CDCsCHSIAnalysis

Caitlin Casey

December 17, 2019

knitr::opts\_chunk$set(echo = TRUE)  
library(cowplot)

## Warning: package 'cowplot' was built under R version 3.6.1

##   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

## Note: As of version 1.0.0, cowplot does not change the

## default ggplot2 theme anymore. To recover the previous

## behavior, execute:  
## theme\_set(theme\_cowplot())

## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

library(png)  
library(googleway)

## Warning: package 'googleway' was built under R version 3.6.1

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.1

library(ggrepel)

## Warning: package 'ggrepel' was built under R version 3.6.1

library(ggspatial)

## Warning: package 'ggspatial' was built under R version 3.6.1

library(lwgeom)

## Warning: package 'lwgeom' was built under R version 3.6.1

## Linking to liblwgeom 2.5.0dev r16016, GEOS 3.6.1, PROJ 4.9.3

library(sf)

## Warning: package 'sf' was built under R version 3.6.1

## Linking to GEOS 3.6.1, GDAL 2.2.3, PROJ 4.9.3

library(rnaturalearth)

## Warning: package 'rnaturalearth' was built under R version 3.6.1

library(rnaturalearthdata)

## Warning: package 'rnaturalearthdata' was built under R version 3.6.1

library(maps)

## Warning: package 'maps' was built under R version 3.6.1

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.6.1

##   
## Attaching package: 'dplyr'

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

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

library(sqldf)

## Warning: package 'sqldf' was built under R version 3.6.1

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

## Warning: package 'RSQLite' was built under R version 3.6.1

library(readr)  
library(reshape)

## Warning: package 'reshape' was built under R version 3.6.1

##   
## Attaching package: 'reshape'

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

## The following object is masked from 'package:cowplot':  
##   
## stamp

library(neuralnet)

## Warning: package 'neuralnet' was built under R version 3.6.1

##   
## Attaching package: 'neuralnet'

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

library(e1071)

## Warning: package 'e1071' was built under R version 3.6.1

library(kernlab)

## Warning: package 'kernlab' was built under R version 3.6.1

##   
## Attaching package: 'kernlab'

## The following object is masked from 'package:ggplot2':  
##   
## alpha

library(rgeos)

## Warning: package 'rgeos' was built under R version 3.6.1

## Loading required package: sp

## Warning: package 'sp' was built under R version 3.6.1

## rgeos version: 0.5-2, (SVN revision 621)  
## GEOS runtime version: 3.6.1-CAPI-1.10.1   
## Linking to sp version: 1.3-2   
## Polygon checking: TRUE

library(rgdal)

## Warning: package 'rgdal' was built under R version 3.6.1

## rgdal: version: 1.4-8, (SVN revision 845)  
## Geospatial Data Abstraction Library extensions to R successfully loaded  
## Loaded GDAL runtime: GDAL 2.2.3, released 2017/11/20  
## Path to GDAL shared files: C:/Users/Caitlin Casey/Documents/R/win-library/3.6/rgdal/gdal  
## GDAL binary built with GEOS: TRUE   
## Loaded PROJ.4 runtime: Rel. 4.9.3, 15 August 2016, [PJ\_VERSION: 493]  
## Path to PROJ.4 shared files: C:/Users/Caitlin Casey/Documents/R/win-library/3.6/rgdal/proj  
## Linking to sp version: 1.3-2

library(maptools)

## Warning: package 'maptools' was built under R version 3.6.1

## Checking rgeos availability: TRUE

library(rsample)

## Warning: package 'rsample' was built under R version 3.6.1

## Loading required package: tidyr

## Warning: package 'tidyr' was built under R version 3.6.1

##   
## Attaching package: 'tidyr'

## The following objects are masked from 'package:reshape':  
##   
## expand, smiths

library(scales)

## Warning: package 'scales' was built under R version 3.6.1

##   
## Attaching package: 'scales'

## The following object is masked from 'package:kernlab':  
##   
## alpha

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

library(shiny)

## Warning: package 'shiny' was built under R version 3.6.1

library(Rcmdr)

## Warning: package 'Rcmdr' was built under R version 3.6.1

## Loading required package: splines

## Loading required package: RcmdrMisc

## Warning: package 'RcmdrMisc' was built under R version 3.6.1

## Loading required package: car

## Warning: package 'car' was built under R version 3.6.1

## Loading required package: carData

## Warning: package 'carData' was built under R version 3.6.1

##   
## Attaching package: 'car'

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

## Loading required package: sandwich

## Warning: package 'sandwich' was built under R version 3.6.1

## Loading required package: effects

## Warning: package 'effects' was built under R version 3.6.1

## Registered S3 methods overwritten by 'lme4':  
## method from  
## cooks.distance.influence.merMod car   
## influence.merMod car   
## dfbeta.influence.merMod car   
## dfbetas.influence.merMod car

## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.

## The Commander GUI is launched only in interactive sessions

##   
## Attaching package: 'Rcmdr'

## The following object is masked from 'package:shiny':  
##   
## radioButtons

## The following object is masked from 'package:base':  
##   
## errorCondition

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.1

## -- Attaching packages --------------------------------------------------------------------------------- tidyverse 1.3.0 --

## <U+2713> tibble 2.1.3 <U+2713> stringr 1.4.0  
## <U+2713> purrr 0.3.3 <U+2713> forcats 0.4.0

## Warning: package 'purrr' was built under R version 3.6.1

## Warning: package 'stringr' was built under R version 3.6.1

## -- Conflicts ------------------------------------------------------------------------------------ tidyverse\_conflicts() --  
## x scales::alpha() masks kernlab::alpha(), ggplot2::alpha()  
## x scales::col\_factor() masks readr::col\_factor()  
## x neuralnet::compute() masks dplyr::compute()  
## x purrr::cross() masks kernlab::cross()  
## x purrr::discard() masks scales::discard()  
## x tidyr::expand() masks reshape::expand()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x purrr::map() masks maps::map()  
## x car::recode() masks dplyr::recode()  
## x reshape::rename() masks dplyr::rename()  
## x purrr::some() masks car::some()

library(RcmdrMisc)  
library(readr)

# Column Key

# Healthypeople is a reference sheet so we won’t join it. Instead, we’ll use it later for comparisons

healthypeople = read.csv("C:/Users/Caitlin Casey/Desktop/SYRACUSE/USA\_CHSI\_DATASET/HEALTHYPEOPLE2010.csv")

# Now we will combine the 5 datasets

demographics = read.csv("C:/Users/Caitlin Casey/Desktop/SYRACUSE/USA\_CHSI\_DATASET/DEMOGRAPHICS.csv")  
vulnpopsandenvhealth = read\_csv("C:/Users/Caitlin Casey/Desktop/SYRACUSE/USA\_CHSI\_DATASET/VUNERABLEPOPSANDENVHEALTH.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## State\_FIPS\_Code = col\_character(),  
## County\_FIPS\_Code = col\_character(),  
## CHSI\_County\_Name = col\_character(),  
## CHSI\_State\_Name = col\_character(),  
## CHSI\_State\_Abbr = col\_character(),  
## EH\_Time\_Span = col\_character()  
## )

## See spec(...) for full column specifications.

riskfactors = read.csv("C:/Users/Caitlin Casey/Desktop/SYRACUSE/USA\_CHSI\_DATASET/RISKFACTORSANDACCESSTOCARE.csv")  
leadingcausesofdeath = read.csv("C:/Users/Caitlin Casey/Desktop/SYRACUSE/USA\_CHSI\_DATASET/LEADINGCAUSESOFDEATH.csv")  
summarymeasures = read.csv("C:/Users/Caitlin Casey/Desktop/SYRACUSE/USA\_CHSI\_DATASET/SUMMARYMEASURESOFHEALTH.csv")  
   
