poverty pct\_adult\_poverty 1  
## fips\_code fips\_code 0  
## state state 0  
## county county 0  
## region region 0  
## adv\_practice\_nurse\_pt adv\_practice\_nurse\_pt 0  
## clinical\_nurse\_pt clinical\_nurse\_pt 0  
## dentist\_pt dentist\_pt 0  
## anesthetist\_nurse\_pt anesthetist\_nurse\_pt 0  
## midwife\_pt midwife\_pt 0  
## nurse\_practitioner\_pt nurse\_practitioner\_pt 0  
## pa\_pt pa\_pt 0  
## syringe\_exchange\_pt syringe\_exchange\_pt 0  
## substance\_abuse\_facility\_pt substance\_abuse\_facility\_pt 0  
## mental\_health\_faciliy\_pt mental\_health\_faciliy\_pt 0  
## pct\_highschool\_diploma pct\_highschool\_diploma 0  
## pct\_some\_college pct\_some\_college 0  
## social\_clubs\_per\_10k social\_clubs\_per\_10k 0  
## pct\_high\_housing\_costs pct\_high\_housing\_costs 0  
## pct\_overcrowded\_hh pct\_overcrowded\_hh 0  
## pct\_no\_kitchen\_or\_plumbing pct\_no\_kitchen\_or\_plumbing 0  
## pct\_30\_min\_plus\_commute pct\_30\_min\_plus\_commute 0  
## pct\_food\_insecurities pct\_food\_insecurities 0  
## pct\_home\_owner pct\_home\_owner 0  
## pct\_0\_17\_age pct\_0\_17\_age 0  
## pct\_65\_plus pct\_65\_plus 0

# Due to large amount of NA values in these observations, the following will beremoved 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()  
  
# ideal number of observation, roughly square root of # of observations  
sqrt(nrow(qol\_data))

## [1] 54.35991

# clean workspace  
#rm(na\_counts\_df, na\_counts)

*Correlation evaluation*

# 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)  
  
# Find the indices of correlations greater than 0.7  
high\_cor\_indices <- which(abs(cor\_matrix) > 0.7, arr.ind = TRUE)  
  
# 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))), ]  
  
print(high\_cor\_pairs)

## var1 var2 correlation  
## 3 pct\_home\_owner weighted\_population 0.9768797  
## 6 pct\_female pct\_male -1.0000000  
## 10 pct\_naturalized\_citizens pct\_not\_citizens 0.7460535  
## 11 pct\_adult\_citizens pct\_not\_citizens -0.7604839  
## 12 pct\_no\_english\_spoken pct\_not\_citizens 0.8039729  
## 13 pct\_hispanic pct\_not\_citizens 0.7112025  
## 16 pct\_asian pct\_naturalized\_citizens 0.7205063  
## 19 pct\_0\_17\_age pct\_adult\_citizens -0.7853593  
## 22 pct\_hispanic pct\_no\_english\_spoken 0.7028031  
## 27 pct\_white pct\_black -0.8307521  
## 36 pct\_hh\_smartphone pct\_hh\_no\_computing\_device -0.8745077  
## 37 pct\_hh\_tablet pct\_hh\_no\_computing\_device -0.8329862  
## 38 pct\_hh\_computer pct\_hh\_no\_computing\_device -0.8579791  
## 39 pct\_hh\_internet pct\_hh\_no\_computing\_device -0.8829880  
## 40 pct\_hh\_broadband pct\_hh\_no\_computing\_device -0.8841641  
## 41 pct\_hh\_cell\_data pct\_hh\_no\_computing\_device -0.7829934  
## 42 pct\_hh\_no\_internet pct\_hh\_no\_computing\_device 0.9160206  
## 45 pct\_hh\_tablet pct\_hh\_smartphone 0.7294597  
## 46 pct\_hh\_internet pct\_hh\_smartphone 0.7386430  
## 47 pct\_hh\_broadband pct\_hh\_smartphone 0.7491716  
## 48 pct\_hh\_cell\_data pct\_hh\_smartphone 0.7969493  
## 49 pct\_hh\_no\_internet pct\_hh\_smartphone -0.7612280  
## 53 pct\_hh\_computer pct\_hh\_tablet 0.8641927  
## 54 pct\_hh\_internet pct\_hh\_tablet 0.8455591  
## 55 pct\_hh\_broadband pct\_hh\_tablet 0.8460944  
## 56 pct\_hh\_cell\_data pct\_hh\_tablet 0.7637942  
## 57 pct\_hh\_no\_internet pct\_hh\_tablet -0.8623915  
## 58 pct\_hh\_inc\_100,000 pct\_hh\_tablet 0.7582842  
## 59 median\_hh\_income.x pct\_hh\_tablet 0.7668542  
## 60 pct\_adult\_poverty pct\_hh\_tablet -0.7047426  
## 64 pct\_hh\_internet pct\_hh\_computer 0.8636050  
## 65 pct\_hh\_broadband pct\_hh\_computer 0.8596444  
## 66 pct\_hh\_cell\_data pct\_hh\_computer 0.7091640  
## 67 pct\_hh\_no\_internet pct\_hh\_computer -0.8930643  
## 68 pct\_hh\_inc\_100,000 pct\_hh\_computer 0.7442133  
## 69 per\_capita\_income pct\_hh\_computer 0.7480006  
## 70 median\_hh\_income.x pct\_hh\_computer 0.7642493  
## 71 pct\_people\_in\_poverty pct\_hh\_computer -0.7301866  
## 72 pct\_poor\_to\_fair\_health pct\_hh\_computer -0.7925179  
## 73 pct\_adult\_smokers pct\_hh\_computer -0.7288549  
## 74 pct\_no\_exercise pct\_hh\_computer -0.8008113  
## 75 pct\_some\_college pct\_hh\_computer 0.7203738  
## 76 pct\_adult\_poverty pct\_hh\_computer -0.7708390  
## 83 pct\_hh\_broadband pct\_hh\_internet 0.9982100  
## 84 pct\_hh\_cell\_data pct\_hh\_internet 0.8652018  
## 85 pct\_hh\_no\_internet pct\_hh\_internet -0.9720126  
## 86 median\_hh\_income.x pct\_hh\_internet 0.7107642  
## 87 pct\_adult\_poverty pct\_hh\_internet -0.7112662  
## 94 pct\_hh\_cell\_data pct\_hh\_broadband 0.8720069  
## 95 pct\_hh\_no\_internet pct\_hh\_broadband -0.9704187  
## 96 median\_hh\_income.x pct\_hh\_broadband 0.7111758  
## 97 pct\_adult\_poverty pct\_hh\_broadband -0.7042475  
## 105 pct\_hh\_no\_internet pct\_hh\_cell\_data -0.8380513  
## 114 median\_hh\_income.x pct\_hh\_no\_internet -0.7109353  
## 115 pct\_adult\_poverty pct\_hh\_no\_internet 0.7216922  
## 119 pct\_employed\_government pct\_employed\_admin 0.7410970  
## 131 median\_rent pct\_employed\_professional 0.7117222  
## 137 pct\_unemployed.x pct\_employed -1.0000000  
## 142 pct\_people\_in\_poverty pct\_hh\_inc\_10,000 0.8218685  
## 143 pct\_poor\_to\_fair\_health pct\_hh\_inc\_10,000 0.7349967  
## 144 pct\_adult\_poverty pct\_hh\_inc\_10,000 0.7782551  
## 145 inequality\_ratio pct\_hh\_inc\_10,000 0.7372305  
## 146 pct\_food\_insecurities pct\_hh\_inc\_10,000 0.7375304  
## 150 pct\_hh\_inc\_24,999 pct\_hh\_inc\_100,000 -0.7620918  
## 151 pct\_hh\_inc\_49,999 pct\_hh\_inc\_100,000 -0.7922814  
## 152 per\_capita\_income pct\_hh\_inc\_100,000 0.8745999  
## 153 median\_home\_value pct\_hh\_inc\_100,000 0.7586850  
## 154 median\_rent pct\_hh\_inc\_100,000 0.7942470  
## 155 median\_hh\_income.x pct\_hh\_inc\_100,000 0.9315370  
## 156 pct\_adult\_smokers pct\_hh\_inc\_100,000 -0.7364367  
## 158 median\_hh\_income.x pct\_hh\_inc\_14,999 -0.7039120  
## 159 pct\_people\_in\_poverty pct\_hh\_inc\_14,999 0.7025136  
## 160 pct\_adult\_poverty pct\_hh\_inc\_14,999 0.7047035  
## 163 median\_hh\_income.x pct\_hh\_inc\_24,999 -0.7697360  
## 164 pct\_adult\_poverty pct\_hh\_inc\_24,999 0.7091168  
## 167 median\_hh\_income.x pct\_hh\_inc\_49,999 -0.7282559  
## 172 median\_home\_value per\_capita\_income 0.7445663  
## 173 median\_rent per\_capita\_income 0.7042592  
## 174 median\_hh\_income.x per\_capita\_income 0.8430250  
## 175 pct\_people\_in\_poverty per\_capita\_income -0.7147403  
## 176 pct\_poor\_to\_fair\_health per\_capita\_income -0.7507918  
## 177 pct\_adult\_smokers per\_capita\_income -0.7274287  
## 178 pct\_no\_exercise per\_capita\_income -0.7372847  
## 179 pct\_some\_college per\_capita\_income 0.7004612  
## 183 median\_rent median\_home\_value 0.8494514  
## 184 median\_hh\_income.x median\_home\_value 0.7430840  
## 190 median\_hh\_income.x median\_rent 0.7676485  
## 193 population\_density pct\_public\_transportatin 0.7390805  
## 195 pct\_people\_in\_poverty pct\_w\_medicaid 0.7501432  
## 196 pct\_poor\_to\_fair\_health pct\_w\_medicaid 0.7126802  
## 197 pct\_adult\_poverty pct\_w\_medicaid 0.7579946  
## 198 pct\_food\_insecurities pct\_w\_medicaid 0.7029890  
## 201 anesthetist\_nurse\_pt adv\_practice\_nurse\_pt 0.7561631  
## 202 nurse\_practitioner\_pt adv\_practice\_nurse\_pt 0.9596472  
## 231 pct\_people\_in\_poverty median\_hh\_income.x -0.7771547  
## 232 pct\_adult\_smokers median\_hh\_income.x -0.7401780  
## 233 pct\_adult\_poverty median\_hh\_income.x -0.7728660  
## 241 pct\_poor\_to\_fair\_health pct\_people\_in\_poverty 0.8297380  
## 242 pct\_no\_exercise pct\_people\_in\_poverty 0.7650908  
## 243 pct\_adult\_poverty pct\_people\_in\_poverty 0.9133787  
## 244 pct\_food\_insecurities pct\_people\_in\_poverty 0.7973447  
## 259 pct\_adult\_smokers pct\_poor\_to\_fair\_health 0.7272550  
## 260 pct\_obese\_adults pct\_poor\_to\_fair\_health 0.7220514  
## 261 pct\_no\_exercise pct\_poor\_to\_fair\_health 0.9262070  
## 262 pct\_highschool\_diploma pct\_poor\_to\_fair\_health -0.8215580  
## 263 pct\_some\_college pct\_poor\_to\_fair\_health -0.7396619  
## 264 pct\_adult\_poverty pct\_poor\_to\_fair\_health 0.8413465  
## 265 life\_expectancy\_years pct\_poor\_to\_fair\_health -0.7028386  
## 266 pct\_food\_insecurities pct\_poor\_to\_fair\_health 0.8186087  
## 273 pct\_obese\_adults pct\_adult\_smokers 0.7038164  
## 274 pct\_no\_exercise pct\_adult\_smokers 0.7749808  
## 275 life\_expectancy\_years pct\_adult\_smokers -0.7111940  
## 279 pct\_no\_exercise pct\_obese\_adults 0.7870511  
## 287 pct\_highschool\_diploma pct\_no\_exercise -0.7694530  
## 288 pct\_some\_college pct\_no\_exercise -0.7258065  
## 289 pct\_adult\_poverty pct\_no\_exercise 0.7775970  
## 290 life\_expectancy\_years pct\_no\_exercise -0.7102781  
## 291 pct\_food\_insecurities pct\_no\_exercise 0.7098093  
## 297 pct\_some\_college pct\_highschool\_diploma 0.7441917  
## 318 pct\_food\_insecurities pct\_adult\_poverty 0.8111507

# clean workspace  
#rm(  
# qol\_numeric,  
# cor\_matrix,  
# high\_cor\_indices  
# )  
  
# 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, # keep for EDA  
 -pct\_food\_insecurities,  
 -pct\_hh\_computer,  
 -pct\_naturalized\_citizens,  
 -median\_rent,  
 -median\_home\_value,  
 -pct\_w\_medicaid  
 )

#rm(high\_cor\_pairs)

*Near Zero Variance evaluation*

qol\_numeric <- qol\_data %>%  
 select(  
 -fips\_code,  
 -state,  
 -county,  
 -region  
 )  
  
# identify predictors with NZV (near-zero variance, high collinearity)  
nzv <- nearZeroVar(qol\_numeric, saveMetrics = TRUE)  
  
print(nzv[which(nzv$nzv == TRUE), ])

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

#rm(qol\_numeric)

qol\_data <- qol\_data %>%  
 select(-syringe\_exchange\_pt  
 )

#rm(nzv)

*Varience Inflation Factor evaluation*

# create linear model for vif analysis, no scaling, no centering, numeric values only  
qol\_numeric <- qol\_data %>%   
 select(where(is.numeric))  
  
qol\_lm <- lm(  
 pct\_poor\_to\_fair\_health ~ .,  
 qol\_numeric  
)  
  
vif\_values <- car::vif(qol\_lm)  
  
vif\_values

## fips\_code weighted\_population   
## 1.413956e+00 2.937561e+01   
## average\_hh\_size pct\_male   
## 4.230557e+00 1.991965e+00   
## pct\_adult\_citizens pct\_no\_english\_spoken   
## 8.682471e+00 3.175002e+00   
## pct\_native\_american pct\_asian   
## 1.961932e+01 4.857908e+00   
## pct\_black pct\_hispanic   
## 7.672287e+01 6.551841e+00   
## pct\_other\_race pct\_white   
## 7.261959e+00 9.855127e+01   
## pct\_single\_parent pct\_hh\_other\_computer   
## 3.020044e+00 1.116598e+00   
## pct\_hh\_internet pct\_hh\_satellite   
## 3.677090e+00 1.468289e+00   
## pct\_hh\_dial\_up pct\_employed\_admin   
## 1.198700e+00 7.924150e+04   
## pct\_employed\_arts pct\_employed\_construction   
## 1.051799e+05 5.553576e+04   
## pct\_employed\_education pct\_employed\_finance   
## 1.922058e+05 3.553441e+04   
## pct\_employed\_government pct\_employed\_information   
## 6.164183e+00 5.418630e+03   
## pct\_employed\_manufacturing pct\_employed\_nature   
## 4.514068e+05 3.566457e+05   
## pct\_employed\_other pct\_employed\_professional   
## 1.632713e+04 9.980477e+04   
## pct\_employed\_nonprofit pct\_employed\_retail   
## 2.476122e+00 5.164249e+04   
## pct\_employed\_transportation pct\_employed\_wholesale   
## 3.983079e+04 1.384352e+04   
## pct\_employed gini\_index   
## 2.089750e+00 2.709078e+00   
## pct\_hh\_inc\_99999 pct\_houses\_vacant   
## 2.507535e+00 2.812388e+00   
## pct\_public\_transportatin pct\_w\_medicare   
## 3.354212e+00 2.341255e+00   
## clinical\_nurse\_pt dentist\_pt   
## 1.222284e+00 1.939791e+00   
## anesthetist\_nurse\_pt midwife\_pt   
## 2.434778e+00 1.318057e+00   
## nurse\_practitioner\_pt pa\_pt   
## 2.101674e+00 2.278364e+00   
## substance\_abuse\_facility\_pt mental\_health\_faciliy\_pt   
## 1.543153e+00 1.469924e+00   
## land\_area\_sqm population\_density   
## 1.697197e+00 2.546185e+00   
## days\_over\_90\_f median\_hh\_income.x   
## 3.746303e+00 6.747624e+00   
## median\_er\_dist median\_surgery\_dist   
## 2.101413e+00 2.254958e+00   
## median\_trauma\_center\_dist median\_pediatric\_icu\_dist   
## 1.464692e+00 1.768941e+00   
## median\_obstetrics\_dist median\_health\_clinic\_dist   
## 2.405678e+00 1.220238e+00   
## median\_drug\_alcohol\_care\_dist percent\_grandparents\_as\_guardians   
## 1.906096e+00 2.127187e+00   
## pct\_adult\_smokers pct\_obese\_adults   
## 8.804114e+00 4.239526e+00   
## pct\_binge\_drinkers pct\_under\_65\_no\_health\_insurance   
## 2.435285e+00 3.928932e+00   
## pct\_highschool\_diploma pct\_some\_college   
## 6.303198e+00 4.701598e+00   
## inequality\_ratio social\_clubs\_per\_10k   
## 2.866709e+00 1.787229e+00   
## air\_polution\_metric water\_quality   
## 2.429196e+00 1.146228e+00   
## pct\_high\_housing\_costs pct\_overcrowded\_hh   
## 2.667037e+00 2.818148e+00   
## pct\_no\_kitchen\_or\_plumbing pct\_30\_min\_plus\_commute   
## 1.221599e+00 3.063413e+00   
## life\_expectancy\_years school\_funding\_gap   
## 3.759726e+00 3.321014e+00   
## pct\_income\_to\_childcare pct\_voters   
## 1.862612e+00 4.166233e+00   
## pct\_home\_owner pct\_65\_plus   
## 3.221387e+01 5.977418e+00   
## pct\_rural\_population   
## 4.758681e+00

# Create a data frame with GVIF and Df  
vif\_df <- data.frame(  
 Feature = names(vif\_values),  
 GVIF = vif\_values,  
 Df = rep(1, length(vif\_values)) # Df is typically 1 for univariate cases  
)  
  
# Calculate GVIF^(1/(2\*Df)) and add it to the data frame  
vif\_df$Adjusted\_VIF <- vif\_df$GVIF^(1/(2 \* vif\_df$Df))  
  
# Sort the data frame by Adjusted VIF in ascending order  
vif\_df <- vif\_df[order(vif\_df$Adjusted\_VIF, decreasing = TRUE), ]  
  
# Set a threshold for high VIF  
high\_vif\_threshold <- 5  
  
# Filter for features with high VIF values  
high\_vif\_features <- vif\_df[vif\_df[, "Adjusted\_VIF"] > high\_vif\_threshold, ]  
  
# Print the table  
# Add a column indicating if the Adjusted VIF is above the threshold  
vif\_df <- high\_vif\_features %>%  
 mutate(High\_VIF = if\_else(Adjusted\_VIF > high\_vif\_threshold, "Yes", "No"))  
  
# Print the table  
vif\_df

## Feature GVIF Df  
## pct\_employed\_manufacturing pct\_employed\_manufacturing 451406.79575 1  
## pct\_employed\_nature pct\_employed\_nature 356645.69262 1  
## pct\_employed\_education pct\_employed\_education 192205.84298 1  
## pct\_employed\_arts pct\_employed\_arts 105179.87896 1  
## pct\_employed\_professional pct\_employed\_professional 99804.77089 1  
## pct\_employed\_admin pct\_employed\_admin 79241.49518 1  
## pct\_employed\_construction pct\_employed\_construction 55535.75977 1  
## pct\_employed\_retail pct\_employed\_retail 51642.48523 1  
## pct\_employed\_transportation pct\_employed\_transportation 39830.79133 1  
## pct\_employed\_finance pct\_employed\_finance 35534.41159 1  
## pct\_employed\_other pct\_employed\_other 16327.12931 1  
## pct\_employed\_wholesale pct\_employed\_wholesale 13843.51829 1  
## pct\_employed\_information pct\_employed\_information 5418.62987 1  
## pct\_white pct\_white 98.55127 1  
## pct\_black pct\_black 76.72287 1  
## pct\_home\_owner pct\_home\_owner 32.21387 1  
## weighted\_population weighted\_population 29.37561 1  
## Adjusted\_VIF High\_VIF  
## pct\_employed\_manufacturing 671.868139 Yes  
## pct\_employed\_nature 597.198202 Yes  
## pct\_employed\_education 438.412868 Yes  
## pct\_employed\_arts 324.314475 Yes  
## pct\_employed\_professional 315.918931 Yes  
## pct\_employed\_admin 281.498659 Yes  
## pct\_employed\_construction 235.660263 Yes  
## pct\_employed\_retail 227.249830 Yes  
## pct\_employed\_transportation 199.576530 Yes  
## pct\_employed\_finance 188.505734 Yes  
## pct\_employed\_other 127.777656 Yes  
## pct\_employed\_wholesale 117.658482 Yes  
## pct\_employed\_information 73.611343 Yes  
## pct\_white 9.927299 Yes  
## pct\_black 8.759159 Yes  
## pct\_home\_owner 5.675727 Yes  
## weighted\_population 5.419927 Yes

qol\_data <- qol\_data %>%   
 select(  
 -pct\_adult\_citizens,  
 -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\_retail,  
 -pct\_employed\_transportation,  
 -pct\_employed\_wholesale,  
 -pct\_65\_plus  
 )

#rm(  
# vif\_values,  
# high\_vif,  
# high\_vif\_threshold,  
# high\_vif\_features,  
# vif\_df  
#)

*Final feature cleaning and testing*

qol\_data <- qol\_data %>%   
 select(  
 -pct\_no\_english\_spoken,  
 -pct\_hh\_satellite,  
 -pct\_hh\_dial\_up,  
 -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\_obese\_adults,  
 -pct\_some\_college,  
 -pct\_no\_kitchen\_or\_plumbing,  
 -pct\_income\_to\_childcare,  
 )

# recreate numerc vector with updated features  
qol\_numeric <- qol\_data %>%   
 select(  
 -fips\_code,  
 -state,  
 -county,  
 -region  
 )  
  
# identify best model for response with regsubset  
qol\_regfit\_full <- leaps::regsubsets(  
 response ~ . - weighted\_population,  
 qol\_numeric,  
 really.big = TRUE,  
 nvmax = 45  
 )  
  
reg\_fit\_summary <- summary(qol\_regfit\_full)

# identifying ideal number of variables  
which.min(reg\_fit\_summary$rss)

## [1] 45

which.max(reg\_fit\_summary$adjr2)

## [1] 29

which.min(reg\_fit\_summary$cp)

## [1] 26

which.min(reg\_fit\_summary$bic)

## [1] 14

plot(qol\_regfit\_full)

# extract the metrics  
rss <- reg\_fit\_summary$rss  
adjr2 <- reg\_fit\_summary$adjr2  
cp <- reg\_fit\_summary$cp  
bic <- reg\_fit\_summary$bic  
  
# ideal value labaled  
min\_rss <- which.min(rss)  
max\_adjr2 <- which.max(adjr2)  
min\_cp <- which.min(cp)  
min\_bic <- which.min(bic)

par(mfrow = c(2, 2))  
  
# plot RSS  
plot(  
 rss,  
 type = "b",  
 pch = 19,  
 col = "blue",  
 ylab = "RSS",  
 xlab = "Number of Variables",  
 main = "RSS vs Number of Variables"  
 )  
points(  
 min\_rss, rss[min\_rss],  
 col = "red",  
 pch = 19,  
 cex = 2  
 )  
  
# plot R^2  
plot(  
 adjr2,  
 type = "b",  
 pch = 19,  
 col = "blue",  
 ylab = "Adjusted R²",  
 xlab = "Number of Variables",  
 main = "Adjusted R² vs Number of Variables"  
 )  
points(  
 max\_adjr2,  
 adjr2[max\_adjr2],  
 col = "red",  
 pch = 19,  
 cex = 2  
 )  
  
# plot Cp  
plot(  
 cp,  
 type = "b",  
 pch = 19,  
 col = "blue",  
 ylab = "Cp",  
 xlab = "Number of Variables",  
 main = "Cp vs Number of Variables"  
 )  
points(min\_cp,  
 cp[min\_cp],  
 col = "red",  
 pch = 19,  
 cex = 2  
 )  
  
# plot BIC  
plot(  
 bic,  
 type = "b",  
 pch = 19,  
 col = "blue",  
 ylab = "BIC",  
 xlab = "Number of Variables",  
 main = "BIC vs Number of Variables"  
 )  
points(  
 min\_bic,  
 bic[min\_bic],  
 col = "red",  
 pch = 19,  
 cex = 2  
 )

# clean enviroment  
#rm(  
# qol\_numeric,  
# qol\_regfit\_full,  
# reg\_fit\_summary,  
# adjr2,  
# bic,  
# cp,  
# max\_adjr2,  
# min\_bic,  
# min\_cp,  
# min\_rss,  
# rss  
#)

*Overall Summary*

glimpse(qol\_data)

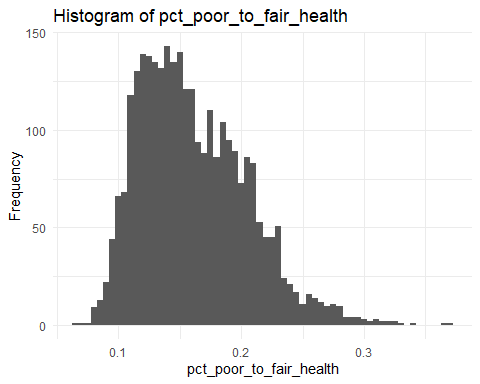
## Rows: 2,955  
## Columns: 51

## Warning in grepl(",", levels(x), fixed = TRUE): input string 2 is invalid UTF-8

## $ fips\_code <dbl> 1001, 1003, 1005, 1007, 1009, 1011, …  
## $ state <fct> Alabama, Alabama, Alabama, Alabama, …  
## $ county <fct> Autauga County, Baldwin County, Barb…  
## $ region <fct> South, South, South, South, South, S…  
## $ weighted\_population <dbl> 55639, 218289, 25026, 22374, 57755, …  
## $ average\_hh\_size <dbl> 2.55, 2.56, 2.37, 2.85, 2.70, 2.84, …  
## $ pct\_male <dbl> 48.62, 48.51, 52.57, 53.73, 49.65, 5…  
## $ pct\_native\_american <dbl> 0.28, 0.69, 0.35, 0.05, 0.10, 0.00, …  
## $ pct\_asian <dbl> 1.17, 0.93, 0.49, 0.25, 0.41, 1.35, …  
## $ pct\_black <dbl> 19.53, 8.77, 47.67, 22.55, 1.40, 68.…  
## $ pct\_hispanic <dbl> 2.88, 4.56, 4.44, 2.68, 9.28, 8.10, …  
## $ pct\_other\_race <dbl> 0.67, 1.56, 3.10, 0.04, 1.80, 3.13, …  
## $ pct\_white <dbl> 75.76, 85.44, 46.30, 76.60, 93.97, 2…  
## $ pct\_single\_parent <dbl> 27.41, 18.10, 52.76, 32.19, 25.75, 5…  
## $ pct\_hh\_other\_computer <dbl> 1.05, 1.75, 2.15, 0.23, 2.16, 1.84, …  
## $ pct\_hh\_internet <dbl> 82.80, 85.52, 65.00, 76.17, 80.03, 6…  
## $ pct\_employed\_nonprofit <dbl> 7.72, 5.68, 4.20, 3.48, 5.30, 6.76, …  
## $ pct\_employed <dbl> 97.09, 96.08, 93.06, 92.56, 94.80, 9…  
## $ pct\_hh\_inc\_99999 <dbl> 30.40, 30.52, 23.63, 34.92, 28.64, 2…  
## $ pct\_w\_medicare <dbl> 6.87, 7.27, 7.51, 7.74, 8.81, 5.87, …  
## $ clinical\_nurse\_pt <dbl> 0.02, 0.02, 0.00, 0.00, 0.00, 0.00, …  
## $ dentist\_pt <dbl> 0.34, 0.49, 0.37, 0.27, 0.19, 0.20, …  
## $ pa\_pt <dbl> 0.04, 0.17, 0.04, 0.23, 0.00, 0.10, …  
## $ mental\_health\_faciliy\_pt <dbl> 0.0178, 0.0174, 0.0813, 0.0000, 0.01…  
## $ population\_density <dbl> 93.60, 137.30, 28.28, 35.94, 89.56, …  
## $ days\_over\_90\_f <dbl> 104, 97, 104, 97, 80, 103, 103, 84, …  
## $ median\_hh\_income.x <dbl> 67565, 71135, 38866, 50907, 55203, 3…  
## $ median\_er\_dist <dbl> 2.25, 6.01, 5.58, 8.44, 10.56, 4.89,…  
## $ median\_trauma\_center\_dist <dbl> 12.07, 25.37, 41.37, 26.40, 26.16, 3…  
## $ median\_pediatric\_icu\_dist <dbl> 55.65, 22.62, 63.76, 21.97, 52.24, 5…  
## $ median\_health\_clinic\_dist <dbl> 9.20, 8.57, 1.17, 3.85, 3.91, 1.63, …  
## $ median\_drug\_alcohol\_care\_dist <dbl> 12.24, 13.21, 35.73, 26.48, 25.22, 4…  
## $ percent\_grandparents\_as\_guardians <dbl> 4.901316, 5.961440, 11.165644, 12.54…  
## $ pct\_poor\_to\_fair\_health <dbl> 0.169, 0.149, 0.275, 0.216, 0.184, 0…  
## $ pct\_adult\_smokers <dbl> 0.183, 0.169, 0.259, 0.228, 0.218, 0…  
## $ pct\_binge\_drinkers <dbl> 0.1665626, 0.1896517, 0.1342877, 0.1…  
## $ pct\_under\_65\_no\_health\_insurance <dbl> 0.1055942, 0.1087488, 0.1436828, 0.1…  
## $ pct\_highschool\_diploma <dbl> 0.8958449, 0.9101416, 0.7567102, 0.8…  
## $ inequality\_ratio <dbl> 4.794400, 4.300710, 5.180597, 5.0349…  
## $ social\_clubs\_per\_10k <dbl> 12.645828, 9.594962, 9.353776, 9.035…  
## $ air\_polution\_metric <dbl> 10.0, 7.6, 9.4, 9.8, 9.6, 9.3, 9.1, …  
## $ water\_quality <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ pct\_high\_housing\_costs <dbl> 0.12635805, 0.10561056, 0.13464346, …  
## $ pct\_overcrowded\_hh <dbl> 0.011217574, 0.012912393, 0.03852327…  
## $ pct\_30\_min\_plus\_commute <dbl> 0.416, 0.376, 0.365, 0.551, 0.595, 0…  
## $ life\_expectancy\_years <dbl> 76.58565, 77.72473, 72.86721, 73.609…  
## $ school\_funding\_gap <dbl> -2077.14200, 343.03810, -13560.45000…  
## $ pct\_voters <dbl> 0.6618208, 0.6529095, 0.5402157, 0.5…  
## $ pct\_home\_owner <dbl> 16227, 67242, 5654, 5580, 16865, 222…  
## $ pct\_rural\_population <dbl> 0.4200216, 0.4227910, 0.6778963, 0.6…  
## $ response <fct> worse, worse, worse, worse, worse, w…

qol\_names <- names(qol\_data)

ggplot(qol\_data, aes(x = pct\_poor\_to\_fair\_health)) +  
 geom\_histogram(binwidth = 0.005) +  
 ggtitle("Histogram of pct\_poor\_to\_fair\_health") +  
 xlab("pct\_poor\_to\_fair\_health") +  
 ylab("Frequency") +  
 theme\_minimal()