combined <- sqldf('  
select   
 d.county\_fips\_code ,d.state\_fips\_code ,d.chsi\_county\_name ,d.chsi\_state\_name ,d.chsi\_state\_abbr ,d.strata\_id\_number ,d.strata\_determining\_factors ,d.number\_counties ,d.population\_size ,d.population\_density ,d.poverty ,d.Age\_19\_Under ,d.Age\_19\_64 ,d.Age\_65\_84, d.Age\_85\_and\_Over, d.White, d.Black, d.Native\_American, d.Asian, d.Hispanic, v.No\_HS\_Diploma, v.Unemployed, v.Sev\_Work\_Disabled, v.Major\_Depression, v.Recent\_Drug\_Use, v.Ecol\_Rpt, v.Ecol\_Rpt\_Ind, v.Ecol\_Exp, v.Salm\_Rpt, v.Salm\_Rpt\_Ind , v.Salm\_Exp, v.Shig\_Rpt, v.Shig\_Rpt\_Ind, v.Shig\_Exp, v.Toxic\_Chem, v.Carbon\_Monoxide\_Ind, v.Nitrogen\_Dioxide\_Ind, v.Sulfur\_Dioxide\_Ind, v.Ozone\_Ind, v.Particulate\_Matter\_Ind, v.Lead\_Ind, v.EH\_Time\_Span, r.No\_Exercise, r.Few\_Fruit\_Veg, r.Obesity, r.High\_Blood\_Pres, r.Smoker, r.Diabetes, r.Uninsured, r.Elderly\_Medicare, r.Disabled\_Medicare, r.Prim\_Care\_Phys\_Rate, r.Dentist\_Rate, r.Community\_Health\_Center\_Ind, r.HPSA\_Ind, l.A\_Wh\_Comp, l.A\_Bl\_Comp, l.A\_Ot\_Comp, l.A\_Hi\_Comp, l.A\_Wh\_BirthDef, l.A\_Bl\_BirthDef, l.A\_Ot\_BirthDef, l.A\_Hi\_BirthDef, l.B\_Wh\_Injury, l.B\_Bl\_Injury, l.B\_Ot\_Injury, l.B\_Hi\_Injury, l.B\_Wh\_Cancer, l.B\_Bl\_Cancer, l.B\_Ot\_Cancer, l.B\_Hi\_Cancer, l.B\_Wh\_Homicide, l.B\_Bl\_Homicide, l.B\_Ot\_Homicide, l.B\_Hi\_Homicide, l.C\_Wh\_Injury, l.C\_Bl\_Injury, l.C\_Ot\_Injury, l.C\_Hi\_Injury, l.C\_Wh\_Homicide, l.C\_Bl\_Homicide, l.C\_Ot\_homicide, l.C\_Hi\_Homicide, l.C\_Wh\_Suicide, l.C\_Bl\_Suicide, l.C\_Ot\_Suicide, l.C\_Hi\_Suicide, l.C\_Wh\_Cancer, l.C\_Bl\_Cancer, l.C\_Ot\_Cancer, l.C\_Hi\_Cancer, l.D\_Wh\_Injury, l.D\_Bl\_Injury, l.D\_Ot\_Injury, l.D\_Hi\_Injury, l.D\_Wh\_Cancer, l.D\_Bl\_Cancer, l.D\_Ot\_Cancer, l.D\_Hi\_Cancer, l.D\_Wh\_HeartDis, l.D\_Bl\_HeartDis, l.D\_Ot\_HeartDis, l.D\_Hi\_HeartDis, l.D\_Wh\_Suicide, l.D\_Bl\_Suicide, l.D\_Ot\_Suicide, l.D\_Hi\_Suicide, l.D\_Wh\_HIV, l.D\_Bl\_HIV, l.D\_Ot\_HIV, l.D\_Hi\_HIV, l.D\_Wh\_Homicide, l.D\_Bl\_Homicide, l.D\_Ot\_Homicide, l.D\_Hi\_Homicide, l.E\_Wh\_Cancer, l.E\_Bl\_Cancer, l.E\_Ot\_Cancer, l.E\_Hi\_Cancer, l.E\_Wh\_HeartDis, l.E\_Bl\_HeartDis, l.E\_Ot\_HeartDis, l.E\_Hi\_HeartDis, l.F\_Wh\_HeartDis, l.F\_Bl\_HeartDis, l.F\_Ot\_HeartDis, l.F\_Hi\_HeartDis, l.F\_Wh\_Cancer, l.F\_Bl\_Cancer, l.F\_Ot\_Cancer, l.F\_Hi\_Cancer, l.LCD\_Time\_Span, s.ALE, s.US\_ALE, s.All\_Death ,s.US\_All\_Death, s.Health\_Status, s.US\_Health\_Status, s.Unhealthy\_Days  
,s.US\_Unhealthy\_Days   
from demographics d   
left join vulnpopsandenvhealth v on v.county\_fips\_code=d.county\_fips\_code and v.state\_fips\_code=d.state\_fips\_code  
left join riskfactors r on r.county\_fips\_code=d.county\_fips\_code and r.state\_fips\_code=d.state\_fips\_code  
left join leadingcausesofdeath l on l.county\_fips\_code=d.county\_fips\_code and l.state\_fips\_code=d.state\_fips\_code  
left join summarymeasures s on s.county\_fips\_code=d.county\_fips\_code and s.state\_fips\_code=d.state\_fips\_code  
' )

# We know that negative numbers are really NAs; let’s replace them

combined[,-1:-7] <- data.frame(lapply(combined[,-1:-7], function(x){  
 as.numeric(gsub("-1\*|-2\*|-9\*",NA,x))  
}))  
  
str(combined)

## 'data.frame': 3141 obs. of 140 variables:  
## $ County\_FIPS\_Code : int 1 3 5 7 9 11 13 15 17 19 ...  
## $ State\_FIPS\_Code : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ CHSI\_County\_Name : chr "Autauga" "Baldwin" "Barbour" "Bibb" ...  
## $ CHSI\_State\_Name : chr "Alabama" "Alabama" "Alabama" "Alabama" ...  
## $ CHSI\_State\_Abbr : chr "AL" "AL" "AL" "AL" ...  
## $ Strata\_ID\_Number : int 29 16 51 42 28 75 76 6 50 64 ...  
## $ Strata\_Determining\_Factors : Factor w/ 4 levels "frontier status, population size",..: 3 3 4 3 3 4 4 2 4 4 ...  
## $ Number\_Counties : num 37 27 33 53 39 37 38 53 27 41 ...  
## $ Population\_Size : num 48612 162586 28414 21516 55725 ...  
## $ Population\_Density : num 82 102 32 35 86 18 27 184 59 44 ...  
## $ Poverty : num 10.4 10.2 22.1 16.8 11.9 26.2 20 16.4 16.2 15.2 ...  
## $ Age\_19\_Under : num 26.9 23.5 24.3 24.6 24.5 24.7 25.6 24.1 24.8 21.9 ...  
## $ Age\_19\_64 : num 62.3 60.3 62.5 63.3 62.1 63.2 58.5 61.6 59.5 61.4 ...  
## $ Age\_65\_84 : num 9.8 14.5 11.6 10.9 12.1 10 13.6 12.7 13.5 15.2 ...  
## $ Age\_85\_and\_Over : num 0.9 1.8 1.6 1.2 1.3 2.2 2.4 1.5 2.2 1.4 ...  
## $ White : num 80.7 88.4 52.2 76.8 97.1 27.8 57.5 78.3 60.8 93.3 ...  
## $ Black : num 17.3 9.9 46.8 22.5 1.5 71.4 41.9 19.7 38.6 5.5 ...  
## $ Native\_American : num 0.5 0.5 0.4 0.3 0.5 0.4 0.2 0.4 0.1 0.3 ...  
## $ Asian : num 0.6 0.4 0.3 0.1 0.2 0.2 0.3 0.8 0.2 0.3 ...  
## $ Hispanic : num 1.7 2.3 3.1 1.4 6.3 5.9 0.9 2 1.2 1.1 ...  
## $ No\_HS\_Diploma : num 6690 20254 6729 5355 11181 ...  
## $ Unemployed : num 774 2533 569 358 819 ...  
## $ Sev\_Work\_Disabled : num 1727 4933 1302 900 2217 ...  
## $ Major\_Depression : num 2680 9354 1618 1218 3164 ...  
## $ Recent\_Drug\_Use : num 2394 7753 1403 1034 2675 ...  
## $ Ecol\_Rpt : num 2 2 0 2 1 0 0 0 0 1 ...  
## $ Ecol\_Rpt\_Ind : num 3 3 3 4 3 3 3 3 3 3 ...  
## $ Ecol\_Exp : num 4 4 0 2 5 1 0 3 2 2 ...  
## $ Salm\_Rpt : num 50 99 53 9 25 17 23 31 32 6 ...  
## $ Salm\_Rpt\_Ind : num 4 4 4 3 3 3 3 3 4 3 ...  
## $ Salm\_Exp : num 31 67 29 32 31 20 39 54 29 35 ...  
## $ Shig\_Rpt : num 4 41 1 3 11 7 5 33 1 6 ...  
## $ Shig\_Rpt\_Ind : num 3 4 3 3 4 4 3 4 3 4 ...  
## $ Shig\_Exp : num 11 13 16 16 5 7 17 33 9 9 ...  
## $ Toxic\_Chem : num 2883197 35664 49874 7302 889020 ...  
## $ Carbon\_Monoxide\_Ind : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Nitrogen\_Dioxide\_Ind : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Sulfur\_Dioxide\_Ind : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Ozone\_Ind : num 1 2 1 1 1 1 1 1 1 1 ...  
## $ Particulate\_Matter\_Ind : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Lead\_Ind : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ EH\_Time\_Span : num NA NA NA NA NA NA NA NA NA NA ...  
## $ No\_Exercise : num 27.8 27.2 NA NA 33.5 NA 24.5 29.2 34.7 NA ...  
## $ Few\_Fruit\_Veg : num 78.6 76.2 NA 86.6 74.6 NA NA 81.9 84.6 NA ...  
## $ Obesity : num 24.5 23.6 25.6 NA 24.2 NA 22 27 NA NA ...  
## $ High\_Blood\_Pres : num 29.1 30.5 NA NA NA NA NA 33.2 NA NA ...  
## $ Smoker : num 26.6 24.6 17.7 NA 23.6 NA 27.3 25.5 11.7 23.6 ...  
## $ Diabetes : num 14.2 7.2 6.6 13.1 8.4 NA 9.5 11.2 17.1 13.3 ...  
## $ Uninsured : num 5690 19798 5126 3315 8131 ...  
## $ Elderly\_Medicare : num 4762 22635 3288 2390 5019 ...  
## $ Disabled\_Medicare : num 1209 3839 1092 974 1300 ...  
## $ Prim\_Care\_Phys\_Rate : num 45.3 67 45.8 41.8 16.2 54.3 43.3 75.8 53.6 28.5 ...  
## $ Dentist\_Rate : num 22.6 30.8 24.6 18.6 10.8 18.1 19.3 41.9 22.6 12.2 ...  
## $ Community\_Health\_Center\_Ind: num 1 1 1 1 2 1 1 1 1 2 ...  
## $ HPSA\_Ind : num 2 2 2 1 1 1 2 2 2 1 ...  
## $ A\_Wh\_Comp : num NA 57 NA NA 34 NA NA 36 NA NA ...  
## $ A\_Bl\_Comp : num NA NA NA NA NA NA NA NA NA NA ...  
## $ A\_Ot\_Comp : num NA NA NA NA NA NA NA NA NA NA ...  
## $ A\_Hi\_Comp : num NA NA NA NA NA NA NA NA NA NA ...  
## $ A\_Wh\_BirthDef : num NA 21 NA NA 34 NA NA 14 NA NA ...  
## $ A\_Bl\_BirthDef : num NA NA NA NA NA NA NA NA NA NA ...  
## $ A\_Ot\_BirthDef : num NA NA NA NA NA NA NA NA NA NA ...  
## $ A\_Hi\_BirthDef : num NA NA NA NA NA NA NA NA NA NA ...  
## $ B\_Wh\_Injury : num NA NA NA NA NA NA NA NA NA NA ...  
## $ B\_Bl\_Injury : num NA NA NA NA NA NA NA NA NA NA ...  
## $ B\_Ot\_Injury : num NA NA NA NA NA NA NA NA NA NA ...  
## $ B\_Hi\_Injury : num NA NA NA NA NA NA NA NA NA NA ...  
## $ B\_Wh\_Cancer : num NA NA NA NA NA NA NA NA NA NA ...  
## $ B\_Bl\_Cancer : num NA NA NA NA NA NA NA NA NA NA ...  
## $ B\_Ot\_Cancer : num NA NA NA NA NA NA NA NA NA NA ...  
## $ B\_Hi\_Cancer : num NA NA NA NA NA NA NA NA NA NA ...  
## $ B\_Wh\_Homicide : num NA NA NA NA NA NA NA NA NA NA ...  
## $ B\_Bl\_Homicide : num NA NA NA NA NA NA NA NA NA NA ...  
## $ B\_Ot\_Homicide : num NA NA NA NA NA NA NA NA NA NA ...  
## $ B\_Hi\_Homicide : num NA NA NA NA NA NA NA NA NA NA ...  
## $ C\_Wh\_Injury : num NA 63 NA 63 64 NA 67 57 NA 68 ...  
## $ C\_Bl\_Injury : num NA NA NA NA NA NA 20 NA NA NA ...  
## $ C\_Ot\_Injury : num NA NA NA NA NA NA NA NA NA NA ...  
## $ C\_Hi\_Injury : num NA NA NA NA NA NA NA NA NA NA ...  
## $ C\_Wh\_Homicide : num NA NA NA NA NA NA NA NA NA NA ...  
## $ C\_Bl\_Homicide : num NA NA NA NA NA NA 13 NA NA NA ...  
## $ C\_Ot\_homicide : num NA NA NA NA NA NA NA NA NA NA ...  
## $ C\_Hi\_Homicide : num NA NA NA NA NA NA NA NA NA NA ...  
## $ C\_Wh\_Suicide : num NA NA NA 13 NA NA 13 11 NA 12 ...  
## $ C\_Bl\_Suicide : num NA NA NA NA NA NA 23 NA NA NA ...  
## $ C\_Ot\_Suicide : num NA NA NA NA NA NA NA NA NA NA ...  
## $ C\_Hi\_Suicide : num NA NA NA NA NA NA NA NA NA NA ...  
## $ C\_Wh\_Cancer : num NA NA NA NA NA NA NA NA NA NA ...  
## $ C\_Bl\_Cancer : num NA NA NA NA NA NA NA NA NA NA ...  
## $ C\_Ot\_Cancer : num NA NA NA NA NA NA NA NA NA NA ...  
## $ C\_Hi\_Cancer : num NA NA NA NA NA NA NA NA NA NA ...  
## $ D\_Wh\_Injury : num 24 30 18 29 22 40 33 33 25 31 ...  
## $ D\_Bl\_Injury : num 24 13 19 24 NA 22 22 13 22 NA ...  
## $ D\_Ot\_Injury : num NA NA NA NA NA NA NA NA NA NA ...  
## $ D\_Hi\_Injury : num NA NA NA NA NA NA NA NA NA NA ...  
## $ D\_Wh\_Cancer : num NA 11 21 16 16 10 12 NA 12 12 ...  
## $ D\_Bl\_Cancer : num 18 NA NA 21 NA NA 10 11 17 NA ...  
## $ D\_Ot\_Cancer : num NA NA NA NA NA NA NA NA NA NA ...  
## $ D\_Hi\_Cancer : num NA NA NA NA NA NA NA NA NA NA ...  
## [list output truncated]