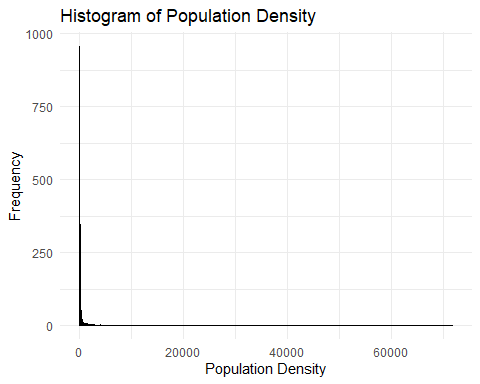


skew\_value <- skewness(  
 qol\_data$pct\_poor\_to\_fair\_health,  
 na.rm = TRUE  
 )  
  
print(skew\_value)

## [1] 0.731473

# data is right skewed

# Create histogram plot of population density  
ggplot(data = qol\_data, aes(x = population\_density)) +  
 geom\_histogram(binwidth = 50, fill = "skyblue", color = "black") +  
 labs(title = "Histogram of Population Density",  
 x = "Population Density",  
 y = "Frequency") +  
 theme\_minimal()



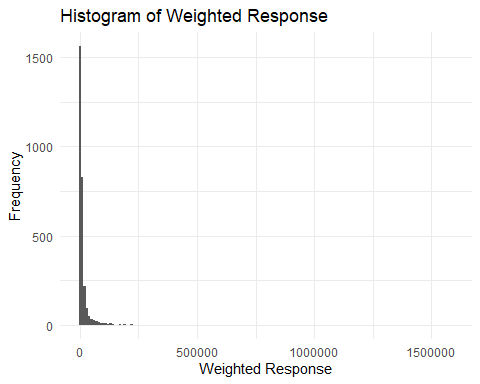
summary(qol\_data$population\_density)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.44 18.76 45.70 213.44 115.72 71895.54

skew\_value <- skewness(  
 qol\_data$population\_density,  
 na.rm = TRUE  
 )  
  
print(skew\_value)

## [1] 39.54178

response\_data <- qol\_data %>%   
 select(  
 pct\_poor\_to\_fair\_health,  
 weighted\_population  
 ) %>%   
 mutate(  
 weighted\_response = pct\_poor\_to\_fair\_health \* weighted\_population  
 )  
  
ggplot(response\_data, aes(x = weighted\_response)) +  
 geom\_histogram(binwidth = 10000) +  
 ggtitle("Histogram of Weighted Response") +  
 xlab("Weighted Response") +  
 ylab("Frequency") +  
 theme\_minimal()



skew\_value <- skewness(  
 response\_data$weighted\_response,  
 na.rm = TRUE  
 )  
  
print(skew\_value)

## [1] 15.47122

# data is still right skewed, failed to correct for skewness

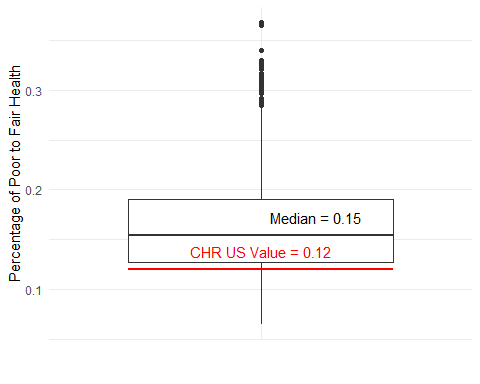
#rm(  
# skew\_value,  
# response\_data  
#)

median\_value <- median(qol\_data$pct\_poor\_to\_fair\_health, na.rm = TRUE)  
  
median\_value

## [1] 0.154

ggplot(qol\_data,  
 aes(x = "",  
 y = pct\_poor\_to\_fair\_health)) +  
 geom\_boxplot() +  
 geom\_segment(  
 aes(x = 0.625,  
 xend = 1.375,  
 y = 0.12,  
 yend = 0.12  
 ),  
 color = "red",  
 size = 1  
 ) +  
 annotate(  
 "text",  
 x = 1,  
 y = 0.12,  
 label = "CHR US Value = 0.12",  
 color = "red",  
 vjust = -1  
 ) +  
 annotate(  
 "text",  
 x = 1,  
 y = median\_value,  
 label = paste(  
 "Median =",  
 round(median\_value, 2)  
 ),  
 color = "black",  
 vjust = -1,  
 hjust = -0.1  
 ) +  
 labs(  
 y = "Percentage of Poor to Fair Health",  
 x = "") +  
 theme\_minimal()

## 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.



# use new median value as it is representative of data in model and not the total us value

qol\_data <- qol\_data %>%   
 mutate(response = ifelse(  
 pct\_poor\_to\_fair\_health >= median\_value,  
 "worse",  
 "better")  
 )

*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))

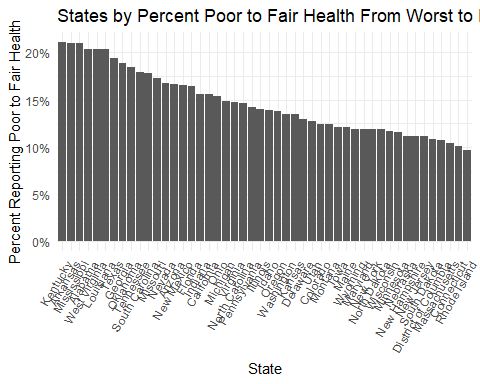
# check for NA  
sum(is.na(qol\_data$weighted\_population))

## [1] 0

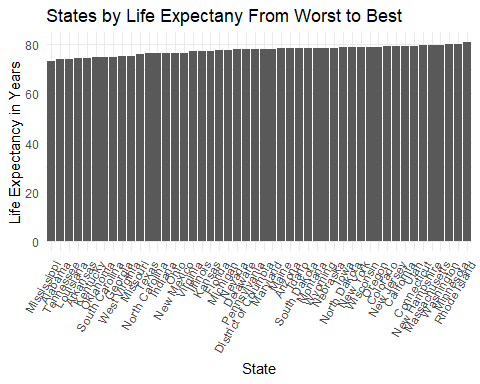
# list of race variables  
race\_vars <- c(  
 "pct\_native\_american",  
 "pct\_asian",  
 "pct\_black",  
 "pct\_hispanic",  
 "pct\_other\_race",  
 "pct\_white"  
 )  
  
# create longer table with race as a column  
qol\_data\_long <- qol\_data %>%  
 pivot\_longer(  
 cols = all\_of(race\_vars),  
 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  
 )) %>%   
 select(  
 race,  
 region,  
 percentage,  
 pct\_poor\_to\_fair\_health,  
 life\_expectancy\_years,  
 pct\_voters,  
 weighted\_population  
 )  
  
# 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")

*Explore individual features*

# Reorder states by median pct\_poor\_to\_fair\_health  
qol\_state\_median <- qol\_state\_median %>%  
 mutate(state = forcats::fct\_reorder(  
 state,  
 pct\_poor\_to\_fair\_health,  
 .desc = TRUE))  
  
# Create the bar plot with vertically rotated labels and percent y-axis  
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))



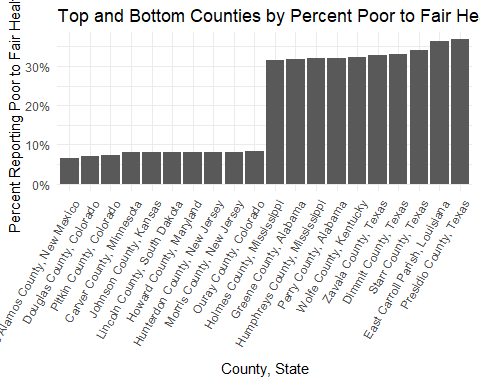
# Reorder states by median pct\_poor\_to\_fair\_health  
qol\_state\_median <- qol\_state\_median %>%  
 mutate(state = forcats::fct\_reorder(  
 state,  
 life\_expectancy\_years,  
 .desc = FALSE))  
  
# Create the bar plot with vertically rotated labels and percent y-axis  
ggplot(data = qol\_state\_median,  
 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))



#rm(  
# qol\_state\_median,  
# qol\_state\_mean  
# )

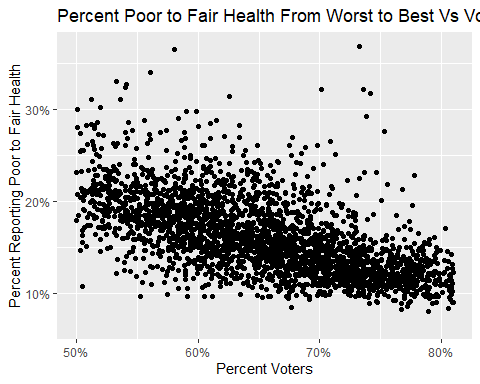
# Remove rows with missing values in pct\_poor\_to\_fair\_health  
#qol\_data\_clean <- qol\_data %>%  
# filter(!is.na(pct\_poor\_to\_fair\_health))  
  
# Arrange qol\_data by pct\_poor\_to\_fair\_health  
qol\_data\_sorted <- qol\_data %>%  
 arrange(pct\_poor\_to\_fair\_health)  
  
# Select top 10 and bottom 10 counties  
top\_bottom\_counties <- qol\_data\_sorted %>%  
 dplyr::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))



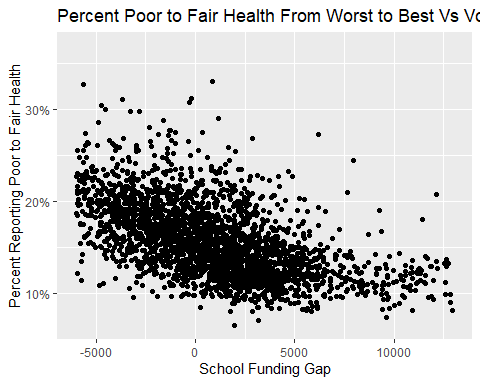
# Create the scatter plot with x-axis limit, percentage formatting, and state labels  
ggplot(data = qol\_data,  
 aes(x = pct\_voters, y = pct\_poor\_to\_fair\_health)) +  
 geom\_point() +  
 #geom\_text(aes(label = state), hjust = 1.2, vjust = 0.5, size = 2) + # Add state labels  
 scale\_y\_continuous(labels = 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()`).

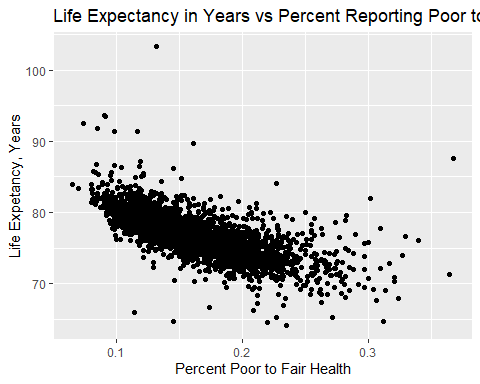


# Create the scatter plot with x-axis limit, percentage formatting, and state labels  
ggplot(data = qol\_data,  
 aes(x = school\_funding\_gap, y = pct\_poor\_to\_fair\_health)) +  
 geom\_point() +  
 # geom\_text(aes(label = state), hjust = 1.2, vjust = 0.5, size = 2) + # Add state labels  
 scale\_y\_continuous(labels = 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()`).

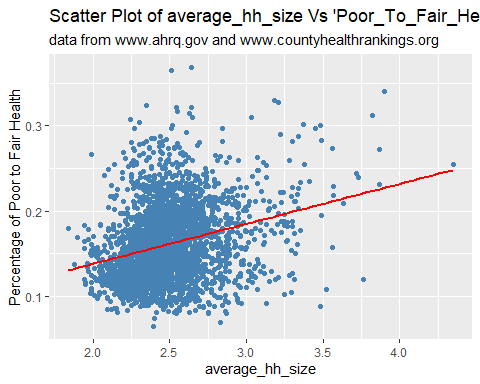
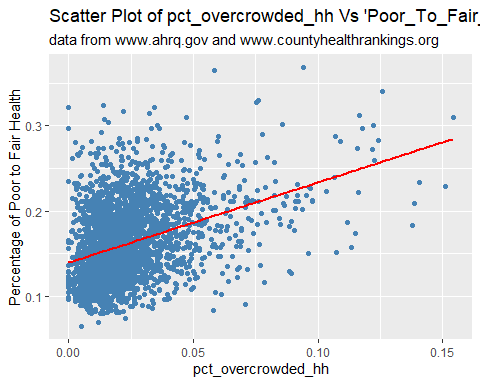
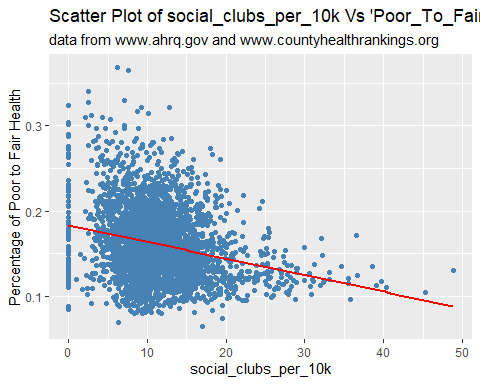
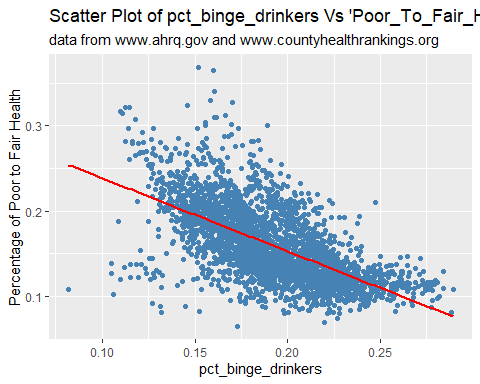
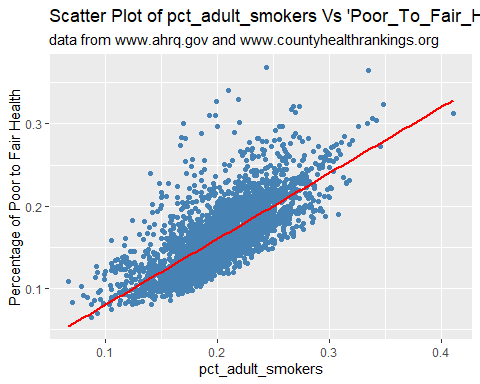
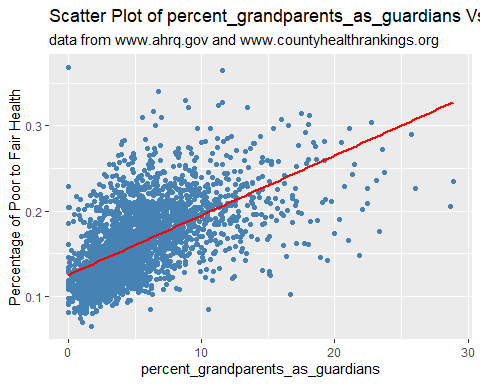
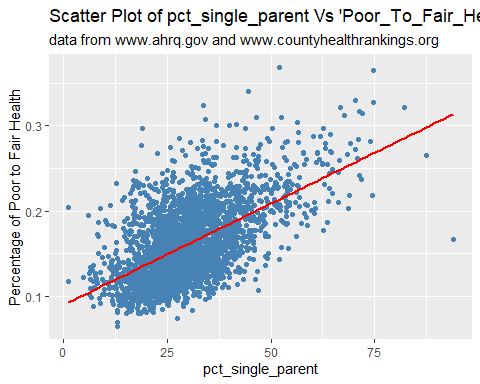


# Create the scatter plot with x-axis limit, percentage formatting, and state labels  
ggplot(data = qol\_data,  
 aes(x = pct\_poor\_to\_fair\_health, y = life\_expectancy\_years)) +  
 geom\_point() +  
 # scale\_y\_continuous() +  
 # scale\_x\_continuous(labels = percent\_format(scale = 100), limits = 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\_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")  
 )  
}

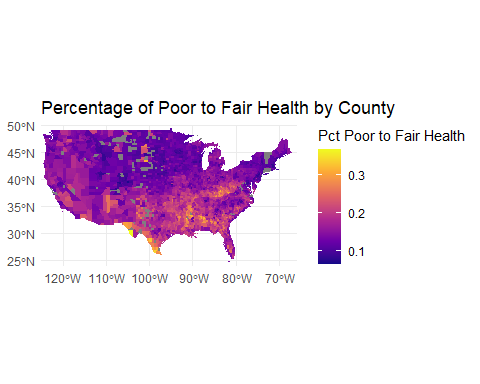


#Exploring the relationship between median household income and health status in US counties along geographical lines by creating heat maps. UO  
  
# Load US counties shapefile  
counties <- tigris::counties(cb = TRUE)

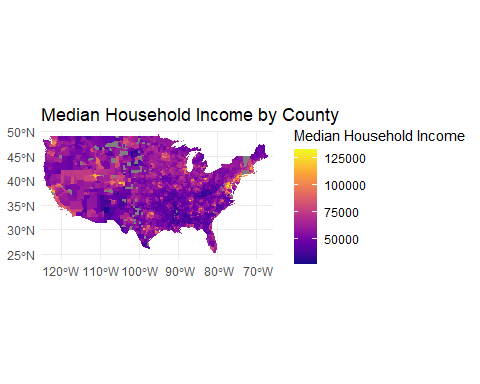
## Retrieving data for the year 2022

## | | | 0% | | | 1% | |= | 1% | |== | 2% | |== | 3% | |=== | 4% | |=== | 5% | |==== | 5% | |==== | 6% | |===== | 6% | |===== | 7% | |===== | 8% | |====== | 8% | |====== | 9% | |======= | 10% | |======== | 11% | |========= | 12% | |========= | 13% | |========== | 14% | |========== | 15% | |=========== | 16% | |============ | 17% | |============ | 18% | |============= | 18% | |============= | 19% | |============== | 20% | |============== | 21% | |=============== | 21% | |=============== | 22% | |================ | 23% | |================= | 24% | |================= | 25% | |================== | 25% | |================== | 26% | |=================== | 27% | |=================== | 28% | |==================== | 28% | |==================== | 29% | |===================== | 29% | |===================== | 30% | |====================== | 31% | |====================== | 32% | |======================= | 32% | |======================= | 33% | |======================== | 34% | |======================== | 35% | |========================= | 35% | |========================= | 36% | |========================== | 37% | |=========================== | 39% | |============================ | 39% | |============================ | 40% | |============================ | 41% | |============================= | 41% | |============================= | 42% | |============================== | 43% | |============================== | 44% | |=============================== | 44% | |================================ | 45% | |================================ | 46% | |================================= | 48% | |================================== | 48% | |================================== | 49% | |=================================== | 49% | |=================================== | 50% | |==================================== | 51% | |==================================== | 52% | |===================================== | 53% | |====================================== | 54% | |========================================= | 58% | |========================================= | 59% | |========================================== | 60% | |=========================================== | 61% | |=========================================== | 62% | |============================================ | 63% | |============================================ | 64% | |============================================= | 64% | |============================================== | 66% | |=============================================== | 67% | |=============================================== | 68% | |================================================ | 68% | |================================================ | 69% | |================================================= | 70% | |================================================== | 71% | |================================================== | 72% | |=================================================== | 72% | |=================================================== | 73% | |==================================================== | 74% | |==================================================== | 75% | |===================================================== | 75% | |===================================================== | 76% | |====================================================== | 77% | |====================================================== | 78% | |======================================================= | 78% | |======================================================= | 79% | |======================================================== | 80% | |========================================================= | 81% | |========================================================== | 82% | |========================================================== | 83% | |=========================================================== | 85% | |============================================================ | 85% | |============================================================ | 86% | |============================================================= | 87% | |============================================================= | 88% | |============================================================== | 88% | |============================================================== | 89% | |================================================================ | 91% | |================================================================ | 92% | |================================================================= | 92% | |================================================================== | 94% | |=================================================================== | 96% | |==================================================================== | 96% | |==================================================================== | 97% | |==================================================================== | 98% | |===================================================================== | 98% | |======================================================================| 100%

# add 0's to fips code  
qol\_data$fips\_code <- sprintf("%05d", qol\_data$fips\_code)  
  
  
# 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()

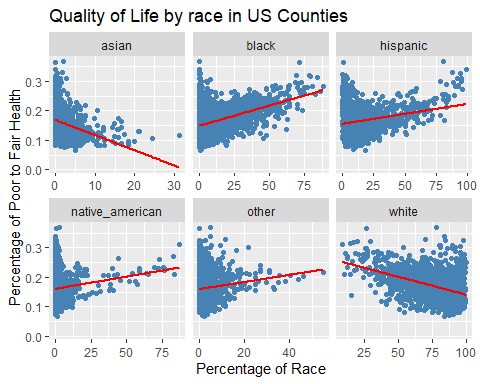


# Create map for Median Household income  
ggplot(data = counties) +  
 geom\_sf(aes(fill = median\_hh\_income.x), color = NA) +  
 scale\_fill\_viridis\_c(option = "plasma") +  
 labs(title = "Median Household Income by County",  
 fill = "Median Household Income") +  
 coord\_sf(xlim = c(-125, -66),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



# Create 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 = "Quality of Life by race 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"  
 )  
  
sum(is.na(data\_mg))

## [1] 0

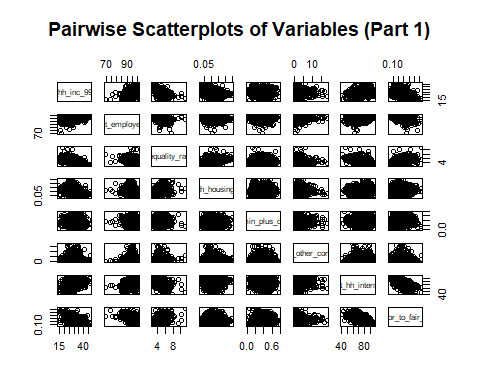
summary(data\_mg)

## 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

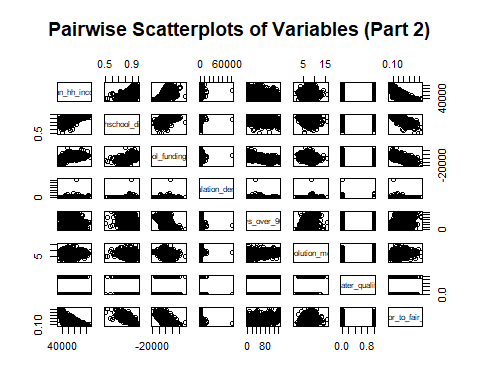
variables <- c(  
 "pct\_hh\_inc\_99999",  
 "pct\_employed",  
 "inequality\_ratio",  
 "pct\_high\_housing\_costs",  
 "pct\_30\_min\_plus\_commute",  
 "pct\_hh\_other\_computer",  
 "pct\_hh\_internet",  
 "median\_hh\_income.x",  
 "pct\_highschool\_diploma",  
 "school\_funding\_gap",  
 "population\_density",  
 "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)")

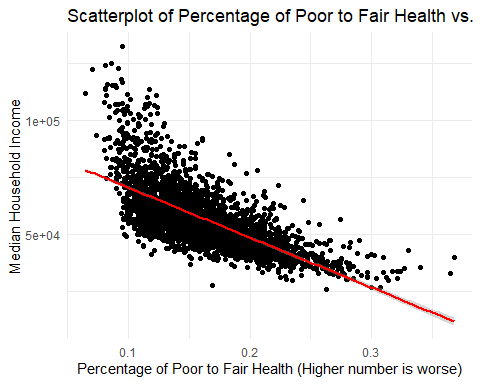


pairs(data\_mg\_part2, main = "Pairwise Scatterplots of Variables (Part 2)")



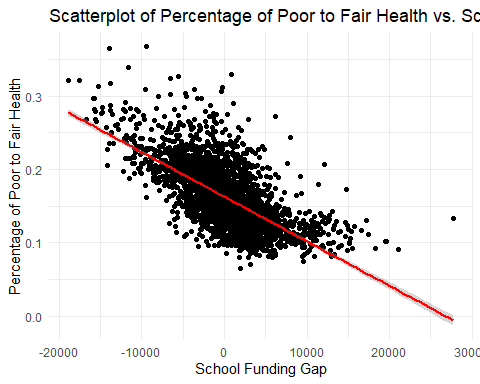
ggplot(data\_mg, aes(x = pct\_poor\_to\_fair\_health, y = median\_hh\_income.x)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", col = "red") +  
 labs(title = "Scatterplot of Percentage of Poor to Fair Health vs. Median Household Income",  
 x = "Percentage of Poor to Fair Health (Higher number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

# list of races  
df\_race\_list <- c("asian", "black", "hispanic", "native\_american", "other", "white")  
  
# list of features to evaluate  
feature\_list <- c("pct\_poor\_to\_fair\_health", "life\_expectancy\_years", "pct\_voters")  
  
# function  
race\_stats\_calc <- function(data, race\_list, feature\_list) {  
 # create list to fill  
 results <- list()  
 # iterate over race 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)  
}

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

## # 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 the summary table  
summary\_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  
summary\_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) | | | | | | |

# explore individual race variables  
summary\_table\_black <- 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  
summary\_table\_black

| **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) | |

summary\_table\_white <- 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  
summary\_table\_white

| **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) | |

# all values same for all races, although percentagers of each race different in county, all feature values reported as the same

qol\_data\_long$population\_of\_race <- qol\_data\_long$percentage/100  
qol\_data\_long$pct\_poor\_to\_fair\_health <- qol\_data\_long$pct\_poor\_to\_fair\_health \* qol\_data\_long$percentage  
qol\_data\_long$life\_expectancy\_years <- qol\_data\_long$life\_expectancy\_years \* qol\_data\_long$percentage  
qol\_data\_long$pct\_voters <-qol\_data\_long$pct\_voters \* qol\_data\_long$percentage  
  
statistic\_table <- race\_stats\_calc(data = qol\_data\_long,  
 race = df\_race\_list,  
 feature = feature\_list)  
  
statistic\_table

## # 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 0.182 0.0994 0.274  
## 2 black 1.75 0.369 3.29   
## 3 hispanic 1.68 0.645 3.14   
## 4 native\_a… 0.296 0.0498 1.40   
## 5 other 0.374 0.156 0.742  
## 6 white 13.0 12.5 3.64   
## # 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 the summary table  
summary\_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  
summary\_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 | 0.1 (0.3) | 0.4 (3.3) | 0.6 (3.1) | 0.0 (1.4) | 0.2 (0.7) | 12.5 (3.6) |
| life\_expectancy\_years | 49 (174) | 184 (1,055) | 334 (1,081) | 25 (443) | 80 (295) | 6,801 (1,301) |
| pct\_voters | 0 (2) | 2 (9) | 3 (8) | 0 (3) | 1 (2) | 56 (14) |
| *1*Median (SD) | | | | | | |

# explore individual race variables  
summary\_table\_white <- 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  
summary\_table\_white

| **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) | |

summary\_table\_black <- 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  
summary\_table\_black

| **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) | |

# does not give mean percent of value for race but contribution of each mean race value to total population value  
# unable to extract these values by race, data gives percentage weigthed by county so all races will appear to have same values in each county, but will contribute to total differently. use um plots for best analysis by race

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 (%) | |

# summary by region  
summary\_table\_region <- tbl\_summary(  
 data = qol\_data,  
 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  
summary\_table\_region

| **Characteristic** | **Midwest** N = 1,018*1* | **Northeast** N = 199*1* | **South** N = 1,354*1* | **West** N = 384*1* |
| --- | --- | --- | --- | --- |
| pct\_poor\_to\_fair\_health | 0.14 (0.03, 1.89, 0.95) | 0.12 (0.02, -0.36, 0.15) | 0.19 (0.04, 0.68, 0.41) | 0.14 (0.03, 0.61, 0.60) |
| life\_expectancy\_years | 77.67 (2.60, 4.64, -0.55) | 78.75 (1.76, 0.00, 0.19) | 75.37 (2.67, 0.77, 0.30) | 78.88 (3.59, 7.85, 1.08) |
| pct\_voters | 0.68 (0.08, 0.65, -0.23) | 0.68 (0.08, 0.63, 0.25) | 0.61 (0.09, 0.24, 0.09) | 0.71 (0.10, -0.01, -0.25) |
| *1*Mean (SD, kurtosis, skewness) | | | | |

#rm(  
# counties,  
# data\_mg,  
# data\_mg\_part1,  
# data\_mg\_part2,  
# mg\_summary\_matrix,  
# qol\_data\_asian,  
# qol\_data\_black,  
# qol\_data\_hispanic,  
# qol\_data\_long,  
# qol\_data\_na,  
# qol\_data\_other,  
# qol\_data\_sorted,  
# qol\_data\_white,  
# statistic\_table,  
# summary\_table,  
# summary\_table\_black,  
# summary\_table\_white,  
# table\_1,  
# summary\_table\_region,  
# summary\_matrix,  
# summary\_stats,  
# top\_bottom\_counties,  
# df\_race\_list,  
# feature,  
# features,  
# feature\_list,  
# i,  
# qol\_names,  
# race\_vars,  
# variables,  
# mg\_sum\_stats,  
# race\_stats\_calc  
#)

*Data Prep for Model builds*

# Remove features with no analytical value namely Fipscode and county, state and pct\_poor\_to\_fair\_health  
qol\_data <- qol\_data %>%  
 select(  
 -c(fips\_code,  
 county,  
 state,  
 pct\_poor\_to\_fair\_health)  
 )

calcSplitRatio <- function(p = NA, df) {  
 ## @p = the number of parameters. by default, if none are provided, the number of columns (predictors) in the dataset are used  
 ## @df = the dataframe that will be used for the analysis  
   
 ## If the number of parameters isn't supplied, set it to the number of features minus 1 for the target  
 if(is.na(p)) {  
 p <- ncol(df) -1 ## COMMENT HERE  
 }  
   
 ## Calculate the ideal number of testing set  
 test\_N <- (1/sqrt(p))\*nrow(df)  
 ## Turn that into a testing proportion  
 test\_prop <- round((1/sqrt(p))\*nrow(df)/nrow(df), 2)  
 ## And find the training proportion  
 train\_prop <- 1-test\_prop  
   
 ## Tell us the results!  
 print(paste0("The ideal split ratio is ", train\_prop, ":", test\_prop, " training:testing"))  
   
 ## Return the size of the training set  
 return(train\_prop)  
}  
  
calcSplitRatio(df = qol\_data)

## [1] "The ideal split ratio is 0.85:0.15 training:testing"

## [1] 0.85

# Split the data into training and testing sets  
set.seed(123)  
split <- sample(2, nrow(qol\_data), replace = TRUE, prob = c(0.86, 0.14))  
qol\_train <- qol\_data[split == 1, ]  
qol\_test <- qol\_data[split == 2, ]  
  
# Check the distribution of the response variable in the training and testing sets  
table(qol\_train$response)

##   
## better worse   
## 1267 1293

table(qol\_test$response)

##   
## better worse   
## 190 205

# Create dummy variables for 'region' in training data  
dummies\_train <- dummyVars(~ region, data = qol\_train, fullRank = FALSE)  
qol\_train\_encoded <- predict(dummies\_train, newdata = qol\_train)  
qol\_train <- cbind(qol\_train, qol\_train\_encoded) %>%  
 select(-region) # Remove the original 'region' column  
  
# Create dummy variables for 'region' in test data  
dummies\_test <- dummyVars(~ region, data = qol\_test, fullRank = FALSE)  
qol\_test\_encoded <- predict(dummies\_test, newdata = qol\_test)  
qol\_test <- cbind(qol\_test, qol\_test\_encoded) %>%  
 select(-region) # Remove the original 'region' column  
  
# View the transformed training data  
print("Transformed Training Data:")

## [1] "Transformed Training Data:"

print(head(qol\_train))

## weighted\_population average\_hh\_size pct\_male pct\_native\_american pct\_asian  
## 1 55639 2.55 48.62 0.28 1.17  
## 2 218289 2.56 48.51 0.69 0.93  
## 3 25026 2.37 52.57 0.35 0.49  
## 4 10173 2.84 54.83 0.00 1.35  
## 5 19726 2.92 45.88 0.33 1.32  
## 6 33427 2.42 47.65 0.24 1.10  
## pct\_black pct\_hispanic pct\_other\_race pct\_white pct\_single\_parent  
## 1 19.53 2.88 0.67 75.76 27.41  
## 2 8.77 4.56 1.56 85.44 18.10  
## 3 47.67 4.44 3.10 46.30 52.76  
## 4 68.61 8.10 3.13 26.18 50.41  
## 5 44.59 1.47 0.50 51.62 39.55  
## 6 40.00 2.52 0.79 56.89 44.71  
## pct\_hh\_other\_computer pct\_hh\_internet pct\_employed\_nonprofit pct\_employed  
## 1 1.05 82.80 7.72 97.09  
## 2 1.75 85.52 5.68 96.08  
## 3 2.15 65.00 4.20 93.06  
## 4 1.84 62.79 6.76 96.65  
## 5 5.07 74.31 6.77 93.53  
## 6 1.63 74.84 4.38 96.41  
## pct\_hh\_inc\_99999 pct\_w\_medicare clinical\_nurse\_pt dentist\_pt pa\_pt  
## 1 30.40 6.87 0.02 0.34 0.04  
## 2 30.52 7.27 0.02 0.49 0.17  
## 3 23.63 7.51 0.00 0.37 0.04  
## 4 24.91 5.87 0.00 0.20 0.10  
## 5 25.09 7.49 0.00 0.36 0.00  
## 6 31.75 7.36 0.00 0.21 0.00  
## mental\_health\_faciliy\_pt population\_density days\_over\_90\_f median\_hh\_income.x  
## 1 0.0178 93.60 104 67565  
## 2 0.0174 137.30 97 71135  
## 3 0.0813 28.28 104 38866  
## 4 0.1002 16.33 103 33124  
## 5 0.1538 25.39 103 42268  
## 6 0.0000 56.03 87 39318  
## median\_er\_dist median\_trauma\_center\_dist median\_pediatric\_icu\_dist  
## 1 2.25 12.07 55.65  
## 2 6.01 25.37 22.62  
## 3 5.58 41.37 63.76  
## 4 4.89 39.11 56.79  
## 5 3.51 39.77 94.42  
## 6 4.67 29.43 4.67  
## median\_health\_clinic\_dist median\_drug\_alcohol\_care\_dist  
## 1 9.20 12.24  
## 2 8.57 13.21  
## 3 1.17 35.73  
## 4 1.63 4.89  
## 5 2.83 43.08  
## 6 4.42 23.21  
## percent\_grandparents\_as\_guardians pct\_adult\_smokers pct\_binge\_drinkers  
## 1 4.901316 0.183 0.1665626  
## 2 5.961440 0.169 0.1896517  
## 3 11.165644 0.259 0.1342877  
## 4 15.405228 0.255 0.1277270  
## 5 9.496840 0.223 0.1374772  
## 6 14.975107 0.217 0.1507498  
## pct\_under\_65\_no\_health\_insurance pct\_highschool\_diploma inequality\_ratio  
## 1 0.1055942 0.8958449 4.794400  
## 2 0.1087488 0.9101416 4.300710  
## 3 0.1436828 0.7567102 5.180597  
## 4 0.1236755 0.7760651 6.274395  
## 5 0.1348505 0.8556485 5.335266  
## 6 0.1321949 0.8278154 4.212493  
## social\_clubs\_per\_10k air\_polution\_metric water\_quality pct\_high\_housing\_costs  
## 1 12.645828 10.0 0 0.12635805  
## 2 9.594962 7.6 1 0.10561056  
## 3 9.353776 9.4 1 0.13464346  
## 4 7.016840 9.3 0 0.12785388  
## 5 9.228876 9.1 0 0.09451411  
## 6 18.865054 9.8 0 0.10871873  
## pct\_overcrowded\_hh pct\_30\_min\_plus\_commute life\_expectancy\_years  
## 1 0.01121757 0.416 76.58565  
## 2 0.01291239 0.376 77.72473  
## 3 0.03852328 0.365 72.86721  
## 4 0.00000000 0.494 73.81468  
## 5 0.01537279 0.347 73.47328  
## 6 0.05278810 0.281 73.76110  
## school\_funding\_gap pct\_voters pct\_home\_owner pct\_rural\_population response  
## 1 -2077.1420 0.6618208 16227 0.4200216 worse  
## 2 343.0381 0.6529095 67242 0.4227910 better  
## 3 -13560.4500 0.5402157 5654 0.6778963 worse  
## 4 -15766.8500 0.5906530 2220 0.5137438 worse  
## 5 -9255.5640 0.6291777 4728 0.7123216 worse  
## 6 -11972.6670 0.5852575 9089 0.4914803 worse  
## region.Midwest region.Northeast region.South region.West  
## 1 0 0 1 0  
## 2 0 0 1 0  
## 3 0 0 1 0  
## 4 0 0 1 0  
## 5 0 0 1 0  
## 6 0 0 1 0

# View the transformed test data  
print("Transformed Test Data:")

## [1] "Transformed Test Data:"

print(head(qol\_test))

## weighted\_population average\_hh\_size pct\_male pct\_native\_american pct\_asian  
## 1 22374 2.85 53.73 0.05 0.25  
## 2 57755 2.70 49.65 0.10 0.41  
## 3 114324 2.50 48.04 0.31 0.81  
## 4 44147 2.55 48.75 1.14 0.44  
## 5 52238 2.59 49.42 1.00 1.61  
## 6 37096 2.43 48.61 0.21 0.66  
## pct\_black pct\_hispanic pct\_other\_race pct\_white pct\_single\_parent  
## 1 22.55 2.68 0.04 76.60 32.19  
## 2 1.40 9.28 1.80 93.97 25.75  
## 3 21.16 3.85 1.64 73.30 34.42  
## 4 9.67 7.75 3.56 82.13 34.19  
## 5 17.05 7.68 1.29 74.69 32.20  
## 6 12.69 1.77 0.36 84.27 29.04  
## pct\_hh\_other\_computer pct\_hh\_internet pct\_employed\_nonprofit pct\_employed  
## 1 0.23 76.17 3.48 92.56  
## 2 2.16 80.03 5.30 94.80  
## 3 2.43 79.97 6.03 92.71  
## 4 1.40 74.92 6.60 93.90  
## 5 2.24 83.18 4.92 94.50  
## 6 11.13 75.93 7.04 92.30  
## pct\_hh\_inc\_99999 pct\_w\_medicare clinical\_nurse\_pt dentist\_pt pa\_pt  
## 1 34.92 7.74 0.00 0.27 0.23  
## 2 28.64 8.81 0.00 0.19 0.00  
## 3 31.39 5.18 0.01 0.68 0.10  
## 4 34.07 5.28 0.00 0.29 0.02  
## 5 31.50 4.30 0.00 0.34 0.04  
## 6 28.68 7.69 0.00 0.38 0.08  
## mental\_health\_faciliy\_pt population\_density days\_over\_90\_f median\_hh\_income.x  
## 1 0.0000 35.94 97 50907  
## 2 0.0173 89.56 80 55203  
## 3 0.0353 188.69 84 50259  
## 4 0.0225 63.72 94 52693  
## 5 0.0188 76.94 105 54203  
## 6 0.0812 36.00 107 43544  
## median\_er\_dist median\_trauma\_center\_dist median\_pediatric\_icu\_dist  
## 1 8.44 26.40 21.97  
## 2 10.56 26.16 52.24  
## 3 3.51 24.79 45.01  
## 4 7.27 31.37 28.36  
## 5 2.74 23.44 44.21  
## 6 1.86 12.59 61.42  
## median\_health\_clinic\_dist median\_drug\_alcohol\_care\_dist  
## 1 3.85 26.48  
## 2 3.91 25.22  
## 3 4.08 15.30  
## 4 3.51 30.65  
## 5 4.12 29.84  
## 6 1.86 54.17  
## percent\_grandparents\_as\_guardians pct\_adult\_smokers pct\_binge\_drinkers  
## 1 12.543648 0.228 0.1591628  
## 2 6.226566 0.218 0.1631069  
## 3 5.798538 0.210 0.1618824  
## 4 4.091697 0.221 0.1637214  
## 5 4.730133 0.196 0.1525075  
## 6 8.664498 0.214 0.1538523  
## pct\_under\_65\_no\_health\_insurance pct\_highschool\_diploma inequality\_ratio  
## 1 0.1300788 0.8053808 5.034985  
## 2 0.1325467 0.8364808 4.803718  
## 3 0.1292500 0.8521486 4.819868  
## 4 0.1526298 0.8176132 4.131532  
## 5 0.1369450 0.8635622 4.790742  
## 6 0.1209324 0.8518224 5.483394  
## social\_clubs\_per\_10k air\_polution\_metric water\_quality pct\_high\_housing\_costs  
## 1 9.035056 9.8 0 0.07985348  
## 2 6.738195 9.6 0 0.07380601  
## 3 13.395729 9.7 0 0.09838489  
## 4 14.865869 9.9 0 0.11185830  
## 5 12.586887 9.1 0 0.07904374  
## 6 11.643650 9.1 0 0.09875173  
## pct\_overcrowded\_hh pct\_30\_min\_plus\_commute life\_expectancy\_years  
## 1 0.01161103 0.551 73.60936  
## 2 0.01798993 0.595 74.17146  
## 3 0.01681426 0.300 72.82501  
## 4 0.02983752 0.508 74.39108  
## 5 0.01756587 0.308 75.26854  
## 6 0.02121212 0.279 72.91764  
## school\_funding\_gap pct\_voters pct\_home\_owner pct\_rural\_population response  
## 1 -2659.8410 0.5456355 5580 0.6835261 worse  
## 2 -889.3883 0.6418799 16865 0.8995150 worse  
## 3 -4113.4269 0.5785633 31441 0.3369683 worse  
## 4 -1903.9890 0.5950693 12221 0.8674472 worse  
## 5 -2322.7189 0.5721074 14153 0.4719508 worse  
## 6 -4510.3738 0.6034274 10681 0.6965179 worse  
## region.Midwest region.Northeast region.South region.West  
## 1 0 0 1 0  
## 2 0 0 1 0  
## 3 0 0 1 0  
## 4 0 0 1 0  
## 5 0 0 1 0  
## 6 0 0 1 0

# Encode response variable as 1 and 0. (1 if worse than national median and 0 if better than national median) and convert to factor  
  
#encode response variable in training set  
qol\_train <- qol\_train %>%  
 mutate(response = ifelse(response == "better", 0, 1))  
table(qol\_data$response)

##   
## better worse   
## 1457 1498

# Convert response variable to factor   
qol\_train$response <- as.factor(qol\_train$response)  
  
#encode response variable in testing set  
qol\_test <- qol\_test %>%  
 mutate(response = ifelse(response == "better", 0, 1))  
table(qol\_test$response)

##   
## 0 1   
## 190 205

# Convert response variable to factor   
qol\_test$response <- as.factor(qol\_test$response)

#Handling missing values; identifying and imputing missing characters  
  
# Count missing values in the train set  
total\_missing <- sum(is.na(qol\_train))  
print(paste("Total missing values in dataframe:", total\_missing))

## [1] "Total missing values in dataframe: 0"

# Count missing values in the test set  
total\_missing <- sum(is.na(qol\_train))  
print(paste("Total missing values in dataframe:", total\_missing))

## [1] "Total missing values in dataframe: 0"

#To determine if there are unwanted characters and whitespace and make corrections:  
#We print a sample of the data for visual inspection.  
#We use regular expressions to detect unwanted characters in text columns.  
#We check for leading, trailing, or excessive whitespace in text columns.  
  
# Visual inspection  
#print(head(qol\_train))  
  
# Detect unwanted characters using regular expressions  
unwanted\_chars <- function(x) {  
 if (is.character(x)) {  
 return(grepl("[^[:alnum:][:space:]]", x))  
 } else {  
 return(FALSE)  
 }  
}  
  
# Apply the function to each column qol\_train  
unwanted\_chars\_results <- sapply(qol\_train, function(col) sum(unwanted\_chars(col)))  
print("Number of unwanted characters in each column:")

## [1] "Number of unwanted characters in each column:"

#print(unwanted\_chars\_results)  
  
# Apply the function to each column qol\_test  
unwanted\_chars\_results <- sapply(qol\_train, function(col) sum(unwanted\_chars(col)))  
print("Number of unwanted characters in each column:")

## [1] "Number of unwanted characters in each column:"

#print(unwanted\_chars\_results)  
  
  
  
# Detect leading, trailing, or excessive whitespace  
whitespace\_check <- function(x) {  
 if (is.character(x)) {  
 return(grepl("^\\s+|\\s+$|\\s{2,}", x))  
 } else {  
 return(FALSE)  
 }  
}  
  
# Apply the function to each column of training and test sets  
whitespace\_results <- sapply(qol\_train, function(col) sum(whitespace\_check(col)))  
print("Number of whitespace issues in each column:")

## [1] "Number of whitespace issues in each column:"

#print(whitespace\_results)  
  
  
whitespace\_results <- sapply(qol\_test, function(col) sum(whitespace\_check(col)))  
print("Number of whitespace issues in each column:")

## [1] "Number of whitespace issues in each column:"

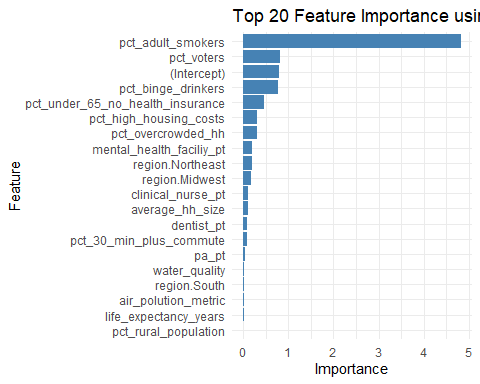
#print(whitespace\_results)

*OlS Model*

# Convert response variable to numeric  
qol\_train$response <- as.numeric(as.character(qol\_train$response))  
# Fit the OLS model  
ols\_model <- lm(response ~ ., data = qol\_train)  
# Predict on the test set  
predictions <- predict(ols\_model, newdata = qol\_test)  
  
# Convert predictions to binary outcomes  
predicted\_classes <- ifelse(predictions > 0.5, 1, 0)  
  
# Evaluate the model  
confusion\_matrix <- table(qol\_test$response, predicted\_classes)  
accuracy\_ols <- sum(diag(confusion\_matrix)) / sum(confusion\_matrix)  
# Extract coefficients  
coefficients <- summary(ols\_model)$coefficients  
  
# Calculate importance (absolute value of coefficients)  
feature\_importance <- abs(coefficients[, "Estimate"])  
  
# Create a data frame for better visualization  
importance\_df <- data.frame(Feature = rownames(coefficients), Importance = feature\_importance)  
  
# Sort by importance  
importance\_df <- importance\_df[order(-importance\_df$Importance), ]  
  
print(importance\_df)

## Feature  
## pct\_adult\_smokers pct\_adult\_smokers  
## pct\_voters pct\_voters  
## (Intercept) (Intercept)  
## pct\_binge\_drinkers pct\_binge\_drinkers  
## pct\_under\_65\_no\_health\_insurance pct\_under\_65\_no\_health\_insurance  
## pct\_high\_housing\_costs pct\_high\_housing\_costs  
## pct\_overcrowded\_hh pct\_overcrowded\_hh  
## mental\_health\_faciliy\_pt mental\_health\_faciliy\_pt  
## region.Northeast region.Northeast  
## region.Midwest region.Midwest  
## clinical\_nurse\_pt clinical\_nurse\_pt  
## average\_hh\_size average\_hh\_size  
## dentist\_pt dentist\_pt  
## pct\_30\_min\_plus\_commute pct\_30\_min\_plus\_commute  
## pa\_pt pa\_pt  
## water\_quality water\_quality  
## region.South region.South  
## air\_polution\_metric air\_polution\_metric  
## life\_expectancy\_years life\_expectancy\_years  
## pct\_rural\_population pct\_rural\_population  
## pct\_other\_race pct\_other\_race  
## pct\_male pct\_male  
## pct\_highschool\_diploma pct\_highschool\_diploma  
## inequality\_ratio inequality\_ratio  
## pct\_hispanic pct\_hispanic  
## percent\_grandparents\_as\_guardians percent\_grandparents\_as\_guardians  
## pct\_black pct\_black  
## median\_health\_clinic\_dist median\_health\_clinic\_dist  
## pct\_w\_medicare pct\_w\_medicare  
## pct\_hh\_inc\_99999 pct\_hh\_inc\_99999  
## pct\_white pct\_white  
## median\_trauma\_center\_dist median\_trauma\_center\_dist  
## pct\_hh\_other\_computer pct\_hh\_other\_computer  
## pct\_native\_american pct\_native\_american  
## median\_drug\_alcohol\_care\_dist median\_drug\_alcohol\_care\_dist  
## pct\_employed pct\_employed  
## pct\_asian pct\_asian  
## days\_over\_90\_f days\_over\_90\_f  
## social\_clubs\_per\_10k social\_clubs\_per\_10k  
## median\_er\_dist median\_er\_dist  
## pct\_hh\_internet pct\_hh\_internet  
## pct\_employed\_nonprofit pct\_employed\_nonprofit  
## pct\_single\_parent pct\_single\_parent  
## median\_pediatric\_icu\_dist median\_pediatric\_icu\_dist  
## school\_funding\_gap school\_funding\_gap  
## population\_density population\_density  
## pct\_home\_owner pct\_home\_owner  
## median\_hh\_income.x median\_hh\_income.x  
## weighted\_population weighted\_population  
## Importance  
## pct\_adult\_smokers 4.828369e+00  
## pct\_voters 8.247546e-01  
## (Intercept) 7.903573e-01  
## pct\_binge\_drinkers 7.839616e-01  
## pct\_under\_65\_no\_health\_insurance 4.602606e-01  
## pct\_high\_housing\_costs 3.170296e-01  
## pct\_overcrowded\_hh 3.000245e-01  
## mental\_health\_faciliy\_pt 2.026490e-01  
## region.Northeast 1.932589e-01  
## region.Midwest 1.756903e-01  
## clinical\_nurse\_pt 1.054102e-01  
## average\_hh\_size 1.051512e-01  
## dentist\_pt 8.346996e-02  
## pct\_30\_min\_plus\_commute 8.052080e-02  
## pa\_pt 4.887201e-02  
## water\_quality 2.211443e-02  
## region.South 2.189672e-02  
## air\_polution\_metric 1.592127e-02  
## life\_expectancy\_years 1.422724e-02  
## pct\_rural\_population 1.076160e-02  
## pct\_other\_race 8.430577e-03  
## pct\_male 8.109387e-03  
## pct\_highschool\_diploma 7.186500e-03  
## inequality\_ratio 6.277583e-03  
## pct\_hispanic 4.537704e-03  
## percent\_grandparents\_as\_guardians 3.596366e-03  
## pct\_black 3.250043e-03  
## median\_health\_clinic\_dist 2.991093e-03  
## pct\_w\_medicare 2.917590e-03  
## pct\_hh\_inc\_99999 2.472234e-03  
## pct\_white 2.407196e-03  
## median\_trauma\_center\_dist 1.965789e-03  
## pct\_hh\_other\_computer 1.500778e-03  
## pct\_native\_american 1.436703e-03  
## median\_drug\_alcohol\_care\_dist 1.374840e-03  
## pct\_employed 1.263213e-03  
## pct\_asian 1.222137e-03  
## days\_over\_90\_f 9.324034e-04  
## social\_clubs\_per\_10k 8.707673e-04  
## median\_er\_dist 7.218458e-04  
## pct\_hh\_internet 3.840578e-04  
## pct\_employed\_nonprofit 2.926322e-04  
## pct\_single\_parent 6.075693e-05  
## median\_pediatric\_icu\_dist 5.675872e-05  
## school\_funding\_gap 3.458337e-06  
## population\_density 1.357713e-06  
## pct\_home\_owner 7.047753e-07  
## median\_hh\_income.x 5.233738e-07  
## weighted\_population 1.620800e-07

# Select the top 20 important features  
top\_20\_features\_ols <- importance\_df %>%  
 arrange(desc(Importance)) %>%  
 head(20)  
  
# Create the bar plot for the top 20 features  
ggplot(top\_20\_features\_ols, aes(x = reorder(Feature, Importance), y = Importance)) +  
 geom\_bar(stat = "identity", fill = "steelblue") +  
 coord\_flip() + # Flip coordinates for better readability  
 labs(title = "Top 20 Feature Importance using ols", x = "Feature", y = "Importance") +  
 theme\_minimal()

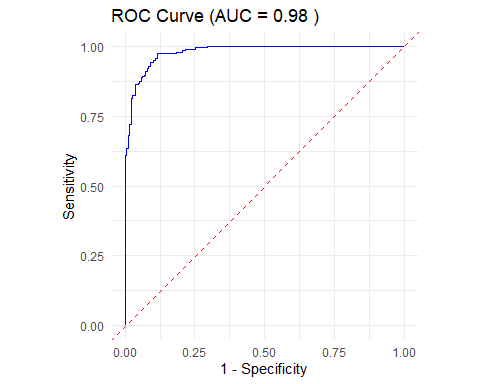


# Calculate ROC curve  
roc\_ols <- pROC::roc(qol\_test$response, predictions)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

# Create a data frame for ggplot  
roc\_data <- data.frame(  
 specificity = rev(roc\_ols$specificities),  
 sensitivity = rev(roc\_ols$sensitivities)  
)  
  
# Plot ROC curve using ggplot2  
ggplot(roc\_data, aes(x = 1-specificity, y = sensitivity)) +  
 geom\_line(color = "blue") +  
 geom\_abline(linetype = "dashed", color = "red") +  
 labs(title = paste("ROC Curve (AUC =", round(pROC::auc(roc\_ols), 2), ")"),  
 x = "1 - Specificity",  
 y = "Sensitivity") +  
 theme\_minimal() +  
 coord\_fixed(ratio = 1)



*Random Forest*

names(qol\_train)

## [1] "weighted\_population" "average\_hh\_size"   
## [3] "pct\_male" "pct\_native\_american"   
## [5] "pct\_asian" "pct\_black"   
## [7] "pct\_hispanic" "pct\_other\_race"   
## [9] "pct\_white" "pct\_single\_parent"   
## [11] "pct\_hh\_other\_computer" "pct\_hh\_internet"   
## [13] "pct\_employed\_nonprofit" "pct\_employed"   
## [15] "pct\_hh\_inc\_99999" "pct\_w\_medicare"   
## [17] "clinical\_nurse\_pt" "dentist\_pt"   
## [19] "pa\_pt" "mental\_health\_faciliy\_pt"   
## [21] "population\_density" "days\_over\_90\_f"   
## [23] "median\_hh\_income.x" "median\_er\_dist"   
## [25] "median\_trauma\_center\_dist" "median\_pediatric\_icu\_dist"   
## [27] "median\_health\_clinic\_dist" "median\_drug\_alcohol\_care\_dist"   
## [29] "percent\_grandparents\_as\_guardians" "pct\_adult\_smokers"   
## [31] "pct\_binge\_drinkers" "pct\_under\_65\_no\_health\_insurance"   
## [33] "pct\_highschool\_diploma" "inequality\_ratio"   
## [35] "social\_clubs\_per\_10k" "air\_polution\_metric"   
## [37] "water\_quality" "pct\_high\_housing\_costs"   
## [39] "pct\_overcrowded\_hh" "pct\_30\_min\_plus\_commute"   
## [41] "life\_expectancy\_years" "school\_funding\_gap"   
## [43] "pct\_voters" "pct\_home\_owner"   
## [45] "pct\_rural\_population" "response"   
## [47] "region.Midwest" "region.Northeast"   
## [49] "region.South" "region.West"

# Train a random forest model for classification  
model\_rf <- randomForest::randomForest(as.factor(response) ~ ., data = qol\_train, ntree = 500)  
  
# Predict the response variable for the testing set  
predictions2 <- predict(model\_rf, qol\_test, type = "response")  
  
# Calculate the accuracy of the random forest model  
accuracy <- mean(qol\_test$response == predictions2)  
print(accuracy)

## [1] 0.9189873

# Calculate the ROC AUC score  
roc\_obj <- pROC::roc(qol\_test$response, as.numeric(predictions2))

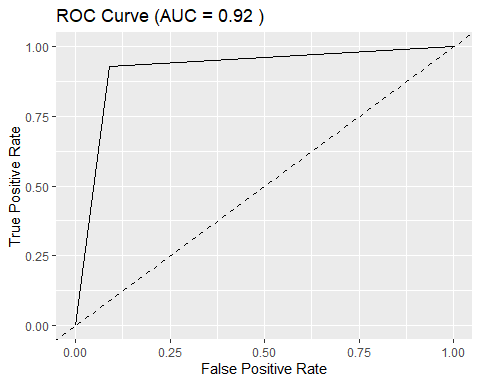
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

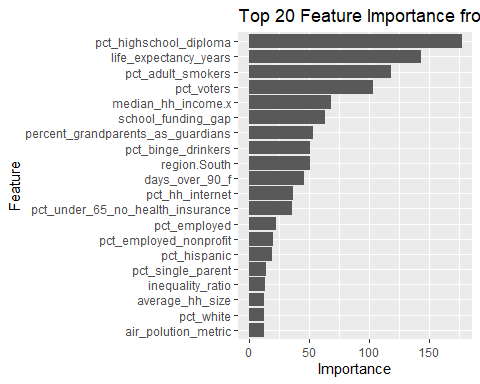
auc\_rf <- pROC::auc(roc\_obj)  
print(auc\_rf)

## Area under the curve: 0.9187

# Plot the ROC curve using ggplot2  
roc\_df <- data.frame(  
 tpr = roc\_obj$sensitivities,  
 fpr = 1 - roc\_obj$specificities  
)  
  
ggplot(roc\_df, aes(x = fpr, y = tpr)) +  
 geom\_line() +  
 geom\_abline(linetype = "dashed") +  
 xlab("False Positive Rate") +  
 ylab("True Positive Rate") +  
 ggtitle(paste("ROC Curve (AUC =", round(auc\_rf, 2), ")"))



# Calculate feature importance  
importance\_rf <- randomForest::importance(model\_rf)  
importance\_df <- data.frame(Feature = rownames(importance\_rf), Importance = importance\_rf[, 1])  
  