write\_csv(combined,'combined.csv')

# Replace NAs with mean of column auto

na\_mean\_swap <- function(x) {  
 replace(x, is.na(x),mean(as.numeric(x),na.rm=TRUE))  
}  
mean\_clean <- cbind(combined[,1:7],replace(combined[,-1:-7],TRUE, lapply(combined[,-1:-7], na\_mean\_swap)))  
str(mean\_clean)

## 'data.frame': 3141 obs. of 140 variables:  
## $ County\_FIPS\_Code : int 1 3 5 7 9 11 13 15 17 19 ...  
## $ State\_FIPS\_Code : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ CHSI\_County\_Name : chr "Autauga" "Baldwin" "Barbour" "Bibb" ...  
## $ CHSI\_State\_Name : chr "Alabama" "Alabama" "Alabama" "Alabama" ...  
## $ CHSI\_State\_Abbr : chr "AL" "AL" "AL" "AL" ...  
## $ Strata\_ID\_Number : int 29 16 51 42 28 75 76 6 50 64 ...  
## $ Strata\_Determining\_Factors : Factor w/ 4 levels "frontier status, population size",..: 3 3 4 3 3 4 4 2 4 4 ...  
## $ Number\_Counties : num 37 27 33 53 39 37 38 53 27 41 ...  
## $ Population\_Size : num 48612 162586 28414 21516 55725 ...  
## $ Population\_Density : num 82 102 32 35 86 18 27 184 59 44 ...  
## $ Poverty : num 10.4 10.2 22.1 16.8 11.9 26.2 20 16.4 16.2 15.2 ...  
## $ Age\_19\_Under : num 26.9 23.5 24.3 24.6 24.5 24.7 25.6 24.1 24.8 21.9 ...  
## $ Age\_19\_64 : num 62.3 60.3 62.5 63.3 62.1 63.2 58.5 61.6 59.5 61.4 ...  
## $ Age\_65\_84 : num 9.8 14.5 11.6 10.9 12.1 10 13.6 12.7 13.5 15.2 ...  
## $ Age\_85\_and\_Over : num 0.9 1.8 1.6 1.2 1.3 2.2 2.4 1.5 2.2 1.4 ...  
## $ White : num 80.7 88.4 52.2 76.8 97.1 27.8 57.5 78.3 60.8 93.3 ...  
## $ Black : num 17.3 9.9 46.8 22.5 1.5 71.4 41.9 19.7 38.6 5.5 ...  
## $ Native\_American : num 0.5 0.5 0.4 0.3 0.5 0.4 0.2 0.4 0.1 0.3 ...  
## $ Asian : num 0.6 0.4 0.3 0.1 0.2 0.2 0.3 0.8 0.2 0.3 ...  
## $ Hispanic : num 1.7 2.3 3.1 1.4 6.3 5.9 0.9 2 1.2 1.1 ...  
## $ No\_HS\_Diploma : num 6690 20254 6729 5355 11181 ...  
## $ Unemployed : num 774 2533 569 358 819 ...  
## $ Sev\_Work\_Disabled : num 1727 4933 1302 900 2217 ...  
## $ Major\_Depression : num 2680 9354 1618 1218 3164 ...  
## $ Recent\_Drug\_Use : num 2394 7753 1403 1034 2675 ...  
## $ Ecol\_Rpt : num 2 2 0 2 1 0 0 0 0 1 ...  
## $ Ecol\_Rpt\_Ind : num 3 3 3 4 3 3 3 3 3 3 ...  
## $ Ecol\_Exp : num 4 4 0 2 5 1 0 3 2 2 ...  
## $ Salm\_Rpt : num 50 99 53 9 25 17 23 31 32 6 ...  
## $ Salm\_Rpt\_Ind : num 4 4 4 3 3 3 3 3 4 3 ...  
## $ Salm\_Exp : num 31 67 29 32 31 20 39 54 29 35 ...  
## $ Shig\_Rpt : num 4 41 1 3 11 7 5 33 1 6 ...  
## $ Shig\_Rpt\_Ind : num 3 4 3 3 4 4 3 4 3 4 ...  
## $ Shig\_Exp : num 11 13 16 16 5 7 17 33 9 9 ...  
## $ Toxic\_Chem : num 2883197 35664 49874 7302 889020 ...  
## $ Carbon\_Monoxide\_Ind : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Nitrogen\_Dioxide\_Ind : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Sulfur\_Dioxide\_Ind : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Ozone\_Ind : num 1 2 1 1 1 1 1 1 1 1 ...  
## $ Particulate\_Matter\_Ind : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Lead\_Ind : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ EH\_Time\_Span : num NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ...  
## $ No\_Exercise : num 27.8 27.2 26.5 26.5 33.5 ...  
## $ Few\_Fruit\_Veg : num 78.6 76.2 78.9 86.6 74.6 ...  
## $ Obesity : num 24.5 23.6 25.6 24.1 24.2 ...  
## $ High\_Blood\_Pres : num 29.1 30.5 26.5 26.5 26.5 ...  
## $ Smoker : num 26.6 24.6 17.7 23.1 23.6 ...  
## $ Diabetes : num 14.2 7.2 6.6 13.1 8.4 ...  
## $ Uninsured : num 5690 19798 5126 3315 8131 ...  
## $ Elderly\_Medicare : num 4762 22635 3288 2390 5019 ...  
## $ Disabled\_Medicare : num 1209 3839 1092 974 1300 ...  
## $ Prim\_Care\_Phys\_Rate : num 45.3 67 45.8 41.8 16.2 54.3 43.3 75.8 53.6 28.5 ...  
## $ Dentist\_Rate : num 22.6 30.8 24.6 18.6 10.8 18.1 19.3 41.9 22.6 12.2 ...  
## $ Community\_Health\_Center\_Ind: num 1 1 1 1 2 1 1 1 1 2 ...  
## $ HPSA\_Ind : num 2 2 2 1 1 1 2 2 2 1 ...  
## $ A\_Wh\_Comp : num 45.1 57 45.1 45.1 34 ...  
## $ A\_Bl\_Comp : num 57.6 57.6 57.6 57.6 57.6 ...  
## $ A\_Ot\_Comp : num 43.4 43.4 43.4 43.4 43.4 ...  
## $ A\_Hi\_Comp : num 49.7 49.7 49.7 49.7 49.7 ...  
## $ A\_Wh\_BirthDef : num 24.2 21 24.2 24.2 34 ...  
## $ A\_Bl\_BirthDef : num 15.4 15.4 15.4 15.4 15.4 15.4 15.4 15.4 15.4 15.4 ...  
## $ A\_Ot\_BirthDef : num 22.7 22.7 22.7 22.7 22.7 ...  
## $ A\_Hi\_BirthDef : num 24.5 24.5 24.5 24.5 24.5 ...  
## $ B\_Wh\_Injury : num 37.4 37.4 37.4 37.4 37.4 ...  
## $ B\_Bl\_Injury : num 32.8 32.8 32.8 32.8 32.8 ...  
## $ B\_Ot\_Injury : num 39 39 39 39 39 ...  
## $ B\_Hi\_Injury : num 34 34 34 34 34 34 34 34 34 34 ...  
## $ B\_Wh\_Cancer : num 16.7 16.7 16.7 16.7 16.7 ...  
## $ B\_Bl\_Cancer : num 13.4 13.4 13.4 13.4 13.4 ...  
## $ B\_Ot\_Cancer : num 15.9 15.9 15.9 15.9 15.9 ...  
## $ B\_Hi\_Cancer : num 16.8 16.8 16.8 16.8 16.8 ...  
## $ B\_Wh\_Homicide : num 12.9 12.9 12.9 12.9 12.9 ...  
## $ B\_Bl\_Homicide : num 15.3 15.3 15.3 15.3 15.3 ...  
## $ B\_Ot\_Homicide : num 10 10 10 10 10 10 10 10 10 10 ...  
## $ B\_Hi\_Homicide : num 12 12 12 12 12 12 12 12 12 12 ...  
## $ C\_Wh\_Injury : num 57.8 63 57.8 63 64 ...  
## $ C\_Bl\_Injury : num 31.4 31.4 31.4 31.4 31.4 ...  
## $ C\_Ot\_Injury : num 47.7 47.7 47.7 47.7 47.7 ...  
## $ C\_Hi\_Injury : num 46.6 46.6 46.6 46.6 46.6 ...  
## $ C\_Wh\_Homicide : num 14.5 14.5 14.5 14.5 14.5 ...  
## $ C\_Bl\_Homicide : num 33.7 33.7 33.7 33.7 33.7 ...  
## $ C\_Ot\_homicide : num 18.5 18.5 18.5 18.5 18.5 ...  
## $ C\_Hi\_Homicide : num 24 24 24 24 24 ...  
## $ C\_Wh\_Suicide : num 17.2 17.2 17.2 13 17.2 ...  
## $ C\_Bl\_Suicide : num 13.6 13.6 13.6 13.6 13.6 ...  
## $ C\_Ot\_Suicide : num 22 22 22 22 22 ...  
## $ C\_Hi\_Suicide : num 15.7 15.7 15.7 15.7 15.7 ...  
## $ C\_Wh\_Cancer : num 12.7 12.7 12.7 12.7 12.7 ...  
## $ C\_Bl\_Cancer : num 12.2 12.2 12.2 12.2 12.2 ...  
## $ C\_Ot\_Cancer : num 13.7 13.7 13.7 13.7 13.7 ...  
## $ C\_Hi\_Cancer : num 12.8 12.8 12.8 12.8 12.8 ...  
## $ D\_Wh\_Injury : num 24 30 18 29 22 40 33 33 25 31 ...  
## $ D\_Bl\_Injury : num 24 13 19 24 18.4 ...  
## $ D\_Ot\_Injury : num 26.2 26.2 26.2 26.2 26.2 ...  
## $ D\_Hi\_Injury : num 30.9 30.9 30.9 30.9 30.9 ...  
## $ D\_Wh\_Cancer : num 17.6 11 21 16 16 ...  
## $ D\_Bl\_Cancer : num 18 15 15 21 15 ...  
## $ D\_Ot\_Cancer : num 20.3 20.3 20.3 20.3 20.3 ...  
## $ D\_Hi\_Cancer : num 15.9 15.9 15.9 15.9 15.9 ...  
## [list output truncated]