# Select top 20 important features  
top\_20\_importance <- importance\_df[order(-importance\_df$Importance), ][1:20, ]  
  
# Plot top 20 feature importance using ggplot2  
ggplot(top\_20\_importance, aes(x = reorder(Feature, Importance), y = Importance)) +  
 geom\_bar(stat = "identity") +  
   
 coord\_flip() +  
 xlab("Feature") +  
 ylab("Importance") +  
 ggtitle("Top 20 Feature Importance from Random Forest Model")

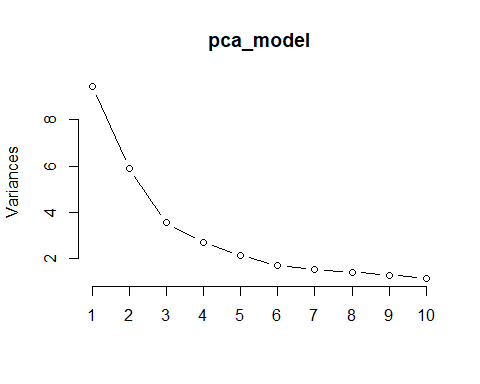


*PCA*

qol\_train\_standardized <- qol\_train %>%  
 select(-response) %>%  
 scale() # standardize the data (exclude the response variable)  
  
pca\_model <- prcomp(  
 qol\_train\_standardized,  
 center = TRUE,  
 scale. = TRUE) # apply PCA  
  
summary(pca\_model) # print summary of PCA model

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 3.0738 2.4277 1.88430 1.64694 1.46286 1.29916 1.23369  
## Proportion of Variance 0.1928 0.1203 0.07246 0.05536 0.04367 0.03445 0.03106  
## Cumulative Proportion 0.1928 0.3131 0.38556 0.44091 0.48459 0.51903 0.55009  
## PC8 PC9 PC10 PC11 PC12 PC13 PC14  
## Standard deviation 1.18920 1.13229 1.06695 1.02135 0.99346 0.97896 0.96790  
## Proportion of Variance 0.02886 0.02617 0.02323 0.02129 0.02014 0.01956 0.01912  
## Cumulative Proportion 0.57895 0.60512 0.62835 0.64964 0.66978 0.68934 0.70846  
## PC15 PC16 PC17 PC18 PC19 PC20 PC21  
## Standard deviation 0.95441 0.93561 0.91171 0.90620 0.85330 0.84947 0.79781  
## Proportion of Variance 0.01859 0.01786 0.01696 0.01676 0.01486 0.01473 0.01299  
## Cumulative Proportion 0.72705 0.74492 0.76188 0.77864 0.79350 0.80822 0.82121  
## PC22 PC23 PC24 PC25 PC26 PC27 PC28  
## Standard deviation 0.77590 0.75804 0.7441 0.73246 0.71214 0.69716 0.68708  
## Proportion of Variance 0.01229 0.01173 0.0113 0.01095 0.01035 0.00992 0.00963  
## Cumulative Proportion 0.83350 0.84523 0.8565 0.86748 0.87783 0.88774 0.89738  
## PC29 PC30 PC31 PC32 PC33 PC34 PC35  
## Standard deviation 0.6677 0.65763 0.65260 0.62780 0.6104 0.59681 0.56805  
## Proportion of Variance 0.0091 0.00883 0.00869 0.00804 0.0076 0.00727 0.00659  
## Cumulative Proportion 0.9065 0.91530 0.92399 0.93204 0.9396 0.94691 0.95349  
## PC36 PC37 PC38 PC39 PC40 PC41 PC42  
## Standard deviation 0.54476 0.53473 0.51540 0.50718 0.47576 0.45814 0.42964  
## Proportion of Variance 0.00606 0.00584 0.00542 0.00525 0.00462 0.00428 0.00377  
## Cumulative Proportion 0.95955 0.96539 0.97081 0.97606 0.98068 0.98496 0.98873  
## PC43 PC44 PC45 PC46 PC47 PC48  
## Standard deviation 0.41021 0.39122 0.35776 0.28203 0.13587 0.07104  
## Proportion of Variance 0.00343 0.00312 0.00261 0.00162 0.00038 0.00010  
## Cumulative Proportion 0.99216 0.99528 0.99790 0.99952 0.99990 1.00000  
## PC49  
## Standard deviation 1.521e-15  
## Proportion of Variance 0.000e+00  
## Cumulative Proportion 1.000e+00

plot(pca\_model, type = "l") # plot explained variance



pca\_components <- predict(pca\_model, qol\_train\_standardized) # extract the principal components  
  
head(pca\_components)

## PC1 PC2 PC3 PC4 PC5 PC6  
## 1 -0.6127441 -0.9361597 1.3031995 1.7763247 0.288469981 -0.9666984  
## 2 0.5807855 -1.1181728 0.3408208 1.9352077 -0.487141027 -0.4551443  
## 3 -7.1467713 -0.4795094 0.6581118 -0.9474059 0.003362109 -1.4972232  
## 4 -7.4119548 -1.1695122 1.8557677 -0.2122726 -0.522624808 -1.4578988  
## 5 -4.8750373 -0.2791846 0.8859073 -0.2640337 -0.509742062 -1.9973717  
## 6 -4.5049055 -0.4989681 1.4467336 0.9313069 1.105113000 -0.9705355  
## PC7 PC8 PC9 PC10 PC11 PC12  
## 1 0.53701912 -0.8816795 0.68944699 -0.059828171 -0.53980355 -0.5759332  
## 2 -0.01007776 -0.9898687 0.10904640 1.077253528 -0.25855318 0.9614640  
## 3 0.17781358 0.3710773 0.07664114 -1.477254937 1.14789471 1.4237628  
## 4 1.21345581 0.2218297 0.50435337 -3.334930626 -0.08728498 0.4034684  
## 5 1.70780036 -0.1748125 0.12255645 0.126536289 -0.42217926 -0.4004569  
## 6 1.57067275 0.4392751 0.23098429 -0.001550246 0.42583037 -0.3558475  
## PC13 PC14 PC15 PC16 PC17 PC18  
## 1 -0.4724690 -0.4802183 -0.9591240 0.6665025 1.01012416 0.1069751  
## 2 -0.6538932 -0.4004458 -0.9221007 0.1991984 0.24266555 0.7797966  
## 3 -0.3099003 -0.0746471 0.2337516 -0.7371858 -0.17164559 0.6194155  
## 4 -0.1877531 -0.7816651 0.6714243 -0.6278555 0.33998562 -0.3538821  
## 5 -0.2354246 1.5582305 0.9067967 -0.8887286 2.05559574 1.3223058  
## 6 -0.6891490 0.8209054 0.2287970 -0.5225074 -0.02587832 -0.9518404  
## PC19 PC20 PC21 PC22 PC23 PC24  
## 1 -0.14797562 -0.5880193 -0.06184543 0.17718739 0.0517747 0.9394226  
## 2 0.14377536 -0.6046132 -0.42721970 -0.09545512 0.5483337 -0.5185464  
## 3 -0.06567448 -0.3454150 0.69206841 -0.27283053 0.2642255 0.6424421  
## 4 -0.86413155 -0.0227908 1.58137196 0.76055409 1.1854040 0.5878847  
## 5 -0.64342697 0.4466650 0.55158604 -0.50086839 -0.3121362 1.9032502  
## 6 0.81541025 -0.8067816 -0.13255422 0.62625918 0.8145652 0.2035835  
## PC25 PC26 PC27 PC28 PC29 PC30 PC31  
## 1 0.24201366 0.2251178 0.5672451 0.1900485 1.1604255 -0.2338644 -0.3552296  
## 2 -0.02834286 -0.2681109 -0.2488004 0.2904646 0.2245030 0.2638751 0.2570353  
## 3 0.45479351 -0.1674047 0.3634690 0.3220357 0.2804890 -0.3541319 0.5994937  
## 4 0.52805139 -1.6586806 0.4394417 1.3127146 1.8098382 -0.3075267 0.5018353  
## 5 0.60317187 -0.7251845 0.7473448 0.2744898 0.4777129 0.6191258 0.1462448  
## 6 -1.19665983 0.2040953 -0.2002272 -0.8517779 0.5873163 -2.1744199 0.3579190  
## PC32 PC33 PC34 PC35 PC36 PC37  
## 1 0.3804850 -0.31967824 0.5519703 -0.01550138 -0.238328628 -0.17132098  
## 2 0.1452856 -0.15100328 -0.3676045 -0.12185567 -0.009529687 -0.38238027  
## 3 -0.7470514 0.38625358 0.5259742 0.01033937 0.751602162 -0.37445727  
## 4 -1.0462428 -0.70628369 -0.3845502 0.28310568 -0.045315955 0.08384884  
## 5 -0.2727371 -0.06422084 -0.8034518 -0.07403673 -0.277491067 -0.35619534  
## 6 -1.3148505 0.23769967 -0.1113828 0.16093579 0.433304265 -0.39327316  
## PC38 PC39 PC40 PC41 PC42 PC43  
## 1 -0.27169033 0.11079301 0.15765945 0.1651158 0.3142116 -0.283652967  
## 2 0.06639147 0.56527128 0.53532927 0.1635970 0.1050017 -0.303198744  
## 3 0.45579101 0.09655068 0.26483966 0.2905933 0.4715022 -0.006811686  
## 4 -1.25431583 -0.73820526 0.04760868 0.2601355 1.3353003 0.310455606  
## 5 -0.52865594 -0.18927769 0.71738584 -0.8451750 0.4062473 -0.070985592  
## 6 0.94273395 0.38217557 0.89028082 0.3844681 0.8621787 -0.104104675  
## PC44 PC45 PC46 PC47 PC48 PC49  
## 1 0.414200587 0.1726176 -0.06081322 0.06246710 0.020140664 -1.609823e-15  
## 2 0.198865965 0.3169539 -0.21709086 -0.12874657 0.031378129 5.551115e-17  
## 3 0.331106290 0.3700642 -0.13055504 -0.02803791 0.019602823 -3.719247e-15  
## 4 -0.001246531 -0.8470838 0.20530970 0.07836619 0.030262054 -4.163336e-15  
## 5 -0.451398036 -0.4900740 0.14337599 0.01096690 0.004142922 -2.053913e-15  
## 6 -0.295310709 -0.1740252 -0.27604566 -0.06742737 0.045074205 -3.719247e-15

qol\_train$response <- as.factor(qol\_train$response) # ensure the response variable is a factor  
  
train\_control <- trainControl(method = "cv", number = 5) # define control for cross-validation  
  
  
set.seed(123)  
ols\_cv <- train(response ~ ., data = qol\_train, method = "glm", family = binomial, trControl = train\_control) # train an OLS model: cross-validation used for classification

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from rank-deficient fit; attr(\*, "non-estim") has doubtful cases

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from rank-deficient fit; attr(\*, "non-estim") has doubtful cases

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from rank-deficient fit; attr(\*, "non-estim") has doubtful cases  
  
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from rank-deficient fit; attr(\*, "non-estim") has doubtful cases

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from rank-deficient fit; attr(\*, "non-estim") has doubtful cases

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

print(ols\_cv)

## Generalized Linear Model   
##   
## 2560 samples  
## 49 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 2049, 2048, 2048, 2048, 2047   
## Resampling results:  
##   
## Accuracy Kappa   
## 0.9402297 0.8804562

print(ols\_cv$results) # OLS cross-validation results

## parameter Accuracy Kappa AccuracySD KappaSD  
## 1 none 0.9402297 0.8804562 0.0118743 0.02374931

print(paste("Best OLS model accuracy:", max(ols\_cv$results$Accuracy)))

## [1] "Best OLS model accuracy: 0.940229710918754"

set.seed(123)  
rf\_cv <- train(response ~ ., data = qol\_train, method = "rf", trControl = train\_control, ntree = 500) # train random forest model using cross-validation for classification  
  
print(rf\_cv)

## Random Forest   
##   
## 2560 samples  
## 49 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 2049, 2048, 2048, 2048, 2047   
## Resampling results across tuning parameters:  
##   
## mtry Accuracy Kappa   
## 2 0.9359275 0.8718624  
## 25 0.9324134 0.8648050  
## 49 0.9257735 0.8515097  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was mtry = 2.

print(rf\_cv$results)

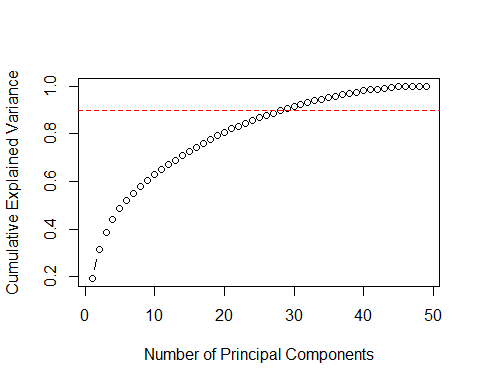
## mtry Accuracy Kappa AccuracySD KappaSD  
## 1 2 0.9359275 0.8718624 0.01054265 0.02110702  
## 2 25 0.9324134 0.8648050 0.01245597 0.02492389  
## 3 49 0.9257735 0.8515097 0.01169499 0.02339492

print(paste("best rf model accuracy = ", max(rf\_cv$results$Accuracy)))

## [1] "best rf model accuracy = 0.935927481911113"

# print random Forest cross-validation results

# Calculate cumulative explained variance  
explained\_variance <- summary(pca\_model)$importance[2, ]  
cumulative\_explained\_variance <- cumsum(explained\_variance)  
  
# Plot cumulative explained variance  
plot(  
 cumulative\_explained\_variance,  
 type = "b",  
 xlab = "Number of Principal Components",  
 ylab = "Cumulative Explained Variance"  
 )  
abline(  
 h = 0.9,  
 col = "red",  
 lty = 2) # line for 95% explained variance



# Find the number of components needed to explain 95% of the variance  
num\_components\_90 <- which(cumulative\_explained\_variance >= 0.9)[1]  
num\_components\_90

## PC29   
## 29

num\_components\_95 <- which(cumulative\_explained\_variance >= 0.95)[1]  
num\_components\_95

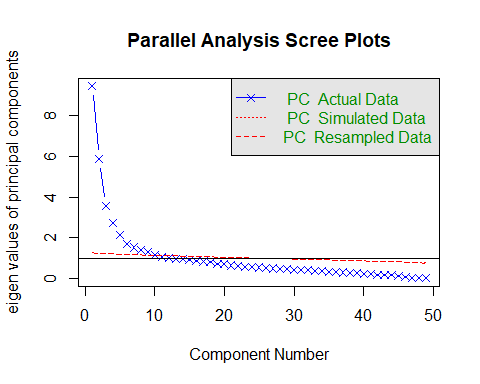
## PC35   
## 35

results\_psych <- psych::fa.parallel(  
 qol\_train\_standardized,  
 fa = "pc",  
 n.iter = 5000,  
 quant = 0.95)

## In smc, smcs < 0 were set to .0  
## In smc, smcs < 0 were set to .0  
## In smc, smcs < 0 were set to .0

## Warning in cor.smooth(r): Matrix was not positive definite, smoothing was done

## In factor.scores, the correlation matrix is singular, the pseudo inverse is used



## Parallel analysis suggests that the number of factors = NA and the number of components = 9

results\_psych$pc.values

## [1] 9.448158e+00 5.893636e+00 3.550599e+00 2.712404e+00 2.139955e+00  
## [6] 1.687826e+00 1.521996e+00 1.414208e+00 1.282091e+00 1.138388e+00  
## [11] 1.043147e+00 9.869625e-01 9.583672e-01 9.368396e-01 9.108978e-01  
## [16] 8.753683e-01 8.312208e-01 8.211898e-01 7.281155e-01 7.215916e-01  
## [21] 6.365071e-01 6.020176e-01 5.746213e-01 5.536883e-01 5.364955e-01  
## [26] 5.071371e-01 4.860293e-01 4.720759e-01 4.457856e-01 4.324792e-01  
## [31] 4.258807e-01 3.941344e-01 3.725534e-01 3.561831e-01 3.226841e-01  
## [36] 2.967647e-01 2.859380e-01 2.656386e-01 2.572278e-01 2.263480e-01  
## [41] 2.098890e-01 1.845947e-01 1.682697e-01 1.530549e-01 1.279926e-01  
## [46] 7.954012e-02 1.846141e-02 5.047163e-03 -9.092872e-17

results\_psych$pc.sim

## [1] 1.2725715 1.2481236 1.2293236 1.2131355 1.1984121 1.1847580 1.1720796  
## [8] 1.1598719 1.1483247 1.1369568 1.1262069 1.1156199 1.1052274 1.0951341  
## [15] 1.0853450 1.0756404 1.0661706 1.0568986 1.0475790 1.0384396 1.0294900  
## [22] 1.0204846 1.0117088 1.0028018 0.9941526 0.9853351 0.9767283 0.9681147  
## [29] 0.9594666 0.9509057 0.9422066 0.9335329 0.9249724 0.9163467 0.9077816  
## [36] 0.8988845 0.8900990 0.8811287 0.8721191 0.8629643 0.8536481 0.8440989  
## [43] 0.8343249 0.8240444 0.8133928 0.8021128 0.7896886 0.7757268 0.7579196

# PCA give low value to features with high collinearity  
  
# 35 features predict 95% of response variability  
# 29 features predict 90% of response variability  
# sqrt(obresvations) = 51, rounded   
  
# Parallel analysis suggests that the number of factors = NA and the number of components = 9

# extract the features (rotation matrix)  
features <- pca\_model$rotation  
  
# subset the loadings for the first 35 principal components  
top\_features <- features[, 1:35]  
  
# find the most important variables for each principal component  
important\_vars <- apply(  
 top\_features,  
 2,  
 function(x) names(x[order(abs(x), decreasing = TRUE)])  
 )  
  
# convert to a data frame  
important\_vars\_df <- as.data.frame(important\_vars)  
  
# most important variables  
important\_vars\_list <- unique(unlist(important\_vars\_df))  
  
important\_vars\_list

## [1] "pct\_highschool\_diploma" "life\_expectancy\_years"   
## [3] "school\_funding\_gap" "pct\_hh\_internet"   
## [5] "region.South" "median\_hh\_income.x"   
## [7] "percent\_grandparents\_as\_guardians" "pct\_adult\_smokers"   
## [9] "pct\_binge\_drinkers" "pct\_voters"   
## [11] "pct\_single\_parent" "days\_over\_90\_f"   
## [13] "pct\_black" "pct\_white"   
## [15] "pct\_employed" "inequality\_ratio"   
## [17] "pct\_under\_65\_no\_health\_insurance" "pct\_hh\_inc\_99999"   
## [19] "pct\_employed\_nonprofit" "region.Midwest"   
## [21] "dentist\_pt" "air\_polution\_metric"   
## [23] "pct\_w\_medicare" "pct\_overcrowded\_hh"   
## [25] "pa\_pt" "region.Northeast"   
## [27] "average\_hh\_size" "pct\_asian"   
## [29] "social\_clubs\_per\_10k" "pct\_30\_min\_plus\_commute"   
## [31] "pct\_rural\_population" "median\_health\_clinic\_dist"   
## [33] "region.West" "median\_trauma\_center\_dist"   
## [35] "pct\_native\_american" "pct\_home\_owner"   
## [37] "pct\_hispanic" "median\_drug\_alcohol\_care\_dist"   
## [39] "pct\_high\_housing\_costs" "weighted\_population"   
## [41] "clinical\_nurse\_pt" "median\_pediatric\_icu\_dist"   
## [43] "median\_er\_dist" "pct\_other\_race"   
## [45] "pct\_hh\_other\_computer" "population\_density"   
## [47] "water\_quality" "mental\_health\_faciliy\_pt"   
## [49] "pct\_male"

*Clean Environment*

#rm(  
# chr\_data,  
# sdoh\_data,  
# coefficients,  
# dummies\_test,  
# dummies\_train,  
# importance\_df,  
# importance\_rf,  
# model\_rf,  
# ols\_cv,  
# ols\_model,  
# pca\_components,  
# pca\_model,  
# qol\_data,  
# qol\_test,  
# qol\_test\_encoded,  
# qol\_train,  
# qol\_train\_encoded,  
# qol\_train\_standardized,  
# rf\_cv,  
# roc\_data,  
# roc\_df,  
# roc\_obj,  
# roc\_ols,  
# top\_20\_features\_ols,  
# top\_20\_importance,  
# train\_control,  
# accuracy,  
# accuracy\_ols,  
# auc\_rf,  
# confusion\_matrix,  
# feature\_importance,  
# predicted\_classes,  
# predictions,  
# predictions2,  
# split,  
# total\_missing,  
# unwanted\_chars\_results,  
# whitespace\_results,  
# unwanted\_chars,  
# whitespace\_check,  
# features,  
# important\_vars,  
# important\_vars\_df,  
# results\_psych,  
# top\_features,  
# cumulative\_explained\_variance,  
# explained\_variance,  
# important\_vars\_list,  
# median\_value,  
# num\_components\_90,  
# num\_components\_95,  
# calcSplitRatio,  
# qol\_lm  
#)

**Upon evaluation of models, it was decided that the model overfit the data, Key features were picked and evaluation was to be done as continuous response with a large feature subset to understand domain causation**

*Cleaned data reload*

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

## [1] 3229 682

# remove unwanted features, create calculated feature, convert fips\_code to data type matching chr\_data  
sdoh\_data <- sdoh\_data %>%   
 select("COUNTYFIPS",   
 "STATE",   
 "COUNTY",   
 "REGION",   
 "ACS\_TOT\_POP\_WT",   
 "ACS\_AVG\_HH\_SIZE",   
 "ACS\_PCT\_MALE",   
 "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\_OTHER\_COMP",   
 "ACS\_PCT\_HH\_INTERNET",  
 "ACS\_PCT\_EMPLOYED",  
 "ACS\_PCT\_HH\_INC\_99999",  
 "ACS\_PCT\_MEDICARE\_ONLY",  
 "AHRF\_CLIN\_NURSE\_SPEC\_RATE",   
 "AHRF\_DENTISTS\_RATE",  
 "AHRF\_PHYSICIAN\_ASSIST\_RATE",  
 "AMFAR\_MHFAC\_RATE",  
 "CEN\_POPDENSITY\_COUNTY",  
 "NEPHTN\_HEATIND\_90",  
 "SAIPE\_MEDIAN\_HH\_INCOME",  
 "POS\_MEDIAN\_DIST\_ED",  
 "POS\_MEDIAN\_DIST\_PED\_ICU",  
 "POS\_MEDIAN\_DIST\_CLINIC",   
 "POS\_MEDIAN\_DIST\_ALC",  
 "ACS\_TOT\_WORKER\_HH", #mg  
 "ACS\_PCT\_VET", #mg  
 "ACS\_PCT\_UNINSURED", #mg  
 "ACS\_PCT\_HH\_PUB\_ASSIST" #mg  
  
 ) %>%   
 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\_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\_other\_computer" = "ACS\_PCT\_HH\_OTHER\_COMP",  
 "pct\_hh\_internet" = "ACS\_PCT\_HH\_INTERNET",  
 "pct\_employed" = "ACS\_PCT\_EMPLOYED",  
 "pct\_hh\_inc\_99999" = "ACS\_PCT\_HH\_INC\_99999", # renamed by mg  
 "pct\_w\_medicare" = "ACS\_PCT\_MEDICARE\_ONLY",  
 "clinical\_nurse\_pt" = "AHRF\_CLIN\_NURSE\_SPEC\_RATE",  
 "dentist\_pt" = "AHRF\_DENTISTS\_RATE",  
 "pa\_pt" = "AHRF\_PHYSICIAN\_ASSIST\_RATE",  
 "mental\_health\_faciliy\_pt" = "AMFAR\_MHFAC\_RATE",  
 "population\_density" = "CEN\_POPDENSITY\_COUNTY",  
 "days\_over\_90\_f" = "NEPHTN\_HEATIND\_90",  
 "median\_hh\_income" = "SAIPE\_MEDIAN\_HH\_INCOME",  
 "median\_er\_dist" = "POS\_MEDIAN\_DIST\_ED",  
 "median\_pediatric\_icu\_dist" = "POS\_MEDIAN\_DIST\_PED\_ICU",  
 "median\_health\_clinic\_dist" = "POS\_MEDIAN\_DIST\_CLINIC",  
 "median\_drug\_alcohol\_care\_dist" = "POS\_MEDIAN\_DIST\_ALC",  
 "hh\_tot\_workers" = "ACS\_TOT\_WORKER\_HH", # mg  
 "pct\_vet" = "ACS\_PCT\_VET", # mg  
 "pct\_uninsured" ="ACS\_PCT\_UNINSURED", # mg   
 "pct\_assistance" = "ACS\_PCT\_HH\_PUB\_ASSIST" # mg  
  
   
 ) %>%   
 mutate(fips\_code = as.numeric(fips\_code))  
  
chr\_data <- read\_csv("data/chr\_data.csv", skip = 1)  
dim(chr\_data)

## [1] 3194 720

# remove unwanted features  
# convert principal care providers from per 100,000 people to per 1,000 people to match other data  
  
chr\_data <- chr\_data %>%  
 select("fipscode",  
 "v002\_rawvalue",  
 "v009\_rawvalue",  
 "v011\_rawvalue",  
 "v070\_rawvalue",   
 "v049\_rawvalue",  
 "v085\_rawvalue",  
 "v168\_rawvalue",   
 "v069\_rawvalue",  
 "v044\_rawvalue",   
 "v140\_rawvalue",  
 "v125\_rawvalue",  
 "v124\_rawvalue",  
 "v136\_other\_data\_1",  
 "v136\_other\_data\_2",  
 "v137\_rawvalue",  
 "v147\_rawvalue",  
 "v139\_rawvalue",  
 "v177\_rawvalue",  
 "v153\_rawvalue",  
 "v053\_rawvalue",   
 "v058\_rawvalue",   
 "v004\_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",  
 "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\_food\_insecurities" = "v139\_rawvalue",  
 "pct\_voters" = "v177\_rawvalue",  
 "pct\_home\_owner" = "v153\_rawvalue",  
 "pct\_65\_plus" = "v053\_rawvalue",  
 "pct\_rural\_population" = "v058\_rawvalue",  
 "life\_expectancy\_years" = "v147\_rawvalue",  
 "pct\_30\_min\_plus\_commute" = "v137\_rawvalue")  
  
# full data sets are extremely large, initial dimension reduction was performed previously

*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 >= 0.154, "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

# create linear model for vif analysis, no scaling, no centering, numeric values only  
qol\_numeric <- qol\_data %>%   
 select(where(is.numeric))  
  
qol\_lm <- lm(  
 pct\_poor\_to\_fair\_health ~ .,  
 qol\_numeric  
)  
  
vif\_values <- car::vif(qol\_lm)  
  
vif\_values

## fips\_code weighted\_population   
## 1.367787 389.483096   
## average\_hh\_size pct\_male   
## 3.915448 1.715235   
## pct\_native\_american pct\_asian   
## 18.999046 5.119266   
## pct\_black pct\_hispanic   
## 83.581892 6.343153   
## pct\_other\_race pct\_white   
## 7.407086 106.195952   
## pct\_single\_parent pct\_hh\_other\_computer   
## 3.111589 1.099768   
## pct\_hh\_internet pct\_employed   
## 3.598688 2.311021   
## pct\_hh\_inc\_99999 pct\_w\_medicare   
## 2.356736 2.287768   
## clinical\_nurse\_pt dentist\_pt   
## 1.158814 1.877711   
## pa\_pt mental\_health\_faciliy\_pt   
## 1.659227 1.161591   
## population\_density days\_over\_90\_f   
## 1.395979 3.383845   
## median\_hh\_income median\_er\_dist   
## 6.491325 1.592515   
## median\_pediatric\_icu\_dist median\_health\_clinic\_dist   
## 1.657468 1.178996   
## median\_drug\_alcohol\_care\_dist hh\_tot\_workers   
## 1.676041 392.722923   
## pct\_vet pct\_uninsured   
## 1.837037 5.588217   
## pct\_assistance percent\_grandparents\_as\_guardians   
## 4.748838 2.057071   
## pct\_adult\_smokers pct\_obese\_adults   
## 11.134324 4.289091   
## pct\_no\_exercise pct\_binge\_drinkers   
## 11.633357 2.309359   
## pct\_under\_65\_no\_health\_insurance pct\_highschool\_diploma   
## 6.929268 6.145603   
## pct\_some\_college inequality\_ratio   
## 4.550463 2.530376   
## social\_clubs\_per\_10k air\_polution\_metric   
## 1.733348 2.093544   
## water\_quality pct\_high\_housing\_costs   
## 1.110293 2.719547   
## pct\_overcrowded\_hh pct\_30\_min\_plus\_commute   
## 2.673865 2.776362   
## life\_expectancy\_years pct\_food\_insecurities   
## 3.721053 5.222556   
## pct\_voters pct\_home\_owner   
## 3.818977 4.109748   
## pct\_65\_plus pct\_rural\_population   
## 5.627477 3.798256   
## pcp\_pt   
## 2.285573

# Create a data frame with GVIF and Df  
vif\_df <- data.frame(  
 Feature = names(vif\_values),  
 GVIF = vif\_values,  
 Df = rep(1, length(vif\_values)) # Df is typically 1 for univariate cases  
)  
  
# Calculate GVIF^(1/(2\*Df)) and add it to the data frame  
vif\_df$Adjusted\_VIF <- vif\_df$GVIF^(1/(2 \* vif\_df$Df))  
  
# Sort the data frame by Adjusted VIF in ascending order  
vif\_df <- vif\_df[order(vif\_df$Adjusted\_VIF, decreasing = TRUE), ]  
  
# Set a threshold for high VIF  
high\_vif\_threshold <- 5  
  
# Filter for features with high VIF values  
high\_vif\_features <- vif\_df[vif\_df[, "Adjusted\_VIF"] > high\_vif\_threshold, ]  
  
# Print the table  
# Add a column indicating if the Adjusted VIF is above the threshold  
vif\_df <- high\_vif\_features %>%  
 mutate(High\_VIF = if\_else(Adjusted\_VIF > high\_vif\_threshold, "Yes", "No"))  
  
# Print the table  
vif\_df

## Feature GVIF Df Adjusted\_VIF High\_VIF  
## hh\_tot\_workers hh\_tot\_workers 392.72292 1 19.817238 Yes  
## weighted\_population weighted\_population 389.48310 1 19.735326 Yes  
## pct\_white pct\_white 106.19595 1 10.305142 Yes  
## pct\_black pct\_black 83.58189 1 9.142313 Yes

*VIF Analysis*

# create linear model for vif analysis, no scaling, no centering, numeric values only  
qol\_numeric <- qol\_data %>%   
 select(where(is.numeric))  
  
qol\_lm <- lm(  
 pct\_poor\_to\_fair\_health ~ .,  
 qol\_numeric  
)  
  
vif\_values <- car::vif(qol\_lm)  
  
vif\_values

## fips\_code weighted\_population   
## 1.367787 389.483096   
## average\_hh\_size pct\_male   
## 3.915448 1.715235   
## pct\_native\_american pct\_asian   
## 18.999046 5.119266   
## pct\_black pct\_hispanic   
## 83.581892 6.343153   
## pct\_other\_race pct\_white   
## 7.407086 106.195952   
## pct\_single\_parent pct\_hh\_other\_computer   
## 3.111589 1.099768   
## pct\_hh\_internet pct\_employed   
## 3.598688 2.311021   
## pct\_hh\_inc\_99999 pct\_w\_medicare   
## 2.356736 2.287768   
## clinical\_nurse\_pt dentist\_pt   
## 1.158814 1.877711   
## pa\_pt mental\_health\_faciliy\_pt   
## 1.659227 1.161591   
## population\_density days\_over\_90\_f   
## 1.395979 3.383845   
## median\_hh\_income median\_er\_dist   
## 6.491325 1.592515   
## median\_pediatric\_icu\_dist median\_health\_clinic\_dist   
## 1.657468 1.178996   
## median\_drug\_alcohol\_care\_dist hh\_tot\_workers   
## 1.676041 392.722923   
## pct\_vet pct\_uninsured   
## 1.837037 5.588217   
## pct\_assistance percent\_grandparents\_as\_guardians   
## 4.748838 2.057071   
## pct\_adult\_smokers pct\_obese\_adults   
## 11.134324 4.289091   
## pct\_no\_exercise pct\_binge\_drinkers   
## 11.633357 2.309359   
## pct\_under\_65\_no\_health\_insurance pct\_highschool\_diploma   
## 6.929268 6.145603   
## pct\_some\_college inequality\_ratio   
## 4.550463 2.530376   
## social\_clubs\_per\_10k air\_polution\_metric   
## 1.733348 2.093544   
## water\_quality pct\_high\_housing\_costs   
## 1.110293 2.719547   
## pct\_overcrowded\_hh pct\_30\_min\_plus\_commute   
## 2.673865 2.776362   
## life\_expectancy\_years pct\_food\_insecurities   
## 3.721053 5.222556   
## pct\_voters pct\_home\_owner   
## 3.818977 4.109748   
## pct\_65\_plus pct\_rural\_population   
## 5.627477 3.798256   
## pcp\_pt   
## 2.285573

# Create a data frame with GVIF and Df  
vif\_df <- data.frame(  
 Feature = names(vif\_values),  
 GVIF = vif\_values,  
 Df = rep(1, length(vif\_values)) # Df is typically 1 for univariate cases  
)  
  
# Calculate GVIF^(1/(2\*Df)) and add it to the data frame  
vif\_df$Adjusted\_VIF <- vif\_df$GVIF^(1/(2 \* vif\_df$Df))  
  
# Sort the data frame by Adjusted VIF in ascending order  
vif\_df <- vif\_df[order(vif\_df$Adjusted\_VIF, decreasing = TRUE), ]  
  
# Set a threshold for high VIF  
high\_vif\_threshold <- 5  
  
# Filter for features with high VIF values  
high\_vif\_features <- vif\_df[vif\_df[, "Adjusted\_VIF"] > high\_vif\_threshold, ]  
  
# Print the table  
# Add a column indicating if the Adjusted VIF is above the threshold  
vif\_df <- vif\_df %>%  
 mutate(High\_VIF = if\_else(Adjusted\_VIF > high\_vif\_threshold, "Yes", "No"))  
  
# Print the table  
vif\_df

## Feature GVIF  
## hh\_tot\_workers hh\_tot\_workers 392.722923  
## weighted\_population weighted\_population 389.483096  
## pct\_white pct\_white 106.195952  
## pct\_black pct\_black 83.581892  
## pct\_native\_american pct\_native\_american 18.999046  
## pct\_no\_exercise pct\_no\_exercise 11.633357  
## pct\_adult\_smokers pct\_adult\_smokers 11.134324  
## pct\_other\_race pct\_other\_race 7.407086  
## pct\_under\_65\_no\_health\_insurance pct\_under\_65\_no\_health\_insurance 6.929268  
## median\_hh\_income median\_hh\_income 6.491325  
## pct\_hispanic pct\_hispanic 6.343153  
## pct\_highschool\_diploma pct\_highschool\_diploma 6.145603  
## pct\_65\_plus pct\_65\_plus 5.627477  
## pct\_uninsured pct\_uninsured 5.588217  
## pct\_food\_insecurities pct\_food\_insecurities 5.222556  
## pct\_asian pct\_asian 5.119266  
## pct\_assistance pct\_assistance 4.748838  
## pct\_some\_college pct\_some\_college 4.550463  
## pct\_obese\_adults pct\_obese\_adults 4.289091  
## pct\_home\_owner pct\_home\_owner 4.109748  
## average\_hh\_size average\_hh\_size 3.915448  
## pct\_voters pct\_voters 3.818977  
## pct\_rural\_population pct\_rural\_population 3.798256  
## life\_expectancy\_years life\_expectancy\_years 3.721053  
## pct\_hh\_internet pct\_hh\_internet 3.598688  
## days\_over\_90\_f days\_over\_90\_f 3.383845  
## pct\_single\_parent pct\_single\_parent 3.111589  
## pct\_30\_min\_plus\_commute pct\_30\_min\_plus\_commute 2.776362  
## pct\_high\_housing\_costs pct\_high\_housing\_costs 2.719547  
## pct\_overcrowded\_hh pct\_overcrowded\_hh 2.673865  
## inequality\_ratio inequality\_ratio 2.530376  
## pct\_hh\_inc\_99999 pct\_hh\_inc\_99999 2.356736  
## pct\_employed pct\_employed 2.311021  
## pct\_binge\_drinkers pct\_binge\_drinkers 2.309359  
## pct\_w\_medicare pct\_w\_medicare 2.287768  
## pcp\_pt pcp\_pt 2.285573  
## air\_polution\_metric air\_polution\_metric 2.093544  
## percent\_grandparents\_as\_guardians percent\_grandparents\_as\_guardians 2.057071  
## dentist\_pt dentist\_pt 1.877711  
## pct\_vet pct\_vet 1.837037  
## social\_clubs\_per\_10k social\_clubs\_per\_10k 1.733348  
## pct\_male pct\_male 1.715235  
## median\_drug\_alcohol\_care\_dist median\_drug\_alcohol\_care\_dist 1.676041  
## pa\_pt pa\_pt 1.659227  
## median\_pediatric\_icu\_dist median\_pediatric\_icu\_dist 1.657468  
## median\_er\_dist median\_er\_dist 1.592515  
## population\_density population\_density 1.395979  
## fips\_code fips\_code 1.367787  
## median\_health\_clinic\_dist median\_health\_clinic\_dist 1.178996  
## mental\_health\_faciliy\_pt mental\_health\_faciliy\_pt 1.161591  
## clinical\_nurse\_pt clinical\_nurse\_pt 1.158814  
## water\_quality water\_quality 1.110293  
## pct\_hh\_other\_computer pct\_hh\_other\_computer 1.099768  
## Df Adjusted\_VIF High\_VIF  
## hh\_tot\_workers 1 19.817238 Yes  
## weighted\_population 1 19.735326 Yes  
## pct\_white 1 10.305142 Yes  
## pct\_black 1 9.142313 Yes  
## pct\_native\_american 1 4.358790 No  
## pct\_no\_exercise 1 3.410771 No  
## pct\_adult\_smokers 1 3.336814 No  
## pct\_other\_race 1 2.721596 No  
## pct\_under\_65\_no\_health\_insurance 1 2.632350 No  
## median\_hh\_income 1 2.547808 No  
## pct\_hispanic 1 2.518562 No  
## pct\_highschool\_diploma 1 2.479033 No  
## pct\_65\_plus 1 2.372230 No  
## pct\_uninsured 1 2.363941 No  
## pct\_food\_insecurities 1 2.285291 No  
## pct\_asian 1 2.262579 No  
## pct\_assistance 1 2.179183 No  
## pct\_some\_college 1 2.133181 No  
## pct\_obese\_adults 1 2.071012 No  
## pct\_home\_owner 1 2.027251 No  
## average\_hh\_size 1 1.978749 No  
## pct\_voters 1 1.954220 No  
## pct\_rural\_population 1 1.948912 No  
## life\_expectancy\_years 1 1.929003 No  
## pct\_hh\_internet 1 1.897021 No  
## days\_over\_90\_f 1 1.839523 No  
## pct\_single\_parent 1 1.763970 No  
## pct\_30\_min\_plus\_commute 1 1.666242 No  
## pct\_high\_housing\_costs 1 1.649105 No  
## pct\_overcrowded\_hh 1 1.635196 No  
## inequality\_ratio 1 1.590716 No  
## pct\_hh\_inc\_99999 1 1.535166 No  
## pct\_employed 1 1.520204 No  
## pct\_binge\_drinkers 1 1.519658 No  
## pct\_w\_medicare 1 1.512537 No  
## pcp\_pt 1 1.511811 No  
## air\_polution\_metric 1 1.446909 No  
## percent\_grandparents\_as\_guardians 1 1.434249 No  
## dentist\_pt 1 1.370296 No  
## pct\_vet 1 1.355373 No  
## social\_clubs\_per\_10k 1 1.316567 No  
## pct\_male 1 1.309670 No  
## median\_drug\_alcohol\_care\_dist 1 1.294620 No  
## pa\_pt 1 1.288110 No  
## median\_pediatric\_icu\_dist 1 1.287427 No  
## median\_er\_dist 1 1.261949 No  
## population\_density 1 1.181516 No  
## fips\_code 1 1.169524 No  
## median\_health\_clinic\_dist 1 1.085816 No  
## mental\_health\_faciliy\_pt 1 1.077771 No  
## clinical\_nurse\_pt 1 1.076482 No  
## water\_quality 1 1.053704 No  
## pct\_hh\_other\_computer 1 1.048698 No

# eliminate variables with no predictive value: fipscode, county and state:  
qol\_data <- qol\_data %>%  
 select(  
 -c(  
 fips\_code,  
 county,  
 state,  
 pct\_no\_exercise,  
 pct\_obese\_adults,  
 life\_expectancy\_years  
 )  
 )

*Train, Test, and fill*

# one hot encode character Variables,center and scale all predictors except one hot encoded variables.  
  
set.seed(12)  
  
# Create a recipe  
  
response\_recipe <- recipe(  
 pct\_poor\_to\_fair\_health ~ .,  
 data = qol\_train  
 ) %>%  
 # One-hot encode the 'region' feature to include all 4 regions  
 step\_dummy(  
 region,  
 one\_hot = TRUE  
 ) %>%   
 # Center all predictors except the one-hot encoded 'region' columns  
 step\_center(  
 all\_predictors(),  
 -starts\_with("region\_")  
 ) %>%   
 # Scale all predictors except the one-hot encoded 'region' columns  
 step\_scale(  
 all\_predictors(),  
 -starts\_with("region\_")  
 ) %>%   
  
 prep(  
 training = qol\_train,  
 retain = TRUE  
 )  
  
# Apply the recipe to the training and testing datasets  
qol\_train <- bake(response\_recipe, new\_data = qol\_train)  
qol\_test <- bake(response\_recipe, new\_data = qol\_test)  
  
head(qol\_test)

## # A tibble: 6 × 53  
## weighted\_population average\_hh\_size pct\_male pct\_native\_american pct\_asian  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.366 0.246 -0.649 -0.166 -0.158  
## 2 -0.239 1.30 1.47 -0.246 -0.399  
## 3 0.0452 0.0273 -0.840 -0.213 -0.201  
## 4 -0.282 0.209 -1.09 -0.194 -0.488  
## 5 -0.216 -0.0454 -0.247 -0.194 -0.467  
## 6 0.970 0.318 -1.01 -0.169 0.175  
## # ℹ 48 more variables: pct\_black <dbl>, pct\_hispanic <dbl>,  
## # pct\_other\_race <dbl>, pct\_white <dbl>, pct\_single\_parent <dbl>,  
## # pct\_hh\_other\_computer <dbl>, pct\_hh\_internet <dbl>, pct\_employed <dbl>,  
## # pct\_hh\_inc\_99999 <dbl>, pct\_w\_medicare <dbl>, clinical\_nurse\_pt <dbl>,  
## # dentist\_pt <dbl>, pa\_pt <dbl>, mental\_health\_faciliy\_pt <dbl>,  
## # population\_density <dbl>, days\_over\_90\_f <dbl>, median\_hh\_income <dbl>,  
## # median\_er\_dist <dbl>, median\_pediatric\_icu\_dist <dbl>, …

*Linear Model*

set.seed(42)  
# Fit the OLS model  
ols\_model <- lm(  
 pct\_poor\_to\_fair\_health ~ .,  
 data = qol\_train  
 )  
  
# Print OLS coefficients  
print(summary(ols\_model)$coefficients)

## Estimate Std. Error t value  
## (Intercept) 1.631698e-01 0.0008603687 189.65101627  
## weighted\_population -3.554968e-03 0.0044319957 -0.80211455  
## average\_hh\_size 2.173630e-03 0.0003948529 5.50491158  
## pct\_male -4.108013e-04 0.0002605797 -1.57648989  
## pct\_native\_american -1.190514e-03 0.0009330662 -1.27591592  
## pct\_asian 1.740648e-04 0.0004821605 0.36101016  
## pct\_black 6.756999e-03 0.0014476294 4.66763021  
## pct\_hispanic 1.037058e-02 0.0004696572 22.08118112  
## pct\_other\_race -3.867298e-04 0.0004681957 -0.82600031  
## pct\_white -2.032211e-05 0.0016969542 -0.01197564  
## pct\_single\_parent 8.071953e-04 0.0003527677 2.28817816  
## pct\_hh\_other\_computer -2.619388e-04 0.0002194899 -1.19339794  
## pct\_hh\_internet -1.311084e-03 0.0003810422 -3.44078495  
## pct\_employed -1.620341e-04 0.0003163026 -0.51227558  
## pct\_hh\_inc\_99999 -2.378838e-03 0.0003036647 -7.83376537  
## pct\_w\_medicare -5.097112e-04 0.0003073571 -1.65836800  
## clinical\_nurse\_pt -3.714287e-04 0.0002209739 -1.68087126  
## dentist\_pt 2.546083e-04 0.0002807244 0.90696905  
## pa\_pt -6.473394e-04 0.0002449297 -2.64296006  
## mental\_health\_faciliy\_pt 5.333537e-04 0.0002266077 2.35364349  
## population\_density -5.456850e-04 0.0002462891 -2.21562788  
## days\_over\_90\_f 1.574312e-03 0.0003845475 4.09393504  
## median\_hh\_income -2.546736e-03 0.0004880927 -5.21773001  
## median\_er\_dist -9.266861e-04 0.0003193877 -2.90144543  
## median\_pediatric\_icu\_dist 4.190861e-05 0.0002993551 0.13999630  
## median\_health\_clinic\_dist -8.203840e-04 0.0002566508 -3.19649908  
## median\_drug\_alcohol\_care\_dist 1.670328e-03 0.0003143288 5.31395169  
## hh\_tot\_workers 3.107802e-03 0.0044591736 0.69694571  
## pct\_vet -1.995943e-03 0.0002661143 -7.50032009  
## pct\_uninsured 1.604472e-05 0.0004703284 0.03411388  
## pct\_assistance -6.055854e-04 0.0004475644 -1.35306866  
## percent\_grandparents\_as\_guardians 1.272539e-04 0.0002935846 0.43344889  
## pct\_adult\_smokers 2.058441e-02 0.0006017111 34.20979636  
## pct\_binge\_drinkers -3.782358e-03 0.0003182414 -11.88518737  
## pct\_under\_65\_no\_health\_insurance -2.844083e-03 0.0005141707 -5.53139905  
## pct\_highschool\_diploma -5.653875e-03 0.0004786748 -11.81151597  
## pct\_some\_college -1.903680e-05 0.0004375243 -0.04351027  
## inequality\_ratio 2.605501e-04 0.0003111952 0.83725624  
## social\_clubs\_per\_10k -1.045761e-03 0.0002614615 -3.99967313  
## air\_polution\_metric 1.919220e-03 0.0002973562 6.45427916  
## water\_quality 6.170447e-04 0.0002208626 2.79379413  
## pct\_high\_housing\_costs -7.684447e-04 0.0003201152 -2.40052559  
## pct\_overcrowded\_hh 1.750911e-03 0.0003693719 4.74023823  
## pct\_30\_min\_plus\_commute -1.069786e-03 0.0003328253 -3.21425687  
## pct\_food\_insecurities 6.830219e-03 0.0004400178 15.52259576  
## pct\_voters -7.758830e-04 0.0003890266 -1.99442179  
## pct\_home\_owner 3.902177e-04 0.0003648637 1.06948898  
## pct\_65\_plus 6.680924e-04 0.0004436399 1.50593377  
## pct\_rural\_population 8.652139e-04 0.0003938698 2.19670030  
## region\_Midwest -8.932044e-03 0.0010320947 -8.65428723  
## region\_Northeast -9.489471e-03 0.0011687795 -8.11912820  
## region\_South 1.977899e-03 0.0010847792 1.82331956  
## Pr(>|t|)  
## (Intercept) 0.000000e+00  
## weighted\_population 4.225587e-01  
## average\_hh\_size 4.048087e-08  
## pct\_male 1.150324e-01  
## pct\_native\_american 2.020971e-01  
## pct\_asian 7.181207e-01  
## pct\_black 3.198667e-06  
## pct\_hispanic 2.585123e-99  
## pct\_other\_race 4.088783e-01  
## pct\_white 9.904460e-01  
## pct\_single\_parent 2.220561e-02  
## pct\_hh\_other\_computer 2.328205e-01  
## pct\_hh\_internet 5.890126e-04  
## pct\_employed 6.085009e-01  
## pct\_hh\_inc\_99999 6.796772e-15  
## pct\_w\_medicare 9.736165e-02  
## clinical\_nurse\_pt 9.290586e-02  
## dentist\_pt 3.645057e-01  
## pa\_pt 8.266897e-03  
## mental\_health\_faciliy\_pt 1.866314e-02  
## population\_density 2.680176e-02  
## days\_over\_90\_f 4.368113e-05  
## median\_hh\_income 1.951468e-07  
## median\_er\_dist 3.745105e-03  
## median\_pediatric\_icu\_dist 8.886736e-01  
## median\_health\_clinic\_dist 1.407441e-03  
## median\_drug\_alcohol\_care\_dist 1.162054e-07  
## hh\_tot\_workers 4.858979e-01  
## pct\_vet 8.636881e-14  
## pct\_uninsured 9.727889e-01  
## pct\_assistance 1.761492e-01  
## percent\_grandparents\_as\_guardians 6.647240e-01  
## pct\_adult\_smokers 9.835304e-213  
## pct\_binge\_drinkers 8.945597e-32  
## pct\_under\_65\_no\_health\_insurance 3.487801e-08  
## pct\_highschool\_diploma 2.061294e-31  
## pct\_some\_college 9.652981e-01  
## inequality\_ratio 4.025241e-01  
## social\_clubs\_per\_10k 6.516687e-05  
## air\_polution\_metric 1.288669e-10  
## water\_quality 5.246838e-03  
## pct\_high\_housing\_costs 1.644015e-02  
## pct\_overcrowded\_hh 2.247611e-06  
## pct\_30\_min\_plus\_commute 1.323544e-03  
## pct\_food\_insecurities 4.482095e-52  
## pct\_voters 4.620834e-02  
## pct\_home\_owner 2.849469e-01  
## pct\_65\_plus 1.322034e-01  
## pct\_rural\_population 2.812823e-02  
## region\_Midwest 8.457803e-18  
## region\_Northeast 7.116239e-16  
## region\_South 6.836769e-02

# Predict on the test set  
predictions\_ols <- predict(  
 ols\_model,  
 new\_data = qol\_test  
 )  
  
# Evaluate the OLS model  
mse\_ols <- mean((qol\_test$pct\_poor\_to\_fair\_health - predictions\_ols)^2)

## Warning in qol\_test$pct\_poor\_to\_fair\_health - predictions\_ols: longer object  
## length is not a multiple of shorter object length

rmse\_ols <- sqrt(mse\_ols)  
r2\_ols <- summary(ols\_model)$r.squared  
  
print(paste("MSE (OLS):", mse\_ols))

## [1] "MSE (OLS): 0.00391915285974174"

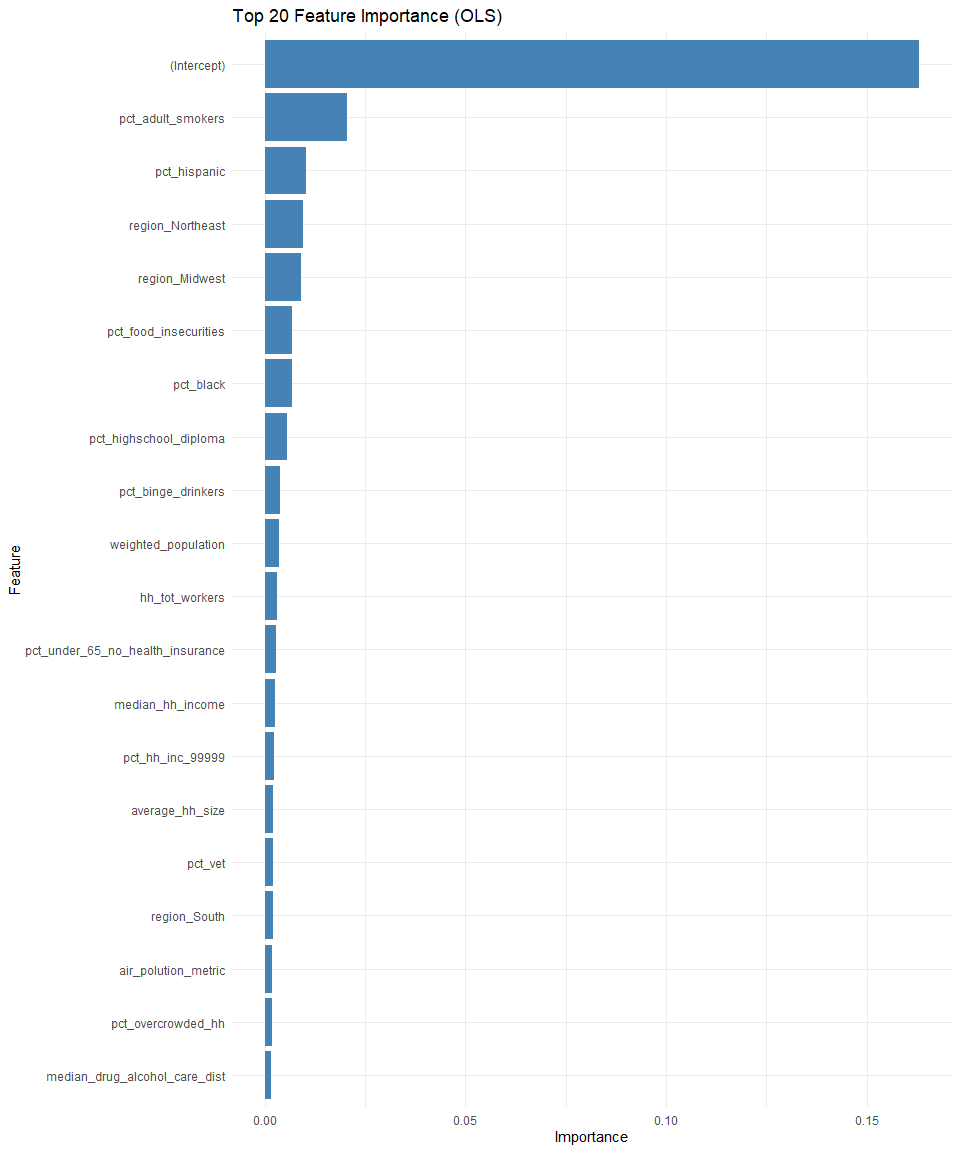
print(paste("RMSE (OLS):", rmse\_ols))

## [1] "RMSE (OLS): 0.0626031377787227"

print(paste("R-squared (OLS):", r2\_ols))

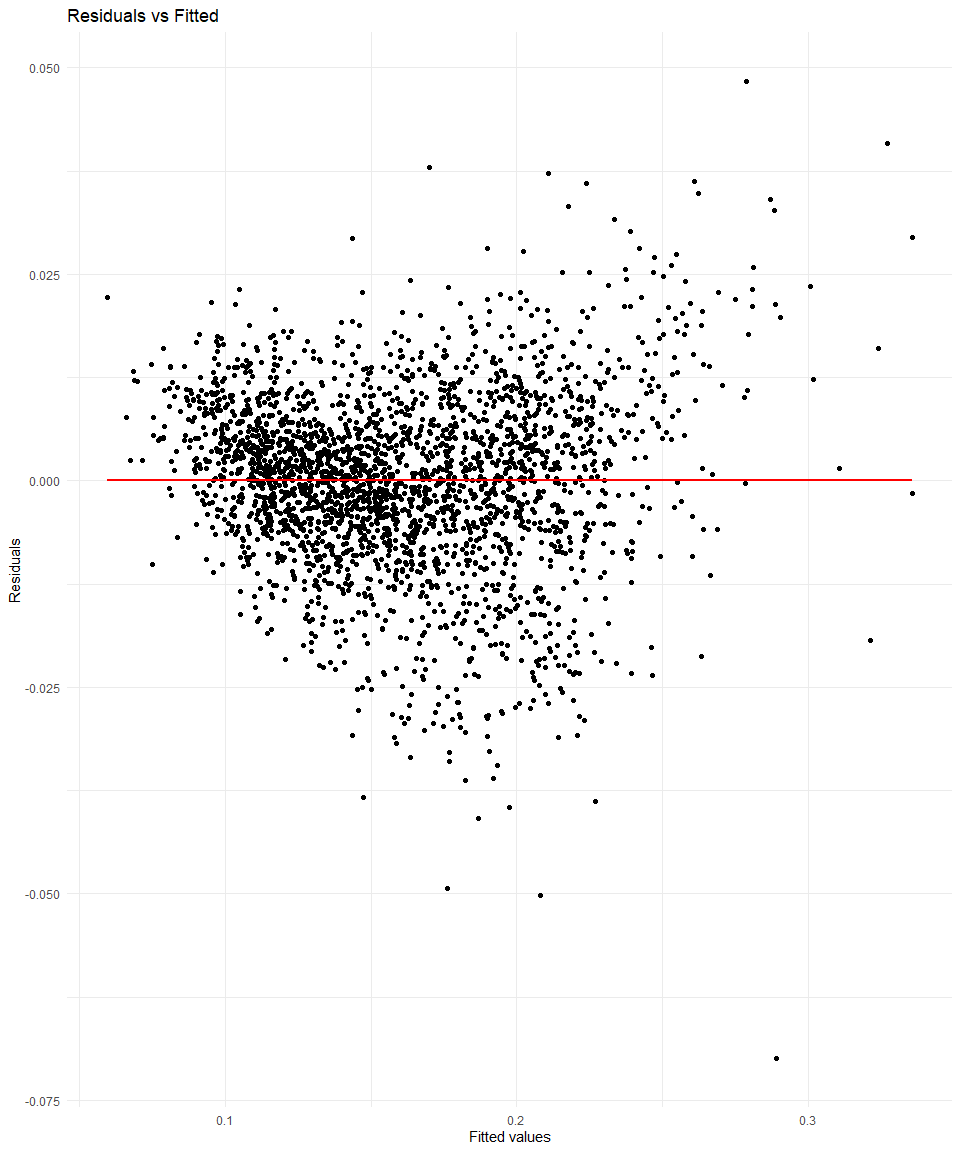
## [1] "R-squared (OLS): 0.941997472989246"

# --- Feature Importance Visualization (OLS) ---  
  
# Calculate feature importance for OLS  
feature\_importance\_ols <- abs(summary(ols\_model)$coefficients[, "Estimate"])  
  
importance\_df\_ols <- data.frame(  
 Feature = names(feature\_importance\_ols),  
 Importance = feature\_importance\_ols  
 )  
  
importance\_df\_ols <- importance\_df\_ols[order(-importance\_df\_ols$Importance), ]  
  
# Select top 20 features  
top\_20\_features\_ols <- head(importance\_df\_ols, 20)  
  
# Plot top 20 feature importance for OLS  
ggplot(  
 top\_20\_features\_ols,  
 aes(  
 x = reorder(Feature, Importance),  
 y = Importance)) +  
 geom\_bar(  
 stat = "identity",  
 fill = "steelblue"  
 ) +  
 coord\_flip() +  
 labs(  
 title = "Top 20 Feature Importance (OLS)",  
 x = "Feature",  
 y = "Importance") +  
 theme\_minimal()