##### Bar Graphs for Demographics

# Transform demographics for valid columns

# subset data to only include some columns

# clean names of the subset

nms\_demo\_dat <- c("county.name","state.name","state.abbr","pop.size","pop.density",  
 "poverty","age.19\_under","age.19\_64","age.65\_84","age.85\_over",  
 "white","black","nat.amer","asian","hispanic")

# change col names

names(demographics)<-nms\_demo\_dat

# This data is a representation at the county level- #we have to do some work to find it at state level

mean(demographics$pop.size)

## [1] 94368.16

max(demographics$pop.size)

## [1] 9935475

min(demographics$pop.size)

## [1] 62

sd(demographics$pop.size)

## [1] 306431.7

# Build vectors

county <- c(demographics$county.name)  
state <- c(demographics$state.name)  
pop <- c(demographics$pop.size)

# now we assemble into data frame from vectors

df <- data.frame(county, state, pop)

# assembled pipe

stateandmeanpop <- df %>% group\_by(state) %>% summarize(mean\_pop = mean(pop))  
stateandsumpop <- df %>% group\_by(state) %>% summarize(mean\_pop = sum(pop))

# Now we have a dataframe we can make different charts from #with each states mean population

# This is the true mean of the population in our dataset

mean(stateandmeanpop$mean\_pop)

## [1] 139279.6

# This is the range of the population in our dataset

max(stateandmeanpop$mean\_pop)

## [1] 622968.1

min(stateandmeanpop$mean\_pop)

## [1] 11756.56

sd(stateandmeanpop$mean\_pop)

## [1] 148119.4

# Show mean age, and ranges

demographics$age.19\_underraw = (demographics$pop.size\*demographics$age.19\_under/100)  
demographics$age.19\_64raw = (demographics$pop.size\*demographics$age.19\_64/100)  
demographics$age.65\_84raw = (demographics$pop.size\*demographics$age.65\_84/100)  
demographics$age.85\_overraw = (demographics$pop.size\*demographics$age.85\_over/100)  
xviiii\_U <- mean(demographics$age.19\_under)  
max(demographics$age.19\_under)

## [1] 47.2

min(demographics$age.19\_under)

## [1] 1.4

xvii\_lxiv <- mean(demographics$age.19\_64)  
max(demographics$age.19\_64)

## [1] 83.3

min(demographics$age.19\_64)

## [1] 47.6

lxv\_lxxxiv <- mean(demographics$age.65\_84)  
max(demographics$age.65\_84)

## [1] 29.2

min(demographics$age.65\_84)

## [1] 2.1

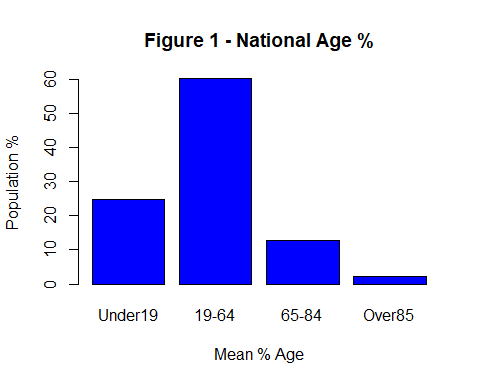
lxxxv\_O <- mean(demographics$age.85\_over)  
max(demographics$age.85\_over)

## [1] 7.6

min(demographics$age.85\_over)

## [1] 0.1

meanofage <- c(xviiii\_U, xvii\_lxiv, lxv\_lxxxiv, lxxxv\_O)  
barplot (meanofage,  
 main = "Figure 1 - National Age %",  
 xlab = "Mean % Age",  
 ylab = "Population %",  
 names.arg = c("Under19", "19-64", "65-84", "Over85"),  
 col = "blue",  
 horiz = FALSE)

 #Show ethnicity nationally #But how do we create raw population numbers for age and ethnicity? #First we have to add new columns in the data.frame with the raw data

demographics$whiteraw = (demographics$pop.size\*demographics$white/100)  
demographics$blackraw = (demographics$pop.size\*demographics$black/100)  
demographics$hispanicraw = (demographics$pop.size\*demographics$hispanic/100)  
demographics$asianraw = (demographics$pop.size\*demographics$asian/100)  
demographics$nat.amerraw = (demographics$pop.size\*demographics$nat.amer/100)  
total\_population <- sum(demographics$whiteraw) + sum(demographics$blackraw) + sum(demographics$hispanicraw) + sum(demographics$asianraw) + sum(demographics$nat.amerraw)

# Create mean vectors for each ethnicity

w <- mean(demographics$whiteraw)  
b <- mean(demographics$blackraw)  
h <- mean(demographics$hispanicraw)  
a <- mean(demographics$asianraw)  
n.a <- mean(demographics$nat.amerraw)

# Create dataframe

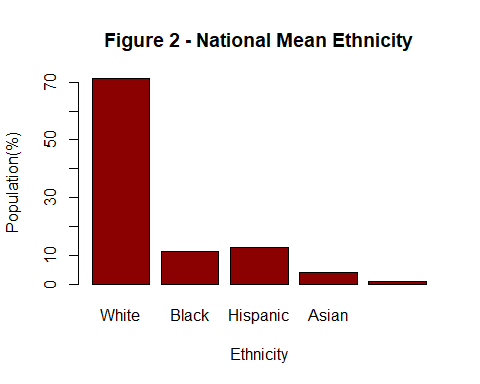
meanofrace <- c(w,b,h,a,n.a)

# converting into percentages

w\_percent <- (sum(demographics$whiteraw)/total\_population)\*100  
b\_percent <- (sum(demographics$blackraw)/total\_population)\*100  
h\_percent <- (sum(demographics$hispanicraw)/total\_population)\*100  
a\_percent <- (sum(demographics$asianraw)/total\_population)\*100  
nat\_percent <- (sum(demographics$nat.amerraw)/total\_population)\*100  
percentofrace <- c(w\_percent,b\_percent,h\_percent,a\_percent,nat\_percent)

# Create bar chart

barplot (percentofrace,  
 main = "Figure 2 - National Mean Ethnicity",  
 xlab = "Ethnicity",  
 ylab = "Population(%)",  
 names.arg = c("White", "Black", "Hispanic", "Asian", "NativeAmerican"),  
 col = "darkred",  
 horiz = FALSE)

 ############# Barchart Summary of Variables Most Affecting Death ############# ##################################################################### #this is from the section which takes population and percent of population into #consideration. It’s a true average of the population as values are originally presented #as percentage of county. #This calculates the count then takes percent of the US

# Convert raw count data into percentages so we can compare it across counties with different populations

vuln$No\_HS\_Diploma <- vuln$No\_HS\_Diploma/vuln$Population\_Size\*100  
vuln$Unemployed <- vuln$Unemployed/vuln$Population\_Size\*100  
vuln$Sev\_Work\_Disabled <-vuln$Sev\_Work\_Disabled/vuln$Population\_Size\*100  
vuln$Major\_Depression <- vuln$Major\_Depression/vuln$Population\_Size\*100  
vuln$Recent\_Drug\_Use <- vuln$Recent\_Drug\_Use/vuln$Population\_Size\*100  
vuln$Uninsured <- vuln$Uninsured/vuln$Population\_Size\*100  
vuln$Elderly\_Medicare <- vuln$Elderly\_Medicare/vuln$Population\_Size\*100  
vuln$Disabled\_Medicare <- vuln$Disabled\_Medicare/vuln$Population\_Size\*100  
vuln$DeathRate <- vuln$All\_Death/vuln$Population\_Size\*100  
  
DF\_Bar <- sqldf( 'select Poverty,Unemployed,Sev\_Work\_Disabled,Recent\_Drug\_Use,No\_Exercise,Obesity,High\_Blood\_Pres,Smoker,Diabetes,Uninsured,Elderly\_Medicare, Disabled\_Medicare   
 from vuln')  
PLot\_Bar <- replace(DF\_Bar, TRUE, lapply(DF\_Bar, na\_mean\_swap))

# Calculate mean of each column

PLot\_Bar <- colMeans(PLot\_Bar)

# Implement melt

library(reshape)

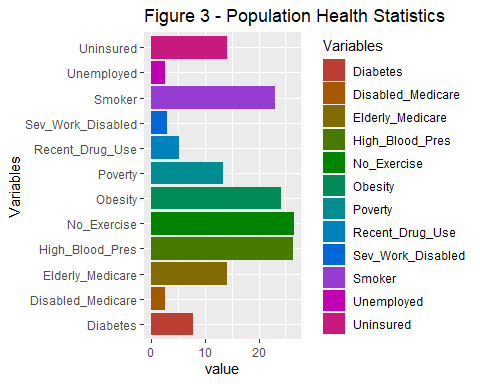
# this makes each row a unique id-variable combination

PLot\_Bar <- melt(PLot\_Bar)  
PLot\_Bar$Variables <- rownames(PLot\_Bar)

# used Cookbook for R, Colors(ggplot2) for assistance

# toned the colors down to be a bit darker and match the colors of our other graphs

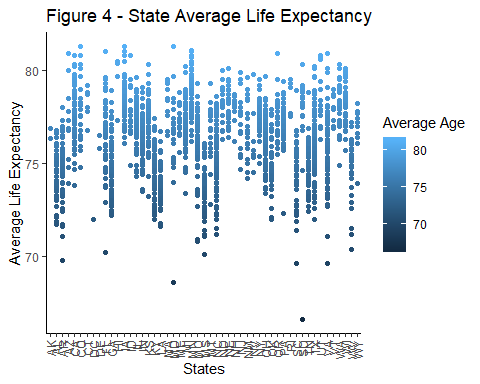
BarPlotFin<- ggplot(data=PLot\_Bar) + geom\_bar(aes(x = Variables,y=value,fill=Variables),stat="identity") + coord\_flip() +scale\_fill\_hue(l=45) + ggtitle('Figure 3 - Population Health Statistics')  
BarPlotFin



##### Plot Average Life Expectancy by State

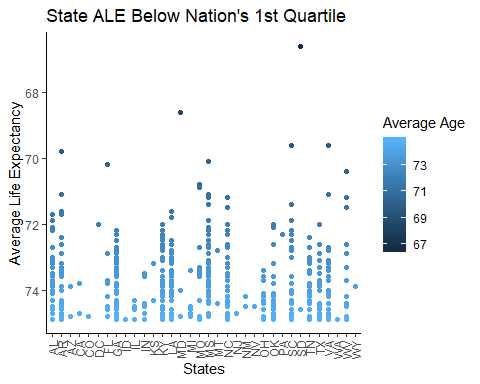
# assistance from <https://www.datanovia.com/en/blog/ggplot-themes-gallery/>

Life\_Exp\_States <- ggplot(data = mean\_clean, aes(x=CHSI\_State\_Abbr, y = ALE, color = ALE, group=CHSI\_State\_Abbr)) +   
 geom\_point() +  
 scale\_color\_continuous(name = "Average Age") +  
 theme\_classic() +  
 xlab(label = "States") +  
 ylab(label = "Average Life Expectancy") +   
 ggtitle(label = "Figure 4 - State Average Life Expectancy") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))  
Life\_Exp\_States



###### plot ALE Under 1st Quartile

Life\_Exp\_75 <- ggplot(data = mean\_clean[mean\_clean$ALE<75,], aes(x=CHSI\_State\_Abbr, y = ALE, color = ALE, group=CHSI\_State\_Abbr)) +   
 geom\_point() +  
 scale\_color\_continuous(name = "Average Age") +  
 theme\_classic() +  
 xlab(label = "States") +  
 ylab(label = "Average Life Expectancy") +   
 scale\_y\_reverse() +  
 ggtitle(label = "State ALE Below Nation's 1st Quartile") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))  
Life\_Exp\_75

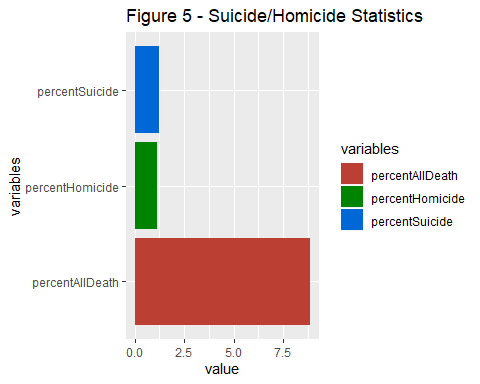


# Suicide

mean\_clean$totalSuicide <- mean\_clean$C\_Bl\_Suicide+mean\_clean$C\_Hi\_Suicide+mean\_clean$C\_Ot\_Suicide+mean\_clean$C\_Wh\_Suicide+mean\_clean$D\_Bl\_Suicide+mean\_clean$D\_Wh\_Suicide+mean\_clean$D\_Hi\_Suicide+mean\_clean$D\_Ot\_Suicide  
mean\_clean$percentSuicide <- (mean\_clean$totalSuicide/mean\_clean$Population\_Size)\*100

# Homicide

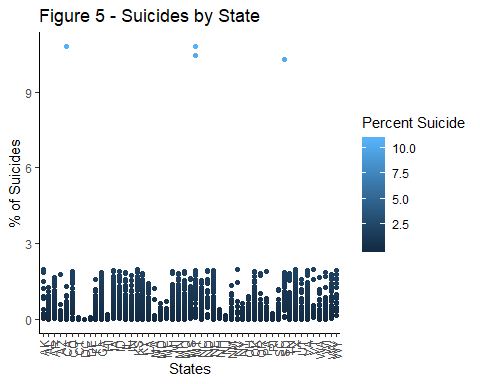
mean\_clean$totalHomicide <- mean\_clean$C\_Bl\_Homicide+mean\_clean$C\_Hi\_Homicide+mean\_clean$C\_Wh\_Homicide+mean\_clean$D\_Bl\_Homicide+mean\_clean$D\_Wh\_Homicide+mean\_clean$D\_Hi\_Homicide  
mean\_clean$percentHomicide <- (mean\_clean$totalHomicide/mean\_clean$Population\_Size)\*100  
mean\_clean$percentAllDeath <- (mean\_clean$All\_Death/mean\_clean$Population\_Size)\*100  
deathDf <- sqldf('select percentSuicide, percentHomicide, percentAllDeath from mean\_clean')  
deathDfMeans <- colMeans(deathDf)  
deathDfMeans <- melt(deathDfMeans)  
deathDfMeans$variables <- rownames(deathDfMeans)  
BarPlotDeath<- ggplot(data=deathDfMeans) + geom\_bar(aes(x = variables,y=value,fill=variables),stat="identity") + coord\_flip() +scale\_fill\_hue(l=45) + ggtitle('Figure 5 - Suicide/Homicide Statistics')  
BarPlotDeath



##### Suicide by State

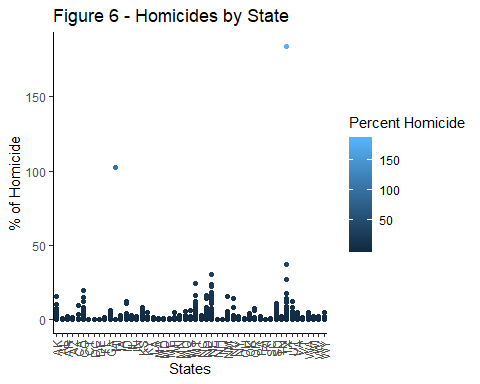
# excluding outliers

mean\_clean\_noOutliers <- mean\_clean[which(mean\_clean$percentSuicide < "100"),]  
Suicide\_by\_States <- ggplot(data = mean\_clean\_noOutliers, aes(x=CHSI\_State\_Abbr, y = percentSuicide, color = percentSuicide, group=CHSI\_State\_Abbr)) +   
 geom\_point() +  
 scale\_color\_continuous(name = "Percent Suicide") +  
 theme\_classic() +  
 xlab(label = "States") +  
 ylab(label = "% of Suicides") +   
 ggtitle(label = "Figure 5 - Suicides by State") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))  
Suicide\_by\_States