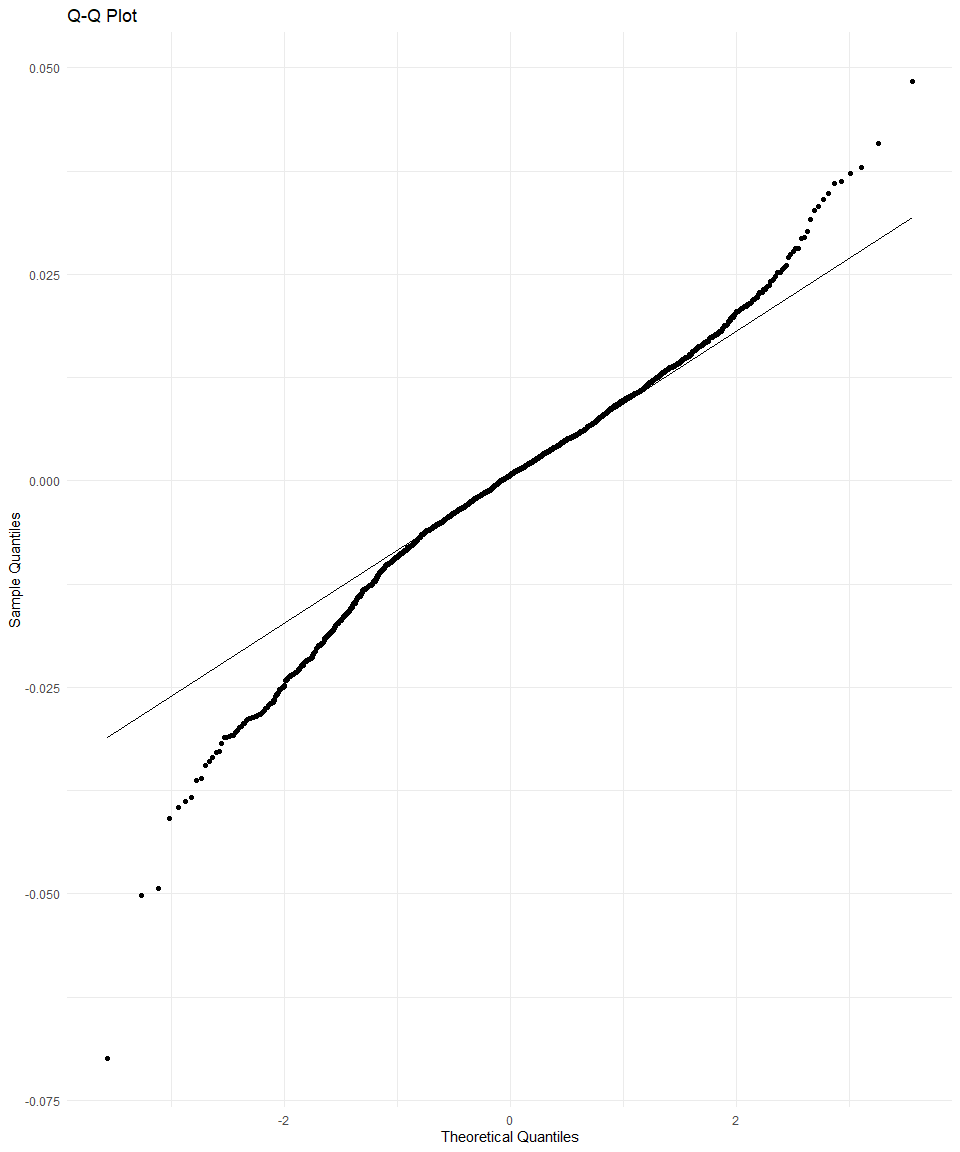


# --- Assumptions Diagnostics for OLS Model ---  
  
# Residuals plot  
ggplot(data.frame(residuals = residuals(ols\_model),  
 fitted = fitted(ols\_model)),  
 aes(x = fitted, y = residuals)) +  
 geom\_point() +  
 geom\_smooth(method = "lm",  
 se = FALSE,  
 color = "red") +  
 labs(title = "Residuals vs Fitted",  
 x = "Fitted values",  
 y = "Residuals") +  
 theme\_minimal()

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



# Q-Q plot  
ggplot(data.frame(sample = residuals(ols\_model)),  
 aes(sample = sample)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 labs(title = "Q-Q Plot",  
 x = "Theoretical Quantiles",  
 y = "Sample Quantiles") +  
 theme\_minimal()



#Durbin-Watson test  
lmtest::dwtest(ols\_model)

## Warning in sqrt(dvar): NaNs produced

##   
## Durbin-Watson test  
##   
## data: ols\_model  
## DW = 0.79957, p-value = NA  
## alternative hypothesis: true autocorrelation is greater than 0

# Perform the Breusch-Godfrey test for autocorrelation  
bg\_test <- lmtest::bgtest(  
 ols\_model,  
 order = 1 # You can change the order as needed  
 )   
  
# Print the test results  
print(bg\_test)

##   
## Breusch-Godfrey test for serial correlation of order up to 1  
##   
## data: ols\_model  
## LM test = 1037.1, df = 1, p-value < 2.2e-16

set.seed(42)  
# Create design matrices (excluding response variable)  
x\_train <- model.matrix(  
 pct\_poor\_to\_fair\_health ~ .,  
 data = qol\_train)[, -1]  
  
x\_test <- model.matrix(  
 pct\_poor\_to\_fair\_health ~ .,  
 data = qol\_test)[, -1]  
  
# Fit the Lasso model  
lasso\_model <- glmnet::glmnet(  
 x\_train,  
 qol\_train$pct\_poor\_to\_fair\_health,  
 alpha = 1  
 )  
  
# Find optimal lambda using cross-validation  
cv\_lasso <- glmnet::cv.glmnet(  
 x\_train,  
 qol\_train$pct\_poor\_to\_fair\_health,  
 alpha = 1  
 )  
  
# Predict on the test set using the optimal lambda  
predictions\_lasso <- predict(lasso\_model, newx = x\_test, s = cv\_lasso$lambda.min)  
  
# Evaluate the Lasso model  
mse\_lasso <- mean((qol\_test$pct\_poor\_to\_fair\_health - predictions\_lasso)^2)  
rmse\_lasso <- sqrt(mse\_lasso)  
r2\_lasso <- 1 - (sum((qol\_test$pct\_poor\_to\_fair\_health - predictions\_lasso)^2) / sum((qol\_test$pct\_poor\_to\_fair\_health - mean(qol\_test$pct\_poor\_to\_fair\_health))^2))  
print(paste("MSE (Lasso):", mse\_lasso))

## [1] "MSE (Lasso): 0.000125332655001026"

print(paste("RMSE (Lasso):", rmse\_lasso))

## [1] "RMSE (Lasso): 0.011195206786881"

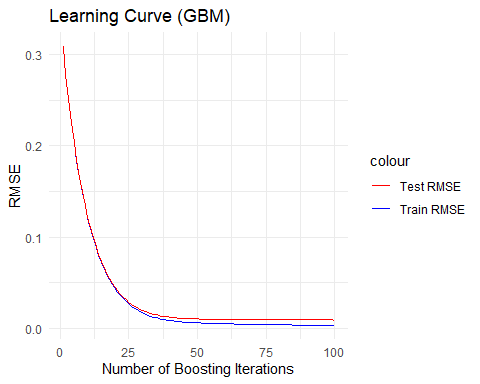
print(paste("R-squared (Lasso):", r2\_lasso))

## [1] "R-squared (Lasso): 0.938981940311687"

set.seed(42)  
# Prepare data for xgboost  
dtrain <- xgboost::xgb.DMatrix(  
 data = as.matrix(  
 qol\_train[, -which(names(qol\_train) == "pct\_poor\_to\_fair\_health")]  
 ),  
 label = qol\_train$pct\_poor\_to\_fair\_health)  
  
dtest <- xgboost::xgb.DMatrix(  
 data = as.matrix(  
 qol\_test[, -which(names(qol\_test) == "pct\_poor\_to\_fair\_health")]  
 ),  
 label = qol\_test$pct\_poor\_to\_fair\_health)  
  
# Set parameters for xgboost  
params <- list(  
 objective = "reg:squarederror",  
 eta = 0.1,  
 max\_depth = 6,  
 subsample = 0.8,  
 colsample\_bytree = 0.8  
)  
  
# Train the GBM model with evaluation metrics at each iteration  
gbm\_model <- xgboost::xgb.train(  
 params,  
 dtrain,  
 nrounds = 100,  
 watchlist = list(train = dtrain, eval = dtest),  
 print\_every\_n = 10,  
 eval\_metric = "rmse",  
 early\_stopping\_rounds = 10  
)

## [1] train-rmse:0.308474 eval-rmse:0.308643   
## Multiple eval metrics are present. Will use eval\_rmse for early stopping.  
## Will train until eval\_rmse hasn't improved in 10 rounds.  
##   
## [11] train-rmse:0.109068 eval-rmse:0.110050   
## [21] train-rmse:0.039527 eval-rmse:0.041200   
## [31] train-rmse:0.015850 eval-rmse:0.018473   
## [41] train-rmse:0.008330 eval-rmse:0.011926   
## [51] train-rmse:0.006060 eval-rmse:0.010398   
## [61] train-rmse:0.005148 eval-rmse:0.009864   
## [71] train-rmse:0.004494 eval-rmse:0.009583   
## [81] train-rmse:0.003957 eval-rmse:0.009484   
## [91] train-rmse:0.003591 eval-rmse:0.009377   
## [100] train-rmse:0.003284 eval-rmse:0.009319

# Extract the evaluation metrics from the watchlist  
eval\_metrics <- gbm\_model$evaluation\_log  
  
# Create a data frame for plotting the learning curve  
learning\_curve <- data.frame(  
 Iteration = 1:nrow(eval\_metrics),  
 Train\_RMSE = eval\_metrics$train\_rmse,  
 Test\_RMSE = eval\_metrics$eval\_rmse  
)  
  
# Plot the learning curve  
ggplot(  
 learning\_curve,  
 aes(x = Iteration)  
 ) +  
 geom\_line(  
 aes(  
 y = Train\_RMSE,  
 color = "Train RMSE")  
 ) +  
 geom\_line(  
 aes(  
 y = Test\_RMSE,  
 color = "Test RMSE"  
 )  
 ) +  
 labs(  
 title = "Learning Curve (GBM)",  
 x = "Number of Boosting Iterations",  
 y = "RMSE"  
 ) +  
 theme\_minimal() +  
 scale\_color\_manual(  
 values = c(  
 "Train RMSE" = "blue",  
 "Test RMSE" = "red"  
 )  
 )



# Make predictions  
predictions\_gbm <- predict(gbm\_model, dtest)  
  
# Evaluate the GBM model  
mse\_gbm <- mean((qol\_test$pct\_poor\_to\_fair\_health - predictions\_gbm)^2)  
rmse\_gbm <- sqrt(mse\_gbm)  
r2\_gbm <- 1 - (sum((qol\_test$pct\_poor\_to\_fair\_health - predictions\_gbm)^2) / sum((qol\_test$pct\_poor\_to\_fair\_health - mean(qol\_test$pct\_poor\_to\_fair\_health))^2))  
  
# Print the evaluation metrics  
print(paste("MSE (GBM):", mse\_gbm))

## [1] "MSE (GBM): 8.68445996745972e-05"

print(paste("RMSE (GBM):", rmse\_gbm))

## [1] "RMSE (GBM): 0.00931904499799186"

print(paste("R-squared (GBM):", r2\_gbm))

## [1] "R-squared (GBM): 0.95771980601138"

set.seed(42)  
# Train a random forest model for classification  
model\_rf <- randomForest::randomForest(  
 pct\_poor\_to\_fair\_health ~ .,  
 data = qol\_train,  
 ntree = 500  
 )  
  
# Predict the response variable for the testing set  
predictions\_rf <- predict(  
 model\_rf,  
 qol\_test,  
 type = "response"  
 )  
  
# Calculate Mean Squared Error (MSE) for Random Forest model  
mse\_rf <- mean(  
 (qol\_test$pct\_poor\_to\_fair\_health - predictions\_rf)^2  
 )  
  
# Calculate Root Mean Squared Error (RMSE) for Random Forest model  
rmse\_rf <- sqrt(mse\_rf)  
  
# Calculate R-squared (R^2) for Random Forest model  
ss\_total <- sum(  
 (qol\_test$pct\_poor\_to\_fair\_health - mean(qol\_train$pct\_poor\_to\_fair\_health))^2  
 )  
ss\_residual <- sum(  
 (qol\_test$pct\_poor\_to\_fair\_health - predictions\_rf)^2  
 )  
r2\_rf <- 1 - (ss\_residual / ss\_total)  
  
# Print the results  
print(paste("MSE (Random Forest):", mse\_rf))

## [1] "MSE (Random Forest): 0.00012289285803182"

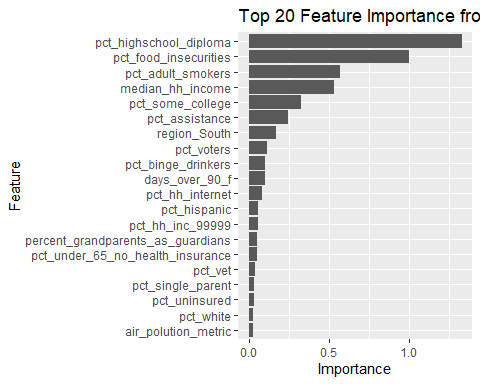
print(paste("RMSE (Random Forest):", rmse\_rf))

## [1] "RMSE (Random Forest): 0.0110857051210927"

print(paste("R-squared (Random Forest):", r2\_rf))

## [1] "R-squared (Random Forest): 0.940169931412358"

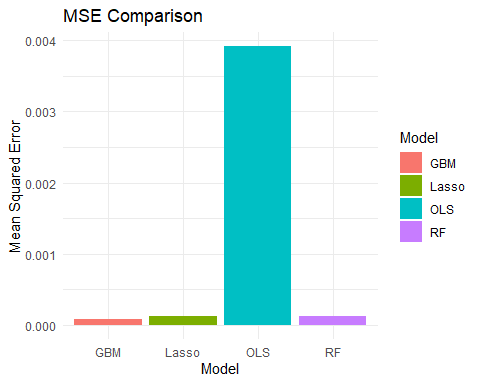
# Calculate feature importance  
importance\_rf <- randomForest::importance(model\_rf)  
importance\_df <- data.frame(  
 Feature = rownames(importance\_rf),  
 Importance = importance\_rf[, 1])  
  
# Select top 20 important features  
top\_20\_importance <- importance\_df[order(-importance\_df$Importance), ][1:20, ]  
  
# Plot top 20 feature importance using ggplot2  
ggplot(top\_20\_importance, aes(x = reorder(Feature, Importance), y = Importance)) +  
 geom\_bar(stat = "identity") +  
 coord\_flip() +  
 xlab("Feature") +  
 ylab("Importance") +  
 ggtitle("Top 20 Feature Importance from Random Forest Model")



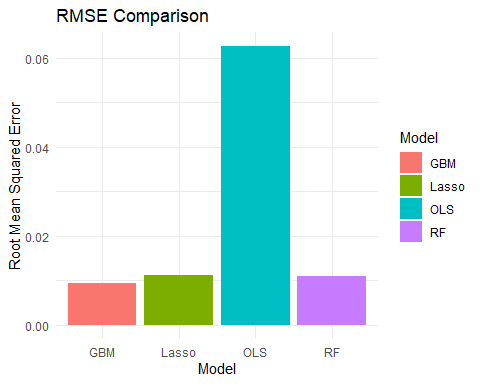
# Create a data frame to store the results  
results <- data.frame(  
 Model = c(  
 "OLS",  
 "Lasso",  
 "GBM",  
 "RF"  
 ),  
 MSE = c(  
 mse\_ols,  
 mse\_lasso,  
 mse\_gbm,  
 mse\_rf  
 ),  
 RMSE = c(  
 rmse\_ols,  
 rmse\_lasso,  
 rmse\_gbm,  
 rmse\_rf  
 ),  
 R\_squared = c(  
 r2\_ols,  
 r2\_lasso,  
 r2\_gbm,  
 r2\_rf  
 )  
 )  
  
# Print the results  
print(results)

## Model MSE RMSE R\_squared  
## 1 OLS 0.0039191529 0.062603138 0.9419975  
## 2 Lasso 0.0001253327 0.011195207 0.9389819  
## 3 GBM 0.0000868446 0.009319045 0.9577198  
## 4 RF 0.0001228929 0.011085705 0.9401699

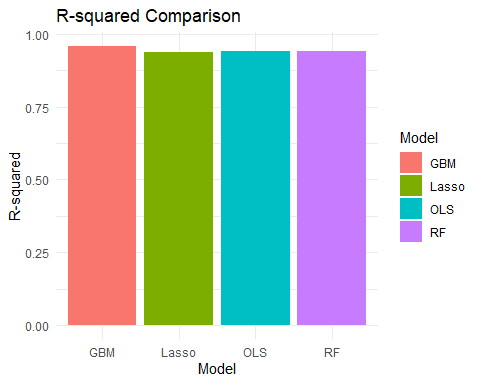
# Plot the results  
library(ggplot2)  
  
# MSE Comparison  
ggplot(  
 results,  
 aes(  
 x = Model,  
 y = MSE,  
 fill = Model  
 )  
 ) +  
 geom\_bar(stat = "identity") +  
 labs(  
 title = "MSE Comparison",  
 x = "Model",  
 y = "Mean Squared Error"  
 ) +  
 theme\_minimal()



# RMSE Comparison  
ggplot(  
 results,  
 aes(  
 x = Model,  
 y = RMSE,  
 fill = Model  
 )  
 ) +  
 geom\_bar(stat = "identity") +  
 labs(  
 title = "RMSE Comparison",  
 x = "Model",  
 y = "Root Mean Squared Error"  
 ) +  
 theme\_minimal()



# R-squared Comparison  
ggplot(  
 results,  
 aes(  
 x = Model,  
 y = R\_squared,  
 fill = Model  
 )  
 ) +  
 geom\_bar(stat = "identity") +  
 labs(  
 title = "R-squared Comparison",  
 x = "Model",  
 y = "R-squared"  
 ) +  
 theme\_minimal()



# Lower MSE and RMSE indicate better model performance.  
# Higher R-squared indicates better model performance.  
# The model with the lowest MSE and RMSE and the highest R-squared would be the recommended choice, the GBM model

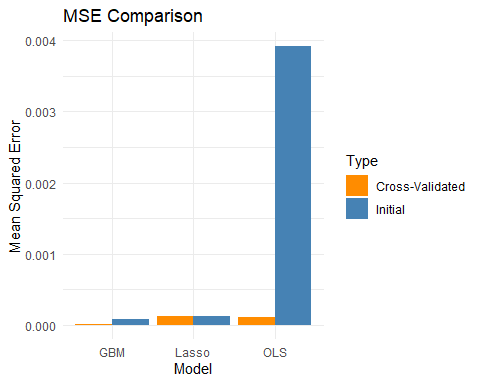
set.seed(42)  
# Combine training and testing data for cross-validation  
data <- rbind(  
 qol\_train,  
 qol\_test  
 )  
  
# Define cross-validation method  
train\_control <- trainControl(  
 method = "repeatedcv",  
 number = 10, # 10-fold cross-validation  
 repeats = 5  
 )   
  
# OLS Model  
ols\_model\_cv <- train(  
 pct\_poor\_to\_fair\_health ~ .,  
 data = data,  
 method = "lm",  
 trControl = train\_control  
 )  
  
ols\_predicted\_cv <- predict(  
 ols\_model\_cv,  
 data  
 )  
  
mse\_ols\_cv <- mean(  
 (data$pct\_poor\_to\_fair\_health - ols\_predicted\_cv)^2  
 )  
rmse\_ols\_cv <- sqrt(mse\_ols\_cv)  
r2\_ols\_cv <- R2(  
 ols\_predicted\_cv,  
 data$pct\_poor\_to\_fair\_health  
 )  
  
# Lasso Model  
lasso\_model\_cv <- train(  
 pct\_poor\_to\_fair\_health ~ .,  
 data = data,  
 method = "glmnet",  
 trControl = train\_control,  
 tuneGrid = expand.grid(  
 alpha = 1,  
 lambda = seq(  
 0.001,  
 0.1,  
 by = 0.001  
 )  
 )  
 )

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,  
## : There were missing values in resampled performance measures.

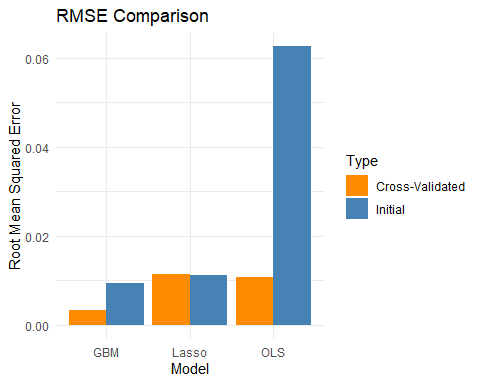
lasso\_predicted\_cv <- predict(  
 lasso\_model\_cv,  
 data  
 )  
  
mse\_lasso\_cv <- mean(  
 (data$pct\_poor\_to\_fair\_health - lasso\_predicted\_cv)^2  
 )  
rmse\_lasso\_cv <- sqrt(mse\_lasso\_cv)  
  
r2\_lasso\_cv <- R2(  
 lasso\_predicted\_cv,  
 data$pct\_poor\_to\_fair\_health  
 )  
  
# GBM Model  
gbm\_model\_cv <- train(  
 pct\_poor\_to\_fair\_health ~ .,  
 data = data,  
 method = "xgbTree",  
 trControl = train\_control,  
 tuneGrid = expand.grid(  
 nrounds = 100,  
 max\_depth = 6,  
 eta = 0.1,  
 gamma = 0,  
 colsample\_bytree = 0.8,  
 min\_child\_weight = 1,  
 subsample = 0.8  
 )  
 )  
  
gbm\_predicted\_cv <- predict(  
 gbm\_model\_cv,  
 data  
 )  
  
mse\_gbm\_cv <- mean(  
 (data$pct\_poor\_to\_fair\_health - gbm\_predicted\_cv)^2  
 )  
  
rmse\_gbm\_cv <- sqrt(mse\_gbm\_cv)  
  
r2\_gbm\_cv <- R2(  
 gbm\_predicted\_cv,  
 data$pct\_poor\_to\_fair\_health  
 )  
  
# Create a data frame to store the results  
cv\_results <- data.frame(  
 Model = rep(  
 c(  
 "OLS",  
 "Lasso",  
 "GBM"),  
 each = 2  
 ),  
 Type = rep(  
 c(  
 "Initial",  
 "Cross-Validated"  
 ),  
 times = 3  
 ),  
 MSE = c(  
 mse\_ols,  
 mse\_ols\_cv,  
 mse\_lasso,  
 mse\_lasso\_cv,  
 mse\_gbm,  
 mse\_gbm\_cv  
 ),  
 RMSE = c(  
 rmse\_ols,  
 rmse\_ols\_cv,  
 rmse\_lasso,  
 rmse\_lasso\_cv,  
 rmse\_gbm,  
 rmse\_gbm\_cv  
 ),  
 R\_squared = c(  
 r2\_ols,  
 r2\_ols\_cv,  
 r2\_lasso,  
 r2\_lasso\_cv,  
 r2\_gbm,  
 r2\_gbm\_cv  
 )  
)  
  
# Print the results  
print(cv\_results)

## Model Type MSE RMSE R\_squared  
## 1 OLS Initial 3.919153e-03 0.062603138 0.9419975  
## 2 OLS Cross-Validated 1.149833e-04 0.010723026 0.9417311  
## 3 Lasso Initial 1.253327e-04 0.011195207 0.9389819  
## 4 Lasso Cross-Validated 1.336389e-04 0.011560231 0.9335559  
## 5 GBM Initial 8.684460e-05 0.009319045 0.9577198  
## 6 GBM Cross-Validated 1.203421e-05 0.003469035 0.9939595

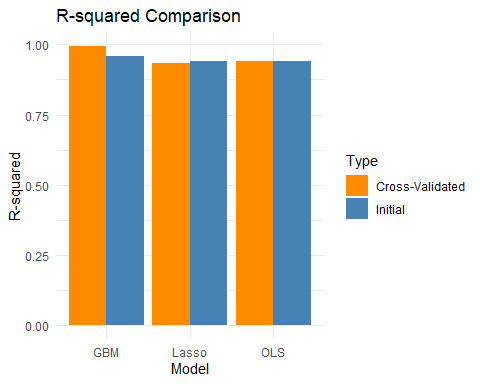
# Plot the results  
# MSE Comparison  
ggplot(  
 cv\_results,  
 aes(  
 x = Model,  
 y = MSE,  
 fill = Type  
 )  
 ) +  
 geom\_bar(  
 stat = "identity",  
 position = position\_dodge(  
 width = 0.9  
 )  
 ) +  
 labs(  
 title = "MSE Comparison",  
 x = "Model",  
 y = "Mean Squared Error"  
 ) +  
 theme\_minimal() +  
 scale\_fill\_manual(  
 values = c(  
 "Initial" = "steelblue",  
 "Cross-Validated" = "darkorange"  
 )  
 )



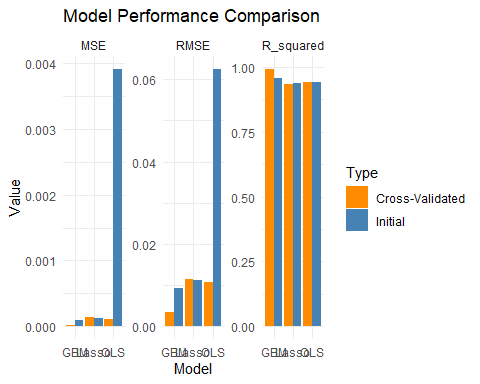
# RMSE Comparison  
ggplot(  
 cv\_results,  
 aes(  
 x = Model,  
 y = RMSE,  
 fill = Type  
 )  
 ) +  
 geom\_bar(  
 stat = "identity",  
 position = position\_dodge(  
 width = 0.9  
 )  
 ) +  
 labs(  
 title = "RMSE Comparison",  
 x = "Model",  
 y = "Root Mean Squared Error"  
 ) +  
 theme\_minimal() +  
 scale\_fill\_manual(  
 values = c(  
 "Initial" = "steelblue",  
 "Cross-Validated" = "darkorange"  
 )  
 )



# R-squared Comparison  
ggplot(  
 cv\_results,  
 aes(  
 x = Model,  
 y = R\_squared,  
 fill = Type  
 )  
 ) +  
 geom\_bar(  
 stat = "identity",  
 position = position\_dodge(  
 width = 0.9  
 )  
 ) +  
 labs(  
 title = "R-squared Comparison",  
 x = "Model",  
 y = "R-squared"  
 ) +  
 theme\_minimal() +  
 scale\_fill\_manual(  
 values = c(  
 "Initial" = "steelblue",  
 "Cross-Validated" = "darkorange"  
 )  
 )



# MSE, RMSE, and R-squared Comparison Side by Side  
cv\_results\_long <- reshape2::melt(  
 cv\_results,  
 id.vars = c(  
 "Model",  
 "Type"  
 ),   
 variable.name = "Metric",  
 value.name = "Value"  
 )  
  
ggplot(  
 cv\_results\_long,  
 aes(  
 x = Model,  
 y = Value,  
 fill = Type  
 )  
 ) +  
 geom\_bar(  
 stat = "identity",  
 position = position\_dodge(  
 width = 0.9  
 )  
 ) +  
 facet\_wrap(  
 ~ Metric,  
 scales = "free\_y"  
 ) +  
 labs(  
 title = "Model Performance Comparison",  
 x = "Model",  
 y = "Value"  
 ) +  
 theme\_minimal() +  
 scale\_fill\_manual(  
 values = c(  
 "Initial" = "steelblue",  
 "Cross-Validated" = "darkorange"  
 )  
 )



# cross validation confirms that the GBM model performs the best

# Combine training and testing data for cross-validation  
data <- rbind(  
 qol\_train,  
 qol\_test)  
  
# Define cross-validation method  
train\_control <- trainControl(  
 method = "cv",  
 number = 10 # 10-fold cross-validation  
 )  
  
# OLS Model  
ols\_model <- train(  
 pct\_poor\_to\_fair\_health ~ .,  
 data = data,  
 method = "lm",  
 trControl = train\_control  
 )  
# Predicted values and residuals for OLS model  
ols\_predicted <- predict(ols\_model, data)  
ols\_residuals <- data$pct\_poor\_to\_fair\_health - ols\_predicted  
  
# Lasso Model  
lasso\_model <- train(  
 pct\_poor\_to\_fair\_health ~ .,  
 data = data,  
 method = "glmnet",  
 trControl = train\_control,  
 tuneGrid = expand.grid(  
 alpha = 1,  
 lambda = seq(0.001, 0.1, by = 0.001)  
 )  
 )

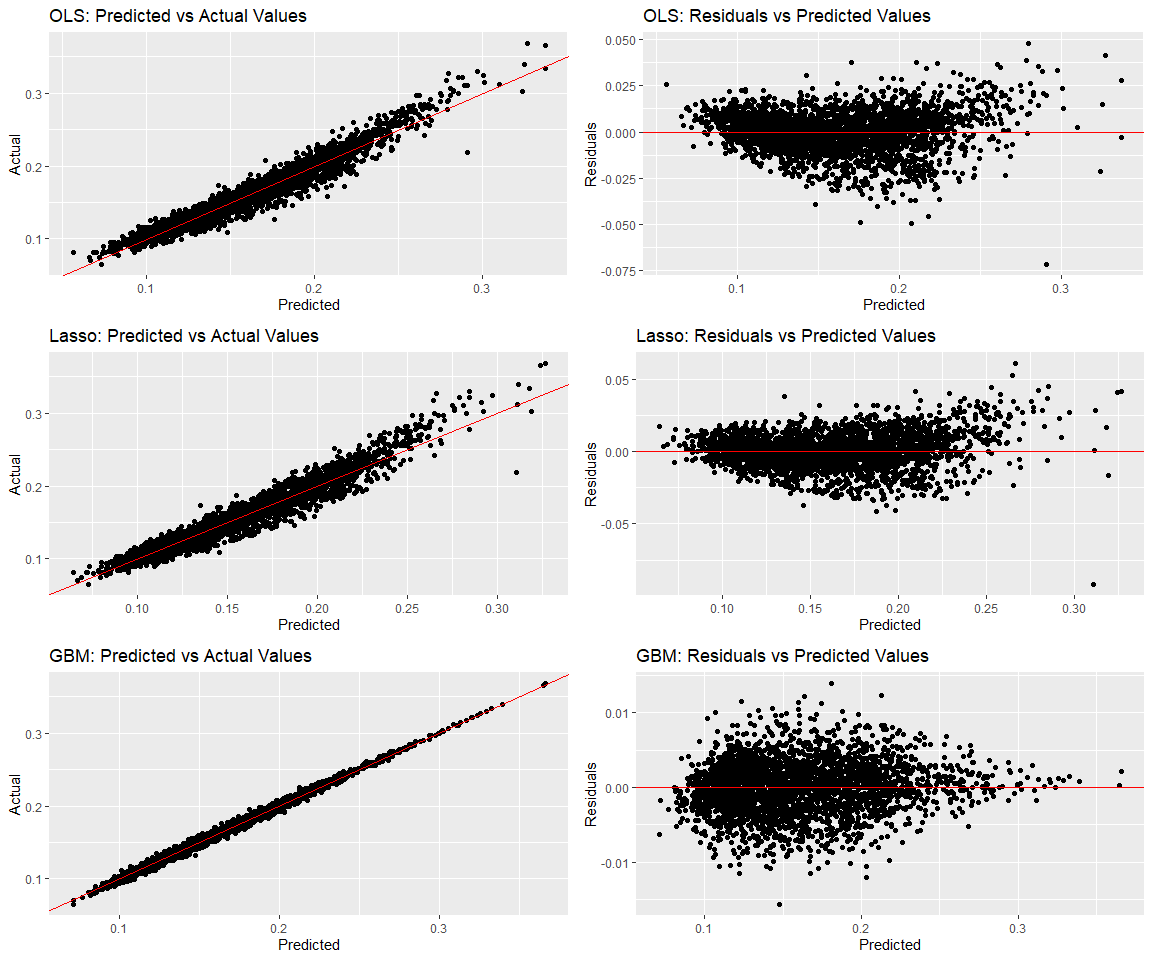
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,  
## : There were missing values in resampled performance measures.