##### Homicide by State

Homicide\_by\_States <- ggplot(data = mean\_clean, aes(x=CHSI\_State\_Abbr, y = percentHomicide, color = percentHomicide, group=CHSI\_State\_Abbr)) +   
 geom\_point() +  
 scale\_color\_continuous(name = "Percent Homicide") +  
 theme\_classic() +  
 xlab(label = "States") +  
 ylab(label = "% of Homicide") +   
 ggtitle(label = "Figure 6 - Homicides by State") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))  
Homicide\_by\_States



Plot of Smokers relating to ALE and Exercise ################

# Shows a plot of no exercise to ALE. Less smoking in green, heavy smoking in blue.

# determine the range of smokers

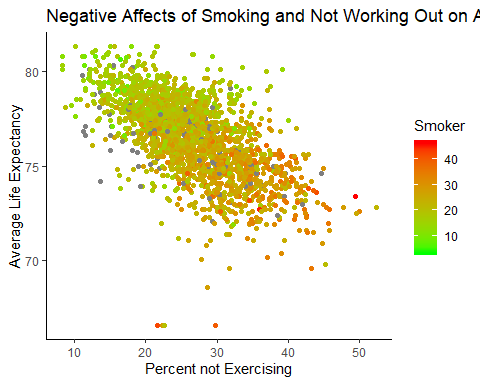
summary(combined$Smoker)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 3.60 19.40 23.00 23.11 26.70 46.20 874

# plot assistance from <https://www.datanovia.com/en/blog/ggplot-themes-gallery/>

gtt <- ggplot(data = combined, aes(x=No\_Exercise, y = ALE, color = Smoker)) +   
 theme\_classic() +   
 geom\_point() +   
 xlab(label = "Percent not Exercising") +  
 #scale\_x\_reverse() + #x\_axis reversed to show positive correlation  
 ylab(label = "Average Life Expectancy") +   
 scale\_colour\_gradient(limits=c(3.6,46.2), low="green", high="red") + #Smoker Range  
 ggtitle(label = "Negative Affects of Smoking and Not Working Out on ALE")  
gtt

## Warning: Removed 935 rows containing missing values (geom\_point).

 # Linear regression A linear model was built using variables like Age, Demographics, Drug Usage, Use of Toxic Chem, No Exercise, Fruit intake, Blood Pressure, Diabetes and an R-squared value of 71.2% was obtained that explained the variability on Average life expectancy in a county.

ndf <-read.csv("C:/Users/Caitlin Casey/Desktop/SYRACUSE/USA\_CHSI\_DATASET/ndf.csv")  
mean\_clean$HIV <- ndf$D\_Hi\_HIV+ndf$D\_Bl\_HIV+ndf$D\_Hi\_HIV  
mean\_clean$E\_Cancer <- ndf$E\_Wh\_Cancer + ndf$E\_Bl\_Cancer + ndf$E\_Ot\_Cancer + ndf$E\_Hi\_Cancer  
mean\_clean$E\_HeartDis <- ndf$E\_Wh\_HeartDis + ndf$E\_Bl\_HeartDis + ndf$E\_Ot\_HeartDis + ndf$E\_Hi\_HeartDis  
mean\_clean$F\_HeartDis <- ndf$F\_Wh\_HeartDis + ndf$F\_Bl\_HeartDis + ndf$F\_Ot\_HeartDis + ndf$F\_Hi\_HeartDis  
mean\_clean$F\_Cancer <- ndf$F\_Wh\_Cancer + ndf$F\_Bl\_Cancer + ndf$F\_Ot\_Cancer + ndf$F\_Hi\_Cancer  
ndf\_lr <- mean\_clean[,c('Poverty','Age\_19\_Under','Age\_19\_64','Age\_65\_84','Age\_85\_and\_Over','White','Black','Native\_American','Asian','Hispanic','Recent\_Drug\_Use','Toxic\_Chem','No\_Exercise','Few\_Fruit\_Veg','Obesity','High\_Blood\_Pres','Smoker','Diabetes','HIV','E\_HeartDis','F\_HeartDis','ALE')]  
linear\_ndf <- lm(ALE~.,ndf\_lr)  
summary(linear\_ndf)$r.squared

## [1] 0.71221

# svm model

Linear and Polynomial kernal gave an error of 0.03% for a cost of 1 and 50. So we decided to go with a cost of 1 with the linear model.

{ r svm results, include=FALSE} str(ndf)

# svm- linear kernel, cost = 1

library(e1071)  
testset1 <- ndf[1:1000,]  
trainset1 <- ndf[1001:3142,]  
svm1 <- svm(ALE ~., data = trainset1, kernel = "linear", type = "eps", cost=1)  
pred=predict(svm1, testset1[,-38], type=c("eps"))  
pred

## 1 2 3 4 5 6 7 8   
## 75.25394 76.15949 73.44320 74.42515 75.86854 71.96952 73.80117 74.06231   
## 9 10 11 12 13 14 15 16   
## 74.22514 75.00704 74.62024 73.80575 72.97592 75.42017 75.47332 75.13808   
## 17 18 19 20 21 22 23 24   
## 75.08845 74.04647 74.53969 74.76101 74.70479 75.06864 74.20613 71.44706   
## 25 26 27 28 29 30 31 32   
## 75.07845 75.54188 73.37800 73.85342 75.28425 75.10304 75.03691 71.52628   
## 33 34 35 36 37 38 39 40   
## 73.17701 74.76699 74.51906 74.28822 74.08083 75.26999 75.35550 74.64550   
## 41 42 43 44 45 46 47 48   
## 75.15333 74.83317 71.92030 71.58173 75.90574 73.66455 75.33283 74.52814   
## 49 50 51 52 53 54 55 56   
## 73.89840 74.34945 73.82901 75.42822 71.96361 73.82449 73.44793 74.38885   
## 57 58 59 60 61 62 63 64   
## 73.45848 74.79051 77.11225 72.37850 73.24546 74.25491 74.35925 73.79336   
## 65 66 67 68 69 70 71 72   
## 73.84774 72.31736 74.82462 80.87995 80.20286 77.14301 76.57673 77.48856   
## 73 74 75 76 77 78 79 80   
## 77.10359 76.46317 76.76959 77.27810 77.20725 76.24916 77.45872 79.00372   
## 81 82 83 84 85 86 87 88   
## 76.55373 76.36120 76.74742 78.43044 69.30535 75.67909 76.94356 76.42080   
## 89 90 91 92 93 94 95 96   
## 75.49852 77.23941 76.35611 77.28731 76.93932 76.50406 75.93154 76.73085   
## 97 98 99 100 101 102 103 104   
## 77.43324 75.00657 76.46335 77.60059 75.69965 77.24454 75.46108 75.61432   
## 105 106 107 108 109 110 111 112   
## 77.39298 76.15272 77.77932 76.88109 76.74324 73.81346 73.70679 76.33607   
## 113 114 115 116 117 118 119 120   
## 77.24563 75.45876 74.15220 75.65465 75.99773 71.59007 75.34228 74.55501   
## 121 122 123 124 125 126 127 128   
## 75.44267 75.03111 74.07578 74.86835 75.14751 75.70620 72.53449 73.60957   
## 129 130 131 132 133 134 135 136   
## 74.47845 73.18677 74.21128 76.52411 75.76995 75.13071 75.37539 75.23974   
## 137 138 139 140 141 142 143 144   
## 74.23657 74.26712 75.20235 74.86671 75.33578 74.93538 73.47701 72.82002   
## 145 146 147 148 149 150 151 152   
## 75.73694 74.94861 74.86285 72.21938 73.65626 74.76292 74.77132 76.29090   
## 153 154 155 156 157 158 159 160   
## 75.45557 75.41020 74.28899 72.25867 73.41533 75.43771 74.71672 75.28849   
## 161 162 163 164 165 166 167 168   
## 73.64680 75.33618 70.98670 76.01741 72.68750 75.19068 75.99316 75.89183   
## 169 170 171 172 173 174 175 176   
## 75.23513 74.81464 71.94889 76.52873 74.51915 74.95745 76.10909 75.13386   
## 177 178 179 180 181 182 183 184   
## 74.60191 74.72438 74.35055 75.40862 76.92484 75.51603 74.14811 76.09408   
## 185 186 187 188 189 190 191 192   
## 80.03884 76.35734 77.56910 77.12203 77.17220 78.17252 79.51297 75.84629   
## 193 194 195 196 197 198 199 200   
## 78.84706 77.29290 77.26939 76.52250 77.73539 77.28816 76.29301 76.93784   
## 201 202 203 204 205 206 207 208   
## 76.48689 75.89824 77.07584 76.79220 80.19171 76.89984 77.42704 76.97144   
## 209 210 211 212 213 214 215 216   
## 75.75736 77.02808 78.32438 79.85840 78.87578 80.06995 79.51741 76.83729   
## 217 218 219 220 221 222 223 224   
## 77.85950 78.18898 79.19985 76.88670 78.75887 82.67512 78.05516 78.72903   
## 225 226 227 228 229 230 231 232   
## 81.69825 78.71461 81.65007 79.11394 77.29936 77.54882 77.04399 79.04736   
## 233 234 235 236 237 238 239 240   
## 79.13061 77.43697 77.66811 76.90695 75.93888 76.46167 77.05900 79.18429   
## 241 242 243 244 245 246 247 248   
## 79.14714 76.37875 77.69865 77.25456 78.41370 76.42732 77.01215 75.33214   
## 249 250 251 252 253 254 255 256   
## 79.32545 78.84508 76.88538 77.49774 77.53141 77.33751 76.47345 75.09292   
## 257 258 259 260 261 262 263 264   
## 75.46613 77.36988 77.50165 77.18907 80.34133 79.82661 78.43777 77.62649   
## 265 266 267 268 269 270 271 272   
## 76.77524 78.56195 77.19416 77.35445 77.59952 76.64577 76.44126 76.49621   
## 273 274 275 276 277 278 279 280   
## 78.58611 78.48240 77.85014 77.73531 78.24843 78.34642 77.43338 76.43412   
## 281 282 283 284 285 286 287 288   
## 76.40230 77.69782 75.69189 77.05665 76.76873 77.39906 77.15927 77.34870   
## 289 290 291 292 293 294 295 296   
## 77.45145 77.41607 78.37543 79.91301 77.67655 77.02913 76.32646 76.93470   
## 297 298 299 300 301 302 303 304   
## 79.05041 76.36592 76.22813 77.38440 77.64095 79.14260 77.83856 76.47650   
## 305 306 307 308 309 310 311 312   
## 77.95078 77.60610 78.89851 78.28906 79.16295 79.21917 78.10714 78.29484   
## 313 314 315 316 317 318 319 320   
## 78.88868 77.77914 75.44447 76.96733 75.81618 75.06774 75.99229 74.14601   
## 321 322 323 324 325 326 327 328   
## 75.66238 74.70562 76.31266 76.99054 75.10574 77.16287 75.66323 77.06492   
## 329 330 331 332 333 334 335 336   
## 77.85641 74.62292 75.01982 74.52555 75.64259 74.18876 77.17808 75.87472   
## 337 338 339 340 341 342 343 344   
## 73.56655 75.57629 75.81864 74.67040 73.84494 75.92525 75.81733 76.18240   
## 345 346 347 348 349 350 351 352   
## 76.22569 76.34822 74.42887 77.99208 74.28273 75.66401 74.79957 76.92852   
## 353 354 355 356 357 358 359 360   
## 76.58587 76.30915 74.96961 75.47811 73.85910 77.26560 75.50147 78.18768   
## 361 362 363 364 365 366 367 368   
## 76.08862 76.69467 76.05340 76.49464 75.28482 76.50822 77.06960 77.62944   
## 369 370 371 372 373 374 375 376   
## 76.99423 77.38509 75.98471 73.79233 77.50553 75.79200 76.35602 78.50980   
## 377 378 379 380 381 382 383 384   
## 77.54774 75.36773 74.62868 74.15276 80.15230 75.95010 76.13929 75.10751   
## 385 386 387 388 389 390 391 392   
## 74.62980 74.69521 75.06782 74.74447 73.34029 74.11514 75.99986 75.78103   
## 393 394 395 396 397 398 399 400   
## 75.48685 74.53354 74.87431 73.33249 75.26958 74.99979 73.99232 76.57530   
## 401 402 403 404 405 406 407 408   
## 74.45213 72.19085 74.48989 72.17699 76.00766 74.88534 75.93602 75.72604   
## 409 410 411 412 413 414 415 416   
## 74.17250 74.43618 75.31146 75.38631 77.88603 75.66659 73.05085 74.20966   
## 417 418 419 420 421 422 423 424   
## 73.64751 77.51684 73.07836 74.08549 78.07631 74.45036 76.48684 75.25311   
## 425 426 427 428 429 430 431 432   
## 73.73405 75.99624 76.42067 74.21539 74.66530 73.56121 73.34494 72.23598   
## 433 434 435 436 437 438 439 440   
## 75.87742 73.00199 75.00942 75.71481 74.06008 73.79473 73.93247 75.49892   
## 441 442 443 444 445 446 447 448   
## 78.09273 75.10470 78.42725 75.34244 75.53316 75.66861 75.74265 75.44667   
## 449 450 451 452 453 454 455 456   
## 76.40275 73.79559 74.14117 78.20761 76.72888 76.60222 71.48582 75.20221   
## 457 458 459 460 461 462 463 464   
## 76.41882 74.78997 75.26263 77.08635 75.13929 74.77281 76.20924 75.21011   
## 465 466 467 468 469 470 471 472   
## 74.44795 73.69332 74.06896 73.71999 75.67687 74.84532 74.42373 73.81030   
## 473 474 475 476 477 478 479 480   
## 77.05835 74.30993 73.93947 74.65106 74.22248 76.28310 74.39506 73.75792   
## 481 482 483 484 485 486 487 488   
## 73.25551 75.40098 74.03149 73.67314 74.96368 72.86690 75.33216 74.56269   
## 489 490 491 492 493 494 495 496   
## 75.34095 75.26615 74.05173 74.93548 77.98322 75.53139 76.49392 73.93754   
## 497 498 499 500 501 502 503 504   
## 76.18664 74.85601 76.01509 75.47041 75.19179 74.33507 73.42467 76.27255   
## 505 506 507 508 509 510 511 512   
## 72.50913 72.77595 75.72939 75.11256 74.30666 73.85617 74.37795 74.88719   
## 513 514 515 516 517 518 519 520   
## 73.44078 73.54911 73.22253 73.86039 73.96922 73.69557 71.65694 73.04066   
## 521 522 523 524 525 526 527 528   
## 74.29088 74.50462 74.03145 76.09774 73.62874 74.70737 73.37595 73.98550   
## 529 530 531 532 533 534 535 536   
## 76.60481 74.59900 75.08766 75.86274 73.73773 73.60665 72.76794 73.76337   
## 537 538 539 540 541 542 543 544   
## 74.34782 73.91662 76.09378 76.27784 73.84252 75.12482 74.24212 73.35404   
## 545 546 547 548 549 550 551 552   
## 78.63017 81.78442 84.07556 81.11394 80.24248 78.34704 76.64957 76.84833   
## 553 554 555 556 557 558 559 560   
## 76.90191 75.78349 77.15339 79.21188 77.04312 76.90129 77.35776 76.99712   
## 561 562 563 564 565 566 567 568   
## 76.07727 77.24653 77.21120 77.19332 77.43578 76.81438 75.86132 76.84917   
## 569 570 571 572 573 574 575 576   
## 76.88873 78.16414 76.71723 77.19721 77.47467 76.62898 77.62264 76.97002   
## 577 578 579 580 581 582 583 584   
## 77.75943 78.43831 76.71792 76.78614 77.18468 78.38479 77.19389 77.39975   
## 585 586 587 588 589 590 591 592   
## 76.77740 75.45433 76.87708 77.02004 75.14577 79.22783 77.60141 76.94816   
## 593 594 595 596 597 598 599 600   
## 76.82256 77.09356 74.24686 76.64654 78.03354 77.05243 78.17517 76.66290   
## 601 602 603 604 605 606 607 608   
## 77.75909 77.32898 77.90794 76.93594 76.83947 77.75700 77.14460 77.19526   
## 609 610 611 612 613 614 615 616   
## 75.99433 76.32106 77.14526 78.16324 77.14813 77.49182 79.68446 77.25347   
## 617 618 619 620 621 622 623 624   
## 77.80054 77.98378 76.25862 77.71339 76.15412 76.53579 76.43369 77.02187   
## 625 626 627 628 629 630 631 632   
## 77.97377 77.27023 77.84322 75.84445 76.24620 78.51582 76.48773 76.76488   
## 633 634 635 636 637 638 639 640   
## 77.49129 75.59765 76.63938 77.73835 75.91660 78.67940 76.37632 79.17765   
## 641 642 643 644 645 646 647 648   
## 76.63075 78.80059 77.15102 76.70818 77.41896 76.63466 77.11818 77.61879   
## 649 650 651 652 653 654 655 656   
## 78.77922 77.93929 75.88309 77.34190 76.46070 76.36389 77.74700 77.14317   
## 657 658 659 660 661 662 663 664   
## 76.47993 77.09421 76.95524 78.59526 77.02841 76.55481 77.25872 77.88397   
## 665 666 667 668 669 670 671 672   
## 75.65708 76.11282 78.29776 77.85532 76.28745 74.61254 77.87172 77.27729   
## 673 674 675 676 677 678 679 680   
## 76.59298 76.83419 74.85886 76.39818 76.93421 77.08630 77.68893 77.42018   
## 681 682 683 684 685 686 687 688   
## 77.84340 77.81698 77.32885 76.70730 75.49746 77.69521 76.92670 77.25852   
## 689 690 691 692 693 694 695 696   
## 77.50711 77.25212 77.38443 77.81250 76.82136 76.17449 78.48736 77.40888   
## 697 698 699 700 701 702 703 704   
## 77.13656 77.45995 77.86856 77.31928 78.14660 77.66415 77.34206 77.32989   
## 705 706 707 708 709 710 711 712   
## 76.46348 76.22315 77.05228 76.28908 76.46785 76.96875 76.51935 77.44948   
## 713 714 715 716 717 718 719 720   
## 76.33019 77.68924 77.31731 76.59822 77.02202 76.49145 76.72867 76.32111   
## 721 722 723 724 725 726 727 728   
## 76.54869 75.52537 76.40741 79.16249 77.66867 76.74349 78.01718 76.68310   
## 729 730 731 732 733 734 735 736   
## 76.43611 77.48544 76.77667 77.51972 76.37602 76.21286 75.89389 76.97284   
## 737 738 739 740 741 742 743 744   
## 75.99744 77.37209 77.18844 74.74700 76.23217 75.90947 76.04967 75.18610   
## 745 746 747 748 749 750 751 752   
## 77.33831 76.95216 76.48826 77.29014 76.24566 76.47468 77.35936 76.95422   
## 753 754 755 756 757 758 759 760   
## 77.49043 75.38208 75.80213 75.47047 76.82917 76.76833 77.51279 77.10195   
## 761 762 763 764 765 766 767 768   
## 77.03133 76.63020 76.24322 77.11201 77.73594 76.58140 75.40218 76.96158   
## 769 770 771 772 773 774 775 776   
## 76.57435 75.62400 77.65599 76.06378 76.49646 77.89383 77.93336 76.93897   
## 777 778 779 780 781 782 783 784   
## 76.27387 76.37996 75.39538 76.90739 76.76519 77.37326 76.00068 76.17938   
## 785 786 787 788 789 790 791 792   
## 77.75735 76.85929 77.23769 78.73976 78.22467 78.27153 76.06747 78.89014   
## 793 794 795 796 797 798 799 800   
## 77.46109 77.38020 77.35810 78.78889 77.02986 78.68794 79.09616 78.59734   
## 801 802 803 804 805 806 807 808   
## 79.03255 78.17134 78.57909 77.82214 78.25621 77.29988 77.46283 78.48754   
## 809 810 811 812 813 814 815 816   
## 78.12160 76.61426 77.79433 78.74551 77.39310 76.11191 78.27209 77.67573   
## 817 818 819 820 821 822 823 824   
## 78.79577 78.33006 78.21503 77.25137 78.08318 78.46426 77.08073 78.07720   
## 825 826 827 828 829 830 831 832   
## 79.39711 78.63379 77.97301 78.78020 78.19897 76.69691 77.86691 78.97450   
## 833 834 835 836 837 838 839 840   
## 78.40375 78.22738 78.11400 78.12972 77.80055 77.69620 79.11712 78.32147   
## 841 842 843 844 845 846 847 848   
## 78.49032 78.58807 76.59870 78.20576 77.52736 77.22579 78.70611 78.37528   
## 849 850 851 852 853 854 855 856   
## 77.58024 77.66072 77.10864 77.04337 78.89675 78.20805 77.41669 77.32192   
## 857 858 859 860 861 862 863 864   
## 77.59506 78.84159 77.55399 76.89339 78.72013 78.51422 78.70717 77.94246   
## 865 866 867 868 869 870 871 872   
## 76.61179 77.84527 77.50360 78.34711 77.25011 78.72890 79.13462 79.07511   
## 873 874 875 876 877 878 879 880   
## 78.52602 78.07483 77.35843 77.81871 77.39709 77.99827 78.32025 77.70968   
## 881 882 883 884 885 886 887 888   
## 76.52563 79.27421 78.91312 77.10151 78.34784 78.57851 75.83371 77.09493   
## 889 890 891 892 893 894 895 896   
## 76.84033 77.65327 76.98191 76.64429 77.12456 77.44969 77.76967 77.35430   
## 897 898 899 900 901 902 903 904   
## 75.58236 79.25829 78.34567 78.28825 77.97589 78.49338 78.28092 76.20858   
## 905 906 907 908 909 910 911 912   
## 76.94646 78.15527 77.14941 76.84439 77.78127 78.37050 77.00689 77.72574   
## 913 914 915 916 917 918 919 920   
## 78.34573 77.00576 78.08187 76.63484 76.77871 77.87391 76.79943 77.80978   
## 921 922 923 924 925 926 927 928   
## 78.39201 77.29813 76.56119 78.70090 77.38973 78.34672 76.98367 77.90141   
## 929 930 931 932 933 934 935 936   
## 76.69268 77.54197 77.93030 79.28812 76.72822 77.73289 78.09918 75.94313   
## 937 938 939 940 941 942 943 944   
## 78.11843 77.06603 76.21337 76.32572 77.39622 77.66245 78.21641 78.77768   
## 945 946 947 948 949 950 951 952   
## 77.48618 78.29830 77.14032 78.66264 76.27143 78.50806 76.97110 78.60897   
## 953 954 955 956 957 958 959 960   
## 76.63890 78.10398 78.95210 76.94832 78.39279 76.96303 78.00487 77.67243   
## 961 962 963 964 965 966 967 968   
## 78.07765 77.81042 78.08656 77.21423 77.92696 77.51862 78.29015 78.01527   
## 969 970 971 972 973 974 975 976   
## 77.64104 77.65193 77.26628 78.43921 76.90703 78.11905 76.85550 76.68413   
## 977 978 979 980 981 982 983 984   
## 76.85129 79.45813 78.25619 77.42012 78.45489 77.39790 77.47045 78.93316   
## 985 986 987 988 989 990 991 992   
## 77.50355 76.76922 78.51806 78.33645 76.72666 77.35616 74.81914 73.88734   
## 993 994 995 996 997 998 999 1000   
## 74.27389 77.10760 74.60191 75.24087 74.26893 72.77296 77.72536 75.90311