# Predicted values and residuals for Lasso model  
lasso\_predicted <- predict(lasso\_model, data)  
lasso\_residuals <- data$pct\_poor\_to\_fair\_health - lasso\_predicted  
  
# GBM Model  
gbm\_model <- train(  
 pct\_poor\_to\_fair\_health ~ .,  
 data = data,  
 method = "xgbTree",  
 trControl = train\_control,  
 tuneGrid = expand.grid(  
 nrounds = 100,  
 max\_depth = 6,  
 eta = 0.1,  
 gamma = 0,  
 colsample\_bytree = 0.8,  
 min\_child\_weight = 1,  
 subsample = 0.8  
 )  
 )  
  
# Predicted values and residuals for GBM model  
gbm\_predicted <- predict(  
 gbm\_model,  
 data  
 )  
  
gbm\_residuals <- data$pct\_poor\_to\_fair\_health - gbm\_predicted  
  
# Create ggplot objects for each model  
p1 <- ggplot(  
 data,  
 aes(  
 x = ols\_predicted,  
 y = pct\_poor\_to\_fair\_health  
 )  
 ) +  
 geom\_point() +  
 geom\_abline(  
 slope = 1,  
 intercept = 0,  
 color = "red"  
 ) +  
 labs(  
 title = "OLS: Predicted vs Actual Values",  
 x = "Predicted",  
 y = "Actual"  
 )  
  
p2 <- ggplot(  
 data,  
 aes(  
 x = ols\_predicted,  
 y = ols\_residuals  
 )  
 ) +  
 geom\_point() +  
 geom\_hline(  
 yintercept = 0,  
 color = "red"  
 ) +  
 labs(  
 title = "OLS: Residuals vs Predicted Values",  
 x = "Predicted",  
 y = "Residuals")  
  
  
p5 <- ggplot(  
 data,  
 aes(  
 x =  
 lasso\_predicted,  
 y = pct\_poor\_to\_fair\_health  
 )  
 ) +  
 geom\_point() +  
 geom\_abline(  
 slope = 1,  
 intercept = 0,  
 color = "red"  
 ) +  
 labs(  
 title = "Lasso: Predicted vs Actual Values",  
 x = "Predicted",  
 y = "Actual"  
 )  
  
p6 <- ggplot(  
 data,  
 aes(  
 x = lasso\_predicted,  
 y = lasso\_residuals  
 )  
 ) +  
 geom\_point() +  
 geom\_hline(  
 yintercept = 0,  
 color = "red"  
 ) +  
 labs(  
 title = "Lasso: Residuals vs Predicted Values",  
 x = "Predicted",  
 y = "Residuals"  
 )  
  
p7 <- ggplot(  
 data,  
 aes(  
 x = gbm\_predicted,  
 y = pct\_poor\_to\_fair\_health  
 )  
 ) +  
 geom\_point() +  
 geom\_abline(  
 slope = 1,  
 intercept = 0,  
 color = "red"  
 ) +  
 labs(  
 title = "GBM: Predicted vs Actual Values",  
 x = "Predicted",  
 y = "Actual")  
  
p8 <- ggplot(  
 data,  
 aes(  
 x = gbm\_predicted,  
 y = gbm\_residuals  
 )  
 ) +  
 geom\_point() +  
 geom\_hline(  
 yintercept = 0,  
 color = "red"  
 ) +  
 labs(  
 title = "GBM: Residuals vs Predicted Values",  
 x = "Predicted",  
 y = "Residuals"  
 )  
  
# Combine plots into one  
library(gridExtra)

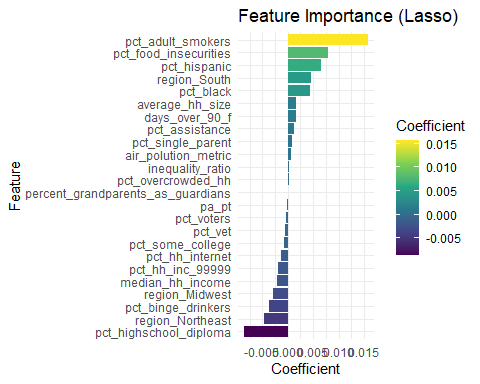
##   
## Attaching package: 'gridExtra'

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

grid.arrange(  
 p1,  
 p2,  
 p5,  
 p6,  
 p7,  
 p8,  
 ncol = 2  
 )



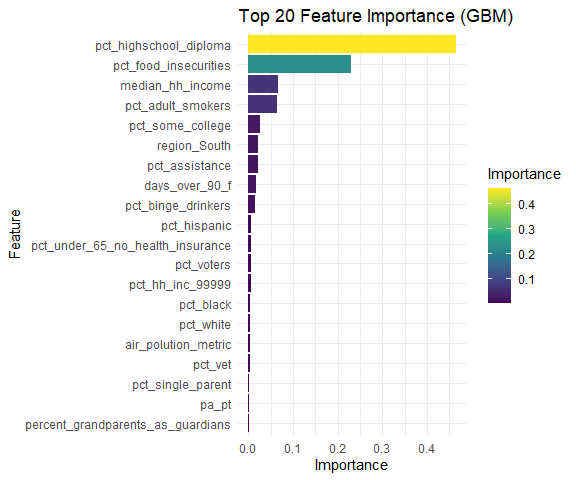
# Extract coefficients from the Lasso model  
lasso\_coefficients <- coef(  
 lasso\_model\_cv$finalModel,  
 s = lasso\_model\_cv$bestTune$lambda  
 )  
  
lasso\_coefficients <- as.data.frame(as.matrix(lasso\_coefficients))  
lasso\_coefficients$Feature <- rownames(lasso\_coefficients)  
colnames(lasso\_coefficients)[1] <- "Coefficient"  
  
# Filter non-zero coefficients and exclude the intercept  
lasso\_coefficients <- lasso\_coefficients[lasso\_coefficients$Coefficient != 0 & lasso\_coefficients$Feature != "(Intercept)", ]  
  
# Sort by magnitude  
lasso\_coefficients <- lasso\_coefficients[order(abs(lasso\_coefficients$Coefficient), decreasing = TRUE), ]  
  
# Plot the feature importance with color  
ggplot(  
 lasso\_coefficients,  
 aes(  
 x = reorder(Feature, Coefficient),  
 y = Coefficient,  
 fill = Coefficient  
 )  
 ) +  
 geom\_bar(stat = "identity") +  
 coord\_flip() +  
 labs(  
 title = "Feature Importance (Lasso)",  
 x = "Feature",  
 y = "Coefficient"  
 ) +  
 theme\_minimal() +  
 scale\_fill\_viridis\_c(  
 option = "viridis",   
 direction = 1,  
 name = "Coefficient"  
 )



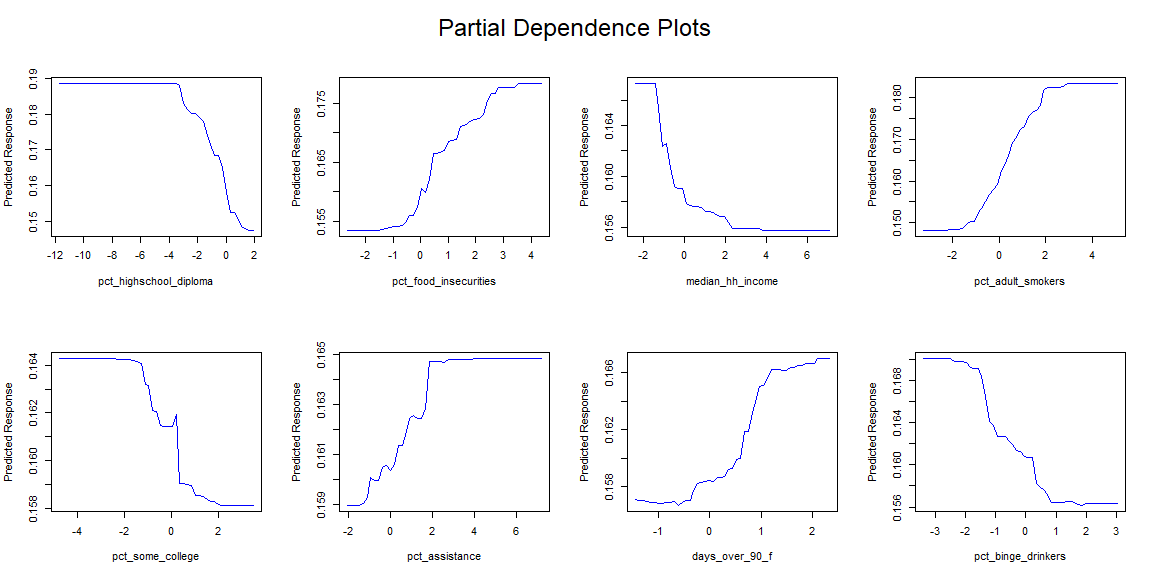
# Extract feature importance  
gbm\_feature\_importance\_cv <- varImp(  
 gbm\_model\_cv,  
 scale = FALSE  
)  
  
# Convert to a data frame  
gbm\_feature\_importance\_cv\_df <- as.data.frame(gbm\_feature\_importance\_cv$importance)  
  
# Add feature names  
gbm\_feature\_importance\_cv\_df$Feature <- rownames(gbm\_feature\_importance\_cv\_df)  
  
# Rename columns for clarity  
colnames(gbm\_feature\_importance\_cv\_df) <- c(  
 "Importance",  
 "Feature"  
 )  
  
# Sort by importance  
gbm\_feature\_importance\_cv\_df <- gbm\_feature\_importance\_cv\_df %>%  
 arrange(desc(Importance))  
  
# Export to a CSV file  
#write.csv(  
# gbm\_feature\_importance\_cv\_df,  
# "gbm\_feature\_importance.csv",  
# row.names = FALSE  
# )  
  
# Print the feature importance table  
print(gbm\_feature\_importance\_cv\_df)

## Importance  
## pct\_highschool\_diploma 4.643326e-01  
## pct\_food\_insecurities 2.297069e-01  
## median\_hh\_income 6.727079e-02  
## pct\_adult\_smokers 6.441548e-02  
## pct\_some\_college 2.592092e-02  
## region\_South 2.253109e-02  
## pct\_assistance 2.201346e-02  
## days\_over\_90\_f 1.801302e-02  
## pct\_binge\_drinkers 1.488110e-02  
## pct\_hispanic 6.557120e-03  
## pct\_under\_65\_no\_health\_insurance 6.374268e-03  
## pct\_voters 5.878899e-03  
## pct\_hh\_inc\_99999 5.689507e-03  
## pct\_black 3.468132e-03  
## pct\_white 3.450485e-03  
## air\_polution\_metric 3.250158e-03  
## pct\_vet 3.075703e-03  
## pct\_single\_parent 2.367880e-03  
## pa\_pt 2.046616e-03  
## percent\_grandparents\_as\_guardians 2.028129e-03  
## pct\_uninsured 1.821077e-03  
## average\_hh\_size 1.736171e-03  
## pct\_overcrowded\_hh 1.441646e-03  
## pct\_65\_plus 1.190921e-03  
## population\_density 1.171810e-03  
## pct\_other\_race 1.129866e-03  
## pct\_native\_american 1.127381e-03  
## median\_er\_dist 1.102490e-03  
## pct\_hh\_other\_computer 1.098252e-03  
## social\_clubs\_per\_10k 9.887691e-04  
## weighted\_population 9.279541e-04  
## median\_drug\_alcohol\_care\_dist 9.147313e-04  
## pct\_home\_owner 8.694203e-04  
## pct\_hh\_internet 8.486502e-04  
## pct\_employed 8.414767e-04  
## pct\_male 8.223924e-04  
## median\_pediatric\_icu\_dist 8.152863e-04  
## pct\_high\_housing\_costs 8.130358e-04  
## pct\_w\_medicare 8.114936e-04  
## inequality\_ratio 7.974180e-04  
## hh\_tot\_workers 7.473830e-04  
## pct\_30\_min\_plus\_commute 7.458699e-04  
## pct\_rural\_population 7.451414e-04  
## region\_Midwest 6.122739e-04  
## pct\_asian 5.155501e-04  
## median\_health\_clinic\_dist 4.873042e-04  
## region\_West 4.423313e-04  
## dentist\_pt 4.297006e-04  
## mental\_health\_faciliy\_pt 3.225875e-04  
## clinical\_nurse\_pt 1.730318e-04  
## region\_Northeast 1.639439e-04  
## water\_quality 7.238083e-05  
## Feature  
## pct\_highschool\_diploma pct\_highschool\_diploma  
## pct\_food\_insecurities pct\_food\_insecurities  
## median\_hh\_income median\_hh\_income  
## pct\_adult\_smokers pct\_adult\_smokers  
## pct\_some\_college pct\_some\_college  
## region\_South region\_South  
## pct\_assistance pct\_assistance  
## days\_over\_90\_f days\_over\_90\_f  
## pct\_binge\_drinkers pct\_binge\_drinkers  
## pct\_hispanic pct\_hispanic  
## pct\_under\_65\_no\_health\_insurance pct\_under\_65\_no\_health\_insurance  
## pct\_voters pct\_voters  
## pct\_hh\_inc\_99999 pct\_hh\_inc\_99999  
## pct\_black pct\_black  
## pct\_white pct\_white  
## air\_polution\_metric air\_polution\_metric  
## pct\_vet pct\_vet  
## pct\_single\_parent pct\_single\_parent  
## pa\_pt pa\_pt  
## percent\_grandparents\_as\_guardians percent\_grandparents\_as\_guardians  
## pct\_uninsured pct\_uninsured  
## average\_hh\_size average\_hh\_size  
## pct\_overcrowded\_hh pct\_overcrowded\_hh  
## pct\_65\_plus pct\_65\_plus  
## population\_density population\_density  
## pct\_other\_race pct\_other\_race  
## pct\_native\_american pct\_native\_american  
## median\_er\_dist median\_er\_dist  
## pct\_hh\_other\_computer pct\_hh\_other\_computer  
## social\_clubs\_per\_10k social\_clubs\_per\_10k  
## weighted\_population weighted\_population  
## median\_drug\_alcohol\_care\_dist median\_drug\_alcohol\_care\_dist  
## pct\_home\_owner pct\_home\_owner  
## pct\_hh\_internet pct\_hh\_internet  
## pct\_employed pct\_employed  
## pct\_male pct\_male  
## median\_pediatric\_icu\_dist median\_pediatric\_icu\_dist  
## pct\_high\_housing\_costs pct\_high\_housing\_costs  
## pct\_w\_medicare pct\_w\_medicare  
## inequality\_ratio inequality\_ratio  
## hh\_tot\_workers hh\_tot\_workers  
## pct\_30\_min\_plus\_commute pct\_30\_min\_plus\_commute  
## pct\_rural\_population pct\_rural\_population  
## region\_Midwest region\_Midwest  
## pct\_asian pct\_asian  
## median\_health\_clinic\_dist median\_health\_clinic\_dist  
## region\_West region\_West  
## dentist\_pt dentist\_pt  
## mental\_health\_faciliy\_pt mental\_health\_faciliy\_pt  
## clinical\_nurse\_pt clinical\_nurse\_pt  
## region\_Northeast region\_Northeast  
## water\_quality water\_quality

# Select the top 20 features based on importance  
top\_20\_features <- gbm\_feature\_importance\_cv\_df %>%  
 arrange(desc(Importance)) %>%  
 head(20)  
  
# Plot the top 20 feature importances with color  
ggplot(  
 top\_20\_features,  
 aes(  
 x = reorder(Feature, Importance),  
 y = Importance,  
 fill = Importance  
 )  
) +  
 geom\_bar(stat = "identity") +  
 coord\_flip() +  
 labs(  
 title = "Top 20 Feature Importance (GBM)",  
 x = "Feature",  
 y = "Importance"  
 ) +  
 theme\_minimal() +  
 scale\_fill\_viridis\_c(  
 option = "viridis",   
 direction = 1,  
 name = "Importance"  
 )



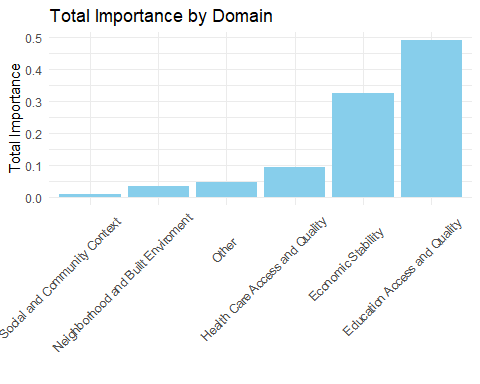
# Define the predictors  
predictors <- c(  
 "pct\_highschool\_diploma",  
 "pct\_food\_insecurities",  
 "median\_hh\_income",  
 "pct\_adult\_smokers",  
 "pct\_some\_college",  
 "pct\_assistance",  
 "days\_over\_90\_f",  
 "pct\_binge\_drinkers"  
)  
  
# Set up plotting area with a grid layout  
par(  
 mfrow = c(2, 4),  
 oma = c(0, 0, 2, 0) # 'oma' for outer margins  
 )   
  
# Create and plot partial dependence plots for each predictor  
for (predictor in predictors) {  
 # Generate partial dependence data  
 partial\_plot <- partial(  
 gbm\_model\_cv,  
 pred.var = predictor,  
 train = qol\_train # Use preprocessed data if needed  
 )  
   
 plot\_data <- as.data.frame(partial\_plot)  
   
 # Plot partial dependence  
 plot(  
 plot\_data[[predictor]], # x-axis  
 plot\_data$yhat, # y-axis  
 type = "l", # Line plot  
 col = "blue", # Line color  
 xlab = predictor, # X-axis label  
 ylab = "Predicted Response", # Y-axis label  
 )  
  
}  
  
# Add a main title to the entire plotting area  
mtext(  
 "Partial Dependence Plots",  
 outer = TRUE,  
 line = -1,  
 cex = 1.5  
 )



gbm\_feature\_importance\_cv\_df <- gbm\_feature\_importance\_cv\_df %>%   
 mutate(  
 Category = case\_when(  
 Feature == "pct\_no\_exercise" ~ "Health Care Access and Quality",  
 Feature == "pct\_highschool\_diploma" ~ "Education Access and Quality",  
 Feature == "pct\_food\_insecurities" ~ "Economic Stability",  
 Feature == "pct\_adult\_smokers" ~"Health Care Access and Quality",  
 Feature == "pct\_obese\_adults" ~ "Health Care Access and Quality",  
 Feature == "life\_expectancy\_years" ~ "Other",  
 Feature == "median\_hh\_income" ~ "Economic Stability",  
 Feature == "pct\_binge\_drinkers" ~ "Health Care Access and Quality",  
 Feature == "region\_Midwest" ~ "Other",  
 Feature == "region\_South" ~ "Other",  
 Feature == "pct\_under\_65\_no\_health\_insurance" ~ "Health Care Access and Quality",  
 Feature == "pct\_hispanic" ~ "Other",  
 Feature == "days\_over\_90\_f" ~ "Neighborhood and Built Enviroment",  
 Feature == "pct\_assistance" ~ "Economic Stability",  
 Feature == "pct\_some\_college" ~ "Education Access and Quality",  
 Feature == "air\_polution\_metric" ~ "Neighborhood and Built Enviroment",  
 Feature == "pct\_employed" ~ "Economic Stability",  
 Feature == "region\_West" ~ "Other",  
 Feature == "pct\_hh\_inc\_99999" ~ "Economic Stability",  
 Feature == "percent\_grandparents\_as\_guardians" ~ "Social and Community Context",  
 Feature == "pct\_uninsured" ~ "Health Care Access and Quality",  
 Feature == "pct\_native\_american" ~ "Other",   
 Feature == "region\_Northeast" ~ "Other",  
 Feature == "pct\_overcrowded\_hh" ~ "Social and Community Context",  
 Feature == "pct\_white" ~ "Other",  
 Feature == "median\_drug\_alcohol\_care\_dist" ~ "Health Care Access and Quality",  
 Feature == "pct\_black" ~ "Other",  
 Feature == "average\_hh\_size" ~ "Social and Community Context",  
 Feature == "pct\_high\_housing\_costs" ~ "Social and Community Context",  
 Feature == "social\_clubs\_per\_10k" ~ "Social and Community Context",  
 Feature == "hh\_tot\_workers" ~ "Economic Stability",  
 Feature == "pct\_home\_owner" ~ "Social and Community Context",  
 Feature == "pa\_pt" ~ "Health Care Access and Quality",  
 Feature == "pct\_30\_min\_plus\_commute" ~ "Neighborhood and Built Enviroment",  
 Feature == "pct\_hh\_other\_computer" ~ "Education Access and Quality",  
 Feature == "pct\_vet" ~ "Neighborhood and Built Enviroment",  
 Feature == "pct\_voters" ~ "Neighborhood and Built Enviroment",  
 Feature == "pct\_hh\_internet" ~ "Education Access and Quality",  
 Feature == "median\_pediatric\_icu\_dist" ~ "Health Care Access and Quality",  
 Feature == "population\_density" ~ "Other",  
 Feature == "pct\_other\_race" ~ "Other",  
 Feature == "pct\_male" ~ "Other",  
 Feature == "pct\_asian" ~ "Other",  
 Feature == "pct\_single\_parent" ~ "Neighborhood and Built Enviroment",  
 Feature == "median\_er\_dist" ~ "Health Care Access and Quality",  
 Feature == "inequality\_ratio" ~ "Economic Stability",  
 Feature == "median\_health\_clinic\_dist" ~ "Health Care Access and Quality",  
 Feature == "weighted\_population" ~ "Other",  
 Feature == "mental\_health\_faciliy\_pt" ~ "Health Care Access and Quality",  
 Feature == "dentist\_pt" ~ "Health Care Access and Quality",  
 Feature == "pct\_rural\_population" ~ "Other",  
 Feature == "clinical\_nurse\_pt" ~ "Health Care Access and Quality",  
 Feature == "pct\_w\_medicare" ~ "Health Care Access and Quality",  
 Feature == "pct\_65\_plus" ~ "Other",  
 Feature == "water\_quality" ~ "Neighborhood and Built Enviroment",  
 Feature == "percent\_grandparents\_as\_guardians" ~ "Neighborhood and Built Enviroment",  
 TRUE ~ NA\_character\_ # Default case if no conditions are met  
 ),  
 Explanation = case\_when(  
 Feature == "pct\_no\_exercise" ~ "Percentage of adults not exercising increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_highschool\_diploma" ~ "Higher percentage of adults with a high school diploma decreases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_food\_insecurities" ~ "Higher food insecurity increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_adult\_smokers" ~"Higher smoking prevalence among adults increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_obese\_adults" ~ "Higher obesity among adults increases the percentage of adults reporting poor or fair health.",  
 Feature == "life\_expectancy\_years" ~ "Higher life expectancy decreases the percentage of adults reporting poor or fair health.",  
 Feature == "median\_hh\_income" ~ "Higher median household income decreases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_binge\_drinkers" ~ "Less binge drinking prevalence among adults increases the percentage of adults reporting poor or fair health.",  
 Feature == "region\_Midwest" ~ "Living in the Midwest region increases the percentage of adults reporting poor or fair health.",  
 Feature == "region\_South" ~ "Living in the South region increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_under\_65\_no\_health\_insurance" ~ "Higher percentage of adults under 65 without health insurance increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_hispanic" ~ "Higher percentage of Hispanic population increases the percentage of adults reporting poor or fair health.",  
 Feature == "days\_over\_90\_f" ~ "More days over 90°F increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_assistance" ~ "Higher percentage of population receiving assistance increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_some\_college" ~ "Higher percentage of adults with some college education decreases the percentage of adults reporting poor or fair health.",  
 Feature == "air\_polution\_metric" ~ "Higher air pollution metric increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_employed" ~ "Higher employment rate decreases the percentage of adults reporting poor or fair health.",  
 Feature == "region\_West" ~ "Living in the West region decreases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_hh\_inc\_99999" ~ "Higher percentage of households with income over $99,999 decreases the percentage of adults reporting poor or fair health.",  
 Feature == "percent\_grandparents\_as\_guardians" ~ "Higher percentage of grandparents as guardians increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_uninsured" ~ "Higher percentage of uninsured population increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_native\_american" ~ "Higher percentage of Native American population increases the percentage of adults reporting poor or fair health.",   
 Feature == "region\_Northeast" ~ "Living in the Northeast region decreases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_overcrowded\_hh" ~ "Higher percentage of overcrowded households increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_white" ~ "Higher percentage of White population decreases the percentage of adults reporting poor or fair health.",  
 Feature == "median\_drug\_alcohol\_care\_dist" ~ "Greater distance to drug and alcohol care facilities increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_black" ~ "Higher percentage of Black population increases the percentage of adults reporting poor or fair health.",  
 Feature == "average\_hh\_size" ~ "Larger average household size increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_high\_housing\_costs" ~ "Higher percentage of households with high housing costs increases the percentage of adults reporting poor or fair health.",  
 Feature == "social\_clubs\_per\_10k" ~ "More social clubs per 10,000 population decreases the percentage of adults reporting poor or fair health.",  
 Feature == "hh\_tot\_workers" ~ "Higher total number of workers in households decreases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_home\_owner" ~ "Higher percentage of home owners decreases the percentage of adults reporting poor or fair health.",  
 Feature == "pa\_pt" ~ "More physician assistants per population decreases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_30\_min\_plus\_commute" ~ "Higher percentage of adults with 30+ minute commute increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_hh\_other\_computer" ~ "Higher percentage of households with other computers decreases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_vet" ~ "Higher percentage of veterans decreases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_voters" ~ "Higher percentage of voters decreases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_hh\_internet" ~ "Higher percentage of households with internet access decreases the percentage of adults reporting poor or fair health.",  
 Feature == "median\_pediatric\_icu\_dist" ~ "Greater distance to pediatric ICU increases the percentage of adults reporting poor or fair health.",  
 Feature == "population\_density" ~ "Higher population density increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_other\_race" ~ "Higher percentage of other races increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_male" ~ "Higher percentage of male population increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_asian" ~ "Higher percentage of Asian population increases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_single\_parent" ~ "Higher percentage of single-parent households increases the percentage of adults reporting poor or fair health.",  
 Feature == "median\_er\_dist" ~ "Greater distance to emergency room increases the percentage of adults reporting poor or fair health.",  
 Feature == "inequality\_ratio" ~"Higher inequality ratio increases the percentage of adults reporting poor or fair health.",  
 Feature == "median\_health\_clinic\_dist" ~ "Greater distance to health clinics increases the percentage of adults reporting poor or fair health.",  
 Feature == "weighted\_population" ~ "Higher weighted population increases the percentage of adults reporting poor or fair health.",  
 Feature == "mental\_health\_faciliy\_pt" ~ "More mental health facilities per population decreases the percentage of adults reporting poor or fair health.",  
 Feature == "dentist\_pt" ~ "More dentists per population decreases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_rural\_population" ~ "Higher percentage of rural population increases the percentage of adults reporting poor or fair health.",  
 Feature == "clinical\_nurse\_pt" ~ "More clinical nurses per population decreases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_w\_medicare" ~ "Higher percentage of population with Medicare decreases the percentage of adults reporting poor or fair health.",  
 Feature == "pct\_65\_plus" ~ "Higher percentage of population aged 65+ increases the percentage of adults reporting poor or fair health.",  
 Feature == "water\_quality" ~ "Better water quality decreases the percentage of adults reporting poor or fair health.",  
 Feature == "percent\_grandparents\_as\_guardians" ~ "More grandparents as guardians increaces the percentage of adults reporting poor or fair health.",  
 TRUE ~ NA\_character\_ # Default case if no conditions are met  
 )  
 )  
  
  
# Export to an Excel file  
openxlsx::write.xlsx(gbm\_feature\_importance\_cv\_df, "gbm\_feature\_importance\_explanations.xlsx")  
  