library(caret)

## Warning: package 'caret' was built under R version 3.6.1

## Loading required package: lattice

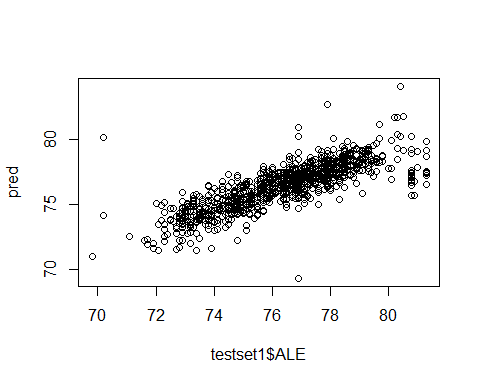
##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

conf <- data.frame(pred,testset1$ALE)  
sqrt((mean(conf$pred-conf$testset1.ALE)^2))

## [1] 0.03546129

plot(pred~testset1$ALE)



# svm-linear, cost = 20

testset1 <- ndf[1:1000,]  
trainset1 <- ndf[1001:3142,]  
svm1 <- svm(ALE ~., data = trainset1, kernel = "linear", type = "eps", cost = 20)  
pred=predict(svm1, testset1, type=c("eps"))  
library(caret)  
conf.linear <- data.frame(pred,testset1$ALE)  
sqrt((mean(conf.linear$pred-conf.linear$testset1.ALE)^2))

## [1] 0.03480293

# svm-polynomial, cost = 50-0.03

testset1 <- ndf[1:1000,]  
trainset1 <- ndf[1001:3142,]  
svm1 <- svm(ALE ~., data = trainset1, kernel = "polynomial", type = "eps", cost = 50)  
pred=predict(svm1, testset1[,-38], type=c("eps"))  
library(caret)  
conf.polynomial <- data.frame(pred,testset1$ALE)  
sqrt((mean(conf.linear$pred-conf.linear$testset1.ALE)^2))

## [1] 0.03480293

# svm-polynomial, cost = 50- 0.039

testset1 <- ndf[1:1000,]  
trainset1 <- ndf[1001:3142,]  
svm1 <- svm(ALE ~., data = trainset1, kernel = "radial", type = "eps", cost = 50)  
pred=predict(svm1, testset1[,-38], type=c("eps"))  
library(caret)  
conf.radial <- data.frame(pred,testset1$ALE)  
sqrt((mean(pred-testset1$ALE)^2))

## [1] 0.03980379

# Decision Tree

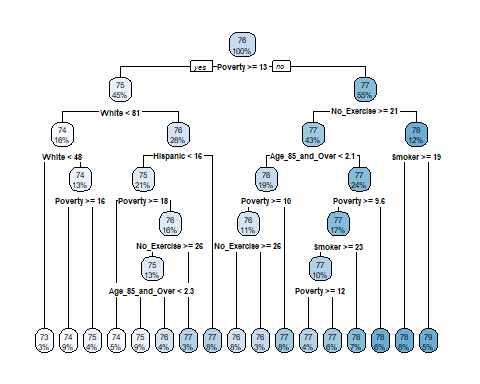
library(rpart)

## Warning: package 'rpart' was built under R version 3.6.1

library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 3.6.1

set.seed(50)  
dt <- rpart(ALE~.,trainset1,method = "anova", control = rpart.control(cp = 0, maxdepth = 7,minsplit = 200))  
rpart.plot(dt)



pred\_dt <- predict(dt, testset1[,-38])  
library(caret)  
conf.dt <- data.frame(pred,testset1$ALE)  
sqrt((mean(pred\_dt-testset1$ALE)^2))

## [1] 0.08064625

# Random Forest

library(randomForest)

## Warning: package 'randomForest' was built under R version 3.6.1

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

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

## The following object is masked from 'package:ggplot2':  
##   
## margin

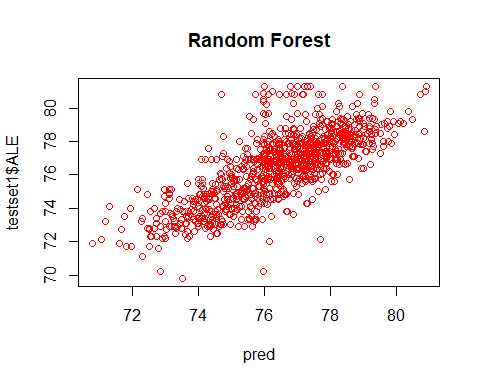
library(caret)  
set.seed(50)  
rfm\_chsi <- randomForest(ALE~., data=trainset1, ntree=11, na.action=na.exclude)  
print(rfm\_chsi)

##   
## Call:  
## randomForest(formula = ALE ~ ., data = trainset1, ntree = 11, na.action = na.exclude)   
## Type of random forest: regression  
## Number of trees: 11  
## No. of variables tried at each split: 12  
##   
## Mean of squared residuals: 1.393254  
## % Var explained: 63.78

predRF <- predict(rfm\_chsi, testset1[,-38])  
conf.RF <- data.frame(pred,testset1$ALE)  
sqrt((mean(predRF-testset1$ALE)^2))

## [1] 0.08227802

plot(pred,testset1$ALE, col='red', main="Random Forest")



# Naive Bayes