# Print the feature importance table with explanations  
print(gbm\_feature\_importance\_cv\_df)

## Importance  
## pct\_highschool\_diploma 4.643326e-01  
## pct\_food\_insecurities 2.297069e-01  
## median\_hh\_income 6.727079e-02  
## pct\_adult\_smokers 6.441548e-02  
## pct\_some\_college 2.592092e-02  
## region\_South 2.253109e-02  
## pct\_assistance 2.201346e-02  
## days\_over\_90\_f 1.801302e-02  
## pct\_binge\_drinkers 1.488110e-02  
## pct\_hispanic 6.557120e-03  
## pct\_under\_65\_no\_health\_insurance 6.374268e-03  
## pct\_voters 5.878899e-03  
## pct\_hh\_inc\_99999 5.689507e-03  
## pct\_black 3.468132e-03  
## pct\_white 3.450485e-03  
## air\_polution\_metric 3.250158e-03  
## pct\_vet 3.075703e-03  
## pct\_single\_parent 2.367880e-03  
## pa\_pt 2.046616e-03  
## percent\_grandparents\_as\_guardians 2.028129e-03  
## pct\_uninsured 1.821077e-03  
## average\_hh\_size 1.736171e-03  
## pct\_overcrowded\_hh 1.441646e-03  
## pct\_65\_plus 1.190921e-03  
## population\_density 1.171810e-03  
## pct\_other\_race 1.129866e-03  
## pct\_native\_american 1.127381e-03  
## median\_er\_dist 1.102490e-03  
## pct\_hh\_other\_computer 1.098252e-03  
## social\_clubs\_per\_10k 9.887691e-04  
## weighted\_population 9.279541e-04  
## median\_drug\_alcohol\_care\_dist 9.147313e-04  
## pct\_home\_owner 8.694203e-04  
## pct\_hh\_internet 8.486502e-04  
## pct\_employed 8.414767e-04  
## pct\_male 8.223924e-04  
## median\_pediatric\_icu\_dist 8.152863e-04  
## pct\_high\_housing\_costs 8.130358e-04  
## pct\_w\_medicare 8.114936e-04  
## inequality\_ratio 7.974180e-04  
## hh\_tot\_workers 7.473830e-04  
## pct\_30\_min\_plus\_commute 7.458699e-04  
## pct\_rural\_population 7.451414e-04  
## region\_Midwest 6.122739e-04  
## pct\_asian 5.155501e-04  
## median\_health\_clinic\_dist 4.873042e-04  
## region\_West 4.423313e-04  
## dentist\_pt 4.297006e-04  
## mental\_health\_faciliy\_pt 3.225875e-04  
## clinical\_nurse\_pt 1.730318e-04  
## region\_Northeast 1.639439e-04  
## water\_quality 7.238083e-05  
## Feature  
## pct\_highschool\_diploma pct\_highschool\_diploma  
## pct\_food\_insecurities pct\_food\_insecurities  
## median\_hh\_income median\_hh\_income  
## pct\_adult\_smokers pct\_adult\_smokers  
## pct\_some\_college pct\_some\_college  
## region\_South region\_South  
## pct\_assistance pct\_assistance  
## days\_over\_90\_f days\_over\_90\_f  
## pct\_binge\_drinkers pct\_binge\_drinkers  
## pct\_hispanic pct\_hispanic  
## pct\_under\_65\_no\_health\_insurance pct\_under\_65\_no\_health\_insurance  
## pct\_voters pct\_voters  
## pct\_hh\_inc\_99999 pct\_hh\_inc\_99999  
## pct\_black pct\_black  
## pct\_white pct\_white  
## air\_polution\_metric air\_polution\_metric  
## pct\_vet pct\_vet  
## pct\_single\_parent pct\_single\_parent  
## pa\_pt pa\_pt  
## percent\_grandparents\_as\_guardians percent\_grandparents\_as\_guardians  
## pct\_uninsured pct\_uninsured  
## average\_hh\_size average\_hh\_size  
## pct\_overcrowded\_hh pct\_overcrowded\_hh  
## pct\_65\_plus pct\_65\_plus  
## population\_density population\_density  
## pct\_other\_race pct\_other\_race  
## pct\_native\_american pct\_native\_american  
## median\_er\_dist median\_er\_dist  
## pct\_hh\_other\_computer pct\_hh\_other\_computer  
## social\_clubs\_per\_10k social\_clubs\_per\_10k  
## weighted\_population weighted\_population  
## median\_drug\_alcohol\_care\_dist median\_drug\_alcohol\_care\_dist  
## pct\_home\_owner pct\_home\_owner  
## pct\_hh\_internet pct\_hh\_internet  
## pct\_employed pct\_employed  
## pct\_male pct\_male  
## median\_pediatric\_icu\_dist median\_pediatric\_icu\_dist  
## pct\_high\_housing\_costs pct\_high\_housing\_costs  
## pct\_w\_medicare pct\_w\_medicare  
## inequality\_ratio inequality\_ratio  
## hh\_tot\_workers hh\_tot\_workers  
## pct\_30\_min\_plus\_commute pct\_30\_min\_plus\_commute  
## pct\_rural\_population pct\_rural\_population  
## region\_Midwest region\_Midwest  
## pct\_asian pct\_asian  
## median\_health\_clinic\_dist median\_health\_clinic\_dist  
## region\_West region\_West  
## dentist\_pt dentist\_pt  
## mental\_health\_faciliy\_pt mental\_health\_faciliy\_pt  
## clinical\_nurse\_pt clinical\_nurse\_pt  
## region\_Northeast region\_Northeast  
## water\_quality water\_quality  
## Category  
## pct\_highschool\_diploma Education Access and Quality  
## pct\_food\_insecurities Economic Stability  
## median\_hh\_income Economic Stability  
## pct\_adult\_smokers Health Care Access and Quality  
## pct\_some\_college Education Access and Quality  
## region\_South Other  
## pct\_assistance Economic Stability  
## days\_over\_90\_f Neighborhood and Built Enviroment  
## pct\_binge\_drinkers Health Care Access and Quality  
## pct\_hispanic Other  
## pct\_under\_65\_no\_health\_insurance Health Care Access and Quality  
## pct\_voters Neighborhood and Built Enviroment  
## pct\_hh\_inc\_99999 Economic Stability  
## pct\_black Other  
## pct\_white Other  
## air\_polution\_metric Neighborhood and Built Enviroment  
## pct\_vet Neighborhood and Built Enviroment  
## pct\_single\_parent Neighborhood and Built Enviroment  
## pa\_pt Health Care Access and Quality  
## percent\_grandparents\_as\_guardians Social and Community Context  
## pct\_uninsured Health Care Access and Quality  
## average\_hh\_size Social and Community Context  
## pct\_overcrowded\_hh Social and Community Context  
## pct\_65\_plus Other  
## population\_density Other  
## pct\_other\_race Other  
## pct\_native\_american Other  
## median\_er\_dist Health Care Access and Quality  
## pct\_hh\_other\_computer Education Access and Quality  
## social\_clubs\_per\_10k Social and Community Context  
## weighted\_population Other  
## median\_drug\_alcohol\_care\_dist Health Care Access and Quality  
## pct\_home\_owner Social and Community Context  
## pct\_hh\_internet Education Access and Quality  
## pct\_employed Economic Stability  
## pct\_male Other  
## median\_pediatric\_icu\_dist Health Care Access and Quality  
## pct\_high\_housing\_costs Social and Community Context  
## pct\_w\_medicare Health Care Access and Quality  
## inequality\_ratio Economic Stability  
## hh\_tot\_workers Economic Stability  
## pct\_30\_min\_plus\_commute Neighborhood and Built Enviroment  
## pct\_rural\_population Other  
## region\_Midwest Other  
## pct\_asian Other  
## median\_health\_clinic\_dist Health Care Access and Quality  
## region\_West Other  
## dentist\_pt Health Care Access and Quality  
## mental\_health\_faciliy\_pt Health Care Access and Quality  
## clinical\_nurse\_pt Health Care Access and Quality  
## region\_Northeast Other  
## water\_quality Neighborhood and Built Enviroment  
## Explanation  
## pct\_highschool\_diploma Higher percentage of adults with a high school diploma decreases the percentage of adults reporting poor or fair health.  
## pct\_food\_insecurities Higher food insecurity increases the percentage of adults reporting poor or fair health.  
## median\_hh\_income Higher median household income decreases the percentage of adults reporting poor or fair health.  
## pct\_adult\_smokers Higher smoking prevalence among adults increases the percentage of adults reporting poor or fair health.  
## pct\_some\_college Higher percentage of adults with some college education decreases the percentage of adults reporting poor or fair health.  
## region\_South Living in the South region increases the percentage of adults reporting poor or fair health.  
## pct\_assistance Higher percentage of population receiving assistance increases the percentage of adults reporting poor or fair health.  
## days\_over\_90\_f More days over 90°F increases the percentage of adults reporting poor or fair health.  
## pct\_binge\_drinkers Less binge drinking prevalence among adults increases the percentage of adults reporting poor or fair health.  
## pct\_hispanic Higher percentage of Hispanic population increases the percentage of adults reporting poor or fair health.  
## pct\_under\_65\_no\_health\_insurance Higher percentage of adults under 65 without health insurance increases the percentage of adults reporting poor or fair health.  
## pct\_voters Higher percentage of voters decreases the percentage of adults reporting poor or fair health.  
## pct\_hh\_inc\_99999 Higher percentage of households with income over $99,999 decreases the percentage of adults reporting poor or fair health.  
## pct\_black Higher percentage of Black population increases the percentage of adults reporting poor or fair health.  
## pct\_white Higher percentage of White population decreases the percentage of adults reporting poor or fair health.  
## air\_polution\_metric Higher air pollution metric increases the percentage of adults reporting poor or fair health.  
## pct\_vet Higher percentage of veterans decreases the percentage of adults reporting poor or fair health.  
## pct\_single\_parent Higher percentage of single-parent households increases the percentage of adults reporting poor or fair health.  
## pa\_pt More physician assistants per population decreases the percentage of adults reporting poor or fair health.  
## percent\_grandparents\_as\_guardians Higher percentage of grandparents as guardians increases the percentage of adults reporting poor or fair health.  
## pct\_uninsured Higher percentage of uninsured population increases the percentage of adults reporting poor or fair health.  
## average\_hh\_size Larger average household size increases the percentage of adults reporting poor or fair health.  
## pct\_overcrowded\_hh Higher percentage of overcrowded households increases the percentage of adults reporting poor or fair health.  
## pct\_65\_plus Higher percentage of population aged 65+ increases the percentage of adults reporting poor or fair health.  
## population\_density Higher population density increases the percentage of adults reporting poor or fair health.  
## pct\_other\_race Higher percentage of other races increases the percentage of adults reporting poor or fair health.  
## pct\_native\_american Higher percentage of Native American population increases the percentage of adults reporting poor or fair health.  
## median\_er\_dist Greater distance to emergency room increases the percentage of adults reporting poor or fair health.  
## pct\_hh\_other\_computer Higher percentage of households with other computers decreases the percentage of adults reporting poor or fair health.  
## social\_clubs\_per\_10k More social clubs per 10,000 population decreases the percentage of adults reporting poor or fair health.  
## weighted\_population Higher weighted population increases the percentage of adults reporting poor or fair health.  
## median\_drug\_alcohol\_care\_dist Greater distance to drug and alcohol care facilities increases the percentage of adults reporting poor or fair health.  
## pct\_home\_owner Higher percentage of home owners decreases the percentage of adults reporting poor or fair health.  
## pct\_hh\_internet Higher percentage of households with internet access decreases the percentage of adults reporting poor or fair health.  
## pct\_employed Higher employment rate decreases the percentage of adults reporting poor or fair health.  
## pct\_male Higher percentage of male population increases the percentage of adults reporting poor or fair health.  
## median\_pediatric\_icu\_dist Greater distance to pediatric ICU increases the percentage of adults reporting poor or fair health.  
## pct\_high\_housing\_costs Higher percentage of households with high housing costs increases the percentage of adults reporting poor or fair health.  
## pct\_w\_medicare Higher percentage of population with Medicare decreases the percentage of adults reporting poor or fair health.  
## inequality\_ratio Higher inequality ratio increases the percentage of adults reporting poor or fair health.  
## hh\_tot\_workers Higher total number of workers in households decreases the percentage of adults reporting poor or fair health.  
## pct\_30\_min\_plus\_commute Higher percentage of adults with 30+ minute commute increases the percentage of adults reporting poor or fair health.  
## pct\_rural\_population Higher percentage of rural population increases the percentage of adults reporting poor or fair health.  
## region\_Midwest Living in the Midwest region increases the percentage of adults reporting poor or fair health.  
## pct\_asian Higher percentage of Asian population increases the percentage of adults reporting poor or fair health.  
## median\_health\_clinic\_dist Greater distance to health clinics increases the percentage of adults reporting poor or fair health.  
## region\_West Living in the West region decreases the percentage of adults reporting poor or fair health.  
## dentist\_pt More dentists per population decreases the percentage of adults reporting poor or fair health.  
## mental\_health\_faciliy\_pt More mental health facilities per population decreases the percentage of adults reporting poor or fair health.  
## clinical\_nurse\_pt More clinical nurses per population decreases the percentage of adults reporting poor or fair health.  
## region\_Northeast Living in the Northeast region decreases the percentage of adults reporting poor or fair health.  
## water\_quality Better water quality decreases the percentage of adults reporting poor or fair health.

sdoh\_category\_data <- gbm\_feature\_importance\_cv\_df %>%   
 group\_by(Category) %>%   
 summarize(  
 total\_importance = sum(Importance)  
 ) %>%   
 arrange(desc(total\_importance))  
  
  
ggplot(  
 data = sdoh\_category\_data,  
 aes(  
 x = reorder(Category, total\_importance), # Reorder categories based on total importance  
 y = total\_importance  
 )  
) +  
 geom\_bar(stat = "identity", fill = "skyblue") + # Use stat = "identity" for actual values  
 labs(  
 x = "",  
 y = "Total Importance",  
 title = "Total Importance by Domain"  
 ) +  
 theme\_minimal() +  
 theme(  
 axis.text.x = element\_text(  
 angle = 45,  
 hjust = 0.8,  
 vjust = 0.8) # Adjust margins to move labels left or right  
 )



# create linear model for VIF (collinearity) analysis, no scaling, no centering, numeric values only  
qol\_numeric <- qol\_data %>%   
 select(where(is.numeric))  
  
qol\_lm <- lm(  
 pct\_poor\_to\_fair\_health ~ .,  
 qol\_numeric  
)  
  
vif\_values <- car::vif(qol\_lm)  
  
vif\_values

## weighted\_population average\_hh\_size   
## 380.541480 3.685275   
## pct\_male pct\_native\_american   
## 1.577336 19.342350   
## pct\_asian pct\_black   
## 4.620903 72.376721   
## pct\_hispanic pct\_other\_race   
## 5.709361 7.042548   
## pct\_white pct\_single\_parent   
## 93.813894 2.976729   
## pct\_hh\_other\_computer pct\_hh\_internet   
## 1.086193 3.457420   
## pct\_employed pct\_hh\_inc\_99999   
## 2.253738 2.228843   
## pct\_w\_medicare clinical\_nurse\_pt   
## 2.175950 1.140358   
## dentist\_pt pa\_pt   
## 1.779709 1.363687   
## mental\_health\_faciliy\_pt population\_density   
## 1.126729 1.347411   
## days\_over\_90\_f median\_hh\_income   
## 2.871957 5.882129   
## median\_er\_dist median\_pediatric\_icu\_dist   
## 1.611019 1.665478   
## median\_health\_clinic\_dist median\_drug\_alcohol\_care\_dist   
## 1.171135 1.734912   
## hh\_tot\_workers pct\_vet   
## 383.791581 1.707336   
## pct\_uninsured pct\_assistance   
## 5.260978 4.603439   
## percent\_grandparents\_as\_guardians pct\_adult\_smokers   
## 1.924987 7.068890   
## pct\_binge\_drinkers pct\_under\_65\_no\_health\_insurance   
## 2.166994 6.129473   
## pct\_highschool\_diploma pct\_some\_college   
## 5.703708 4.183170   
## inequality\_ratio social\_clubs\_per\_10k   
## 2.366425 1.632428   
## air\_polution\_metric water\_quality   
## 2.157830 1.100237   
## pct\_high\_housing\_costs pct\_overcrowded\_hh   
## 2.419277 2.665506   
## pct\_30\_min\_plus\_commute pct\_food\_insecurities   
## 2.534443 4.731278   
## pct\_voters pct\_home\_owner   
## 3.471844 3.589868   
## pct\_65\_plus pct\_rural\_population   
## 5.071392 3.806509

# Create a data frame with GVIF and Df  
vif\_df <- data.frame(  
 Feature = names(vif\_values),  
 GVIF = vif\_values,  
 Df = rep(1, length(vif\_values)) # Df is typically 1 for univariate cases  
)  
  
# Calculate GVIF^(1/(2\*Df)) and add it to the data frame  
vif\_df$Adjusted\_VIF <- vif\_df$GVIF^(1/(2 \* vif\_df$Df))  
  
# Sort the data frame by Adjusted VIF in ascending order  
vif\_df <- vif\_df[order(vif\_df$Adjusted\_VIF, decreasing = TRUE), ]  
  
# Set a threshold for high collinearity   
high\_vif\_threshold <- 5  
  
# Filter for features with high collinearity values  
high\_vif\_features <- vif\_df[vif\_df[, "Adjusted\_VIF"] > high\_vif\_threshold, ]  
  
# Print the table  
print(high\_vif\_features)

## Feature GVIF Df Adjusted\_VIF  
## hh\_tot\_workers hh\_tot\_workers 383.79158 1 19.590599  
## weighted\_population weighted\_population 380.54148 1 19.507472  
## pct\_white pct\_white 93.81389 1 9.685757  
## pct\_black pct\_black 72.37672 1 8.507451

# `sp500` table data  
high\_vif\_features %>%   
 gt::gt() %>%   
 gt::tab\_header(  
 title = "Predictors With High Collineariy",  
 )

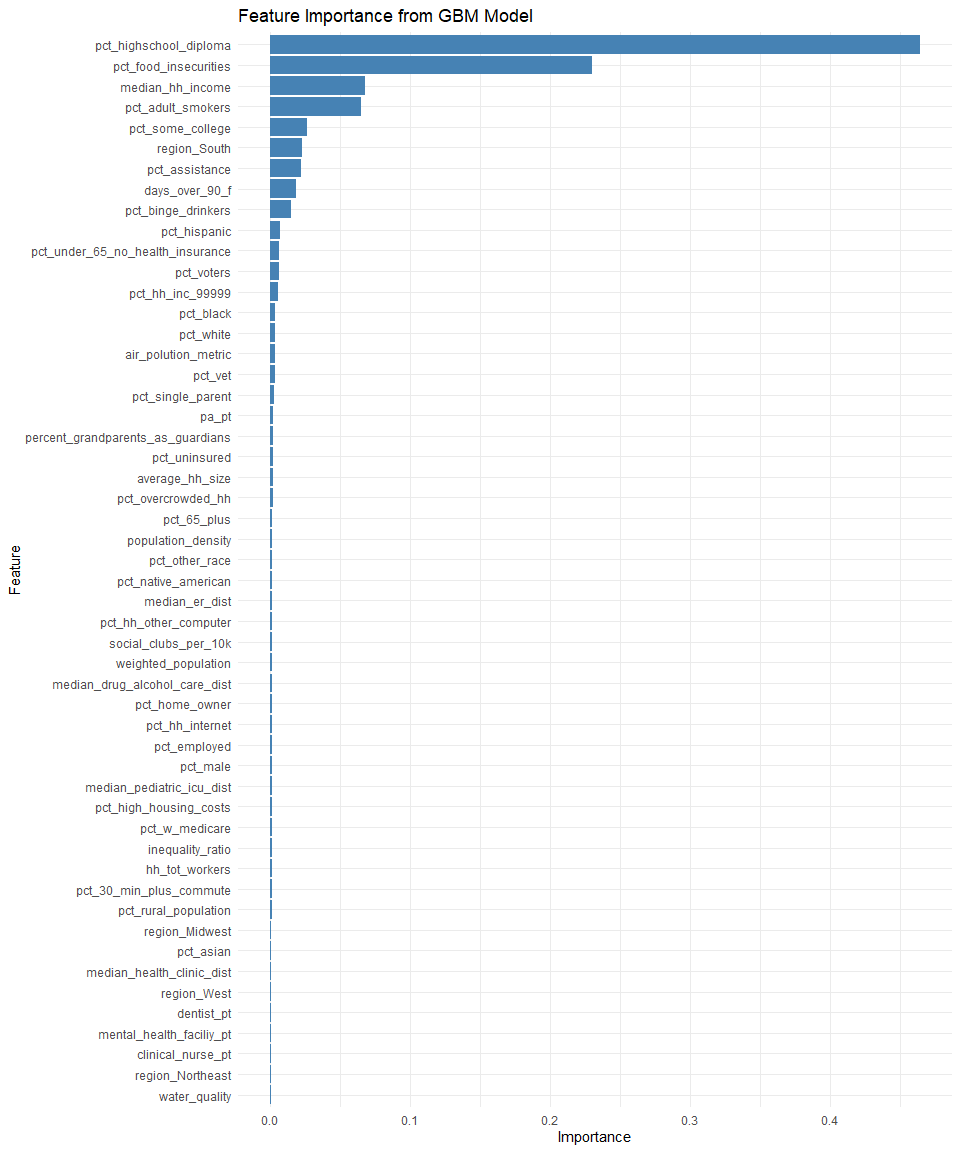
Table 1: Predictors With High Collineariy

| Feature | GVIF | Df | Adjusted\_VIF |
| --- | --- | --- | --- |
| hh\_tot\_workers | 383.79158 | 1 | 19.590599 |
| weighted\_population | 380.54148 | 1 | 19.507472 |
| pct\_white | 93.81389 | 1 | 9.685757 |
| pct\_black | 72.37672 | 1 | 8.507451 |

# Extract feature importance  
gbm\_feature\_importance <- varImp(  
 gbm\_model\_cv,  
 scale = FALSE  
 )  
  
# Convert to a data frame  
gbm\_feature\_importance\_df <- as.data.frame(  
 gbm\_feature\_importance$importance  
 )  
  
# Add feature names  
gbm\_feature\_importance\_df$Feature <- rownames(  
 gbm\_feature\_importance\_df  
 )  
  
# Rename columns for clarity  
colnames(gbm\_feature\_importance\_df) <- c(  
 "Importance",  
 "Feature"  
 )  
  
# Sort by importance  
gbm\_feature\_importance\_df <- gbm\_feature\_importance\_df %>%  
 arrange(desc(Importance))  
  
# Export to a CSV file  
write.csv(  
 gbm\_feature\_importance\_df,  
 "gbm\_feature\_importance.csv",  
 row.names = FALSE  
 )  
  
# Print the feature importance table  
print(gbm\_feature\_importance\_df)

## Importance  
## pct\_highschool\_diploma 4.643326e-01  
## pct\_food\_insecurities 2.297069e-01  
## median\_hh\_income 6.727079e-02  
## pct\_adult\_smokers 6.441548e-02  
## pct\_some\_college 2.592092e-02  
## region\_South 2.253109e-02  
## pct\_assistance 2.201346e-02  
## days\_over\_90\_f 1.801302e-02  
## pct\_binge\_drinkers 1.488110e-02  
## pct\_hispanic 6.557120e-03  
## pct\_under\_65\_no\_health\_insurance 6.374268e-03  
## pct\_voters 5.878899e-03  
## pct\_hh\_inc\_99999 5.689507e-03  
## pct\_black 3.468132e-03  
## pct\_white 3.450485e-03  
## air\_polution\_metric 3.250158e-03  
## pct\_vet 3.075703e-03  
## pct\_single\_parent 2.367880e-03  
## pa\_pt 2.046616e-03  
## percent\_grandparents\_as\_guardians 2.028129e-03  
## pct\_uninsured 1.821077e-03  
## average\_hh\_size 1.736171e-03  
## pct\_overcrowded\_hh 1.441646e-03  
## pct\_65\_plus 1.190921e-03  
## population\_density 1.171810e-03  
## pct\_other\_race 1.129866e-03  
## pct\_native\_american 1.127381e-03  
## median\_er\_dist 1.102490e-03  
## pct\_hh\_other\_computer 1.098252e-03  
## social\_clubs\_per\_10k 9.887691e-04  
## weighted\_population 9.279541e-04  
## median\_drug\_alcohol\_care\_dist 9.147313e-04  
## pct\_home\_owner 8.694203e-04  
## pct\_hh\_internet 8.486502e-04  
## pct\_employed 8.414767e-04  
## pct\_male 8.223924e-04  
## median\_pediatric\_icu\_dist 8.152863e-04  
## pct\_high\_housing\_costs 8.130358e-04  
## pct\_w\_medicare 8.114936e-04  
## inequality\_ratio 7.974180e-04  
## hh\_tot\_workers 7.473830e-04  
## pct\_30\_min\_plus\_commute 7.458699e-04  
## pct\_rural\_population 7.451414e-04  
## region\_Midwest 6.122739e-04  
## pct\_asian 5.155501e-04  
## median\_health\_clinic\_dist 4.873042e-04  
## region\_West 4.423313e-04  
## dentist\_pt 4.297006e-04  
## mental\_health\_faciliy\_pt 3.225875e-04  
## clinical\_nurse\_pt 1.730318e-04  
## region\_Northeast 1.639439e-04  
## water\_quality 7.238083e-05  
## Feature  
## pct\_highschool\_diploma pct\_highschool\_diploma  
## pct\_food\_insecurities pct\_food\_insecurities  
## median\_hh\_income median\_hh\_income  
## pct\_adult\_smokers pct\_adult\_smokers  
## pct\_some\_college pct\_some\_college  
## region\_South region\_South  
## pct\_assistance pct\_assistance  
## days\_over\_90\_f days\_over\_90\_f  
## pct\_binge\_drinkers pct\_binge\_drinkers  
## pct\_hispanic pct\_hispanic  
## pct\_under\_65\_no\_health\_insurance pct\_under\_65\_no\_health\_insurance  
## pct\_voters pct\_voters  
## pct\_hh\_inc\_99999 pct\_hh\_inc\_99999  
## pct\_black pct\_black  
## pct\_white pct\_white  
## air\_polution\_metric air\_polution\_metric  
## pct\_vet pct\_vet  
## pct\_single\_parent pct\_single\_parent  
## pa\_pt pa\_pt  
## percent\_grandparents\_as\_guardians percent\_grandparents\_as\_guardians  
## pct\_uninsured pct\_uninsured  
## average\_hh\_size average\_hh\_size  
## pct\_overcrowded\_hh pct\_overcrowded\_hh  
## pct\_65\_plus pct\_65\_plus  
## population\_density population\_density  
## pct\_other\_race pct\_other\_race  
## pct\_native\_american pct\_native\_american  
## median\_er\_dist median\_er\_dist  
## pct\_hh\_other\_computer pct\_hh\_other\_computer  
## social\_clubs\_per\_10k social\_clubs\_per\_10k  
## weighted\_population weighted\_population  
## median\_drug\_alcohol\_care\_dist median\_drug\_alcohol\_care\_dist  
## pct\_home\_owner pct\_home\_owner  
## pct\_hh\_internet pct\_hh\_internet  
## pct\_employed pct\_employed  
## pct\_male pct\_male  
## median\_pediatric\_icu\_dist median\_pediatric\_icu\_dist  
## pct\_high\_housing\_costs pct\_high\_housing\_costs  
## pct\_w\_medicare pct\_w\_medicare  
## inequality\_ratio inequality\_ratio  
## hh\_tot\_workers hh\_tot\_workers  
## pct\_30\_min\_plus\_commute pct\_30\_min\_plus\_commute  
## pct\_rural\_population pct\_rural\_population  
## region\_Midwest region\_Midwest  
## pct\_asian pct\_asian  
## median\_health\_clinic\_dist median\_health\_clinic\_dist  
## region\_West region\_West  
## dentist\_pt dentist\_pt  
## mental\_health\_faciliy\_pt mental\_health\_faciliy\_pt  
## clinical\_nurse\_pt clinical\_nurse\_pt  
## region\_Northeast region\_Northeast  
## water\_quality water\_quality

# Plot the feature importance  
ggplot(  
 gbm\_feature\_importance\_df,  
 aes(  
 x = reorder(Feature, Importance),  
 y = Importance)  
 ) +  
 geom\_bar(  
 stat = "identity",  
 fill = "steelblue"  
 ) +  
 coord\_flip() +  
 labs(  
 title = "Feature Importance from GBM Model",  
 x = "Feature",  
 y = "Importance"  
 ) +  
 theme\_minimal()



# Train the SVM model  
svm\_model <- e1071::svm(  
 pct\_poor\_to\_fair\_health ~ .,  
 data = qol\_train  
 )  
  
# Evaluate the model on the test data  
svm\_pred <- predict(  
 svm\_model,  
 qol\_test  
 )  
  
# Calculate evaluation metrics  
svm\_mse <- mean(  
 (svm\_pred - qol\_test$pct\_poor\_to\_fair\_health)^2  
 )  
  
svm\_rmse <- sqrt(  
 svm\_mse  
 )  
  
svm\_r\_squared <- 1 - sum(  
 (svm\_pred - qol\_test$pct\_poor\_to\_fair\_health)^2  
 ) /   
 sum(  
 (qol\_test$pct\_poor\_to\_fair\_health - mean(qol\_test$pct\_poor\_to\_fair\_health))^2  
 )  
  
# Print the evaluation metrics  
cat("SVM MSE:", svm\_mse, "\n")

## SVM MSE: 0.0001215293

cat("SVM RMSE:", svm\_rmse, "\n")

## SVM RMSE: 0.01102403

cat("SVM R-squared:", svm\_r\_squared, "\n")

## SVM R-squared: 0.9408336

# Compute variable importance  
svm\_importance <- vip::vip(  
 svm\_model,  
 train = qol\_train,  
 target = "pct\_poor\_to\_fair\_health",  
 method = "permute",  
 metric = "rmse",  
 pred\_wrapper = function(  
 model,  
 newdata  
 )   
 predict(  
 model  
 ,  
 newdata  
 ),   
 num\_features = ncol(qol\_train) - 1  
 )  
  
plot(svm\_importance)

