data\_exploration

Group 21

2022-10-07

library(rstudioapi)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.8 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.1  
## ✔ readr 2.1.2 ✔ forcats 0.5.2  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(gam)

## Loading required package: splines  
## Loading required package: foreach  
##   
## Attaching package: 'foreach'  
##   
## The following objects are masked from 'package:purrr':  
##   
## accumulate, when  
##   
## Loaded gam 1.20.2

library(splines)  
library(splines2)   
library(dplyr)  
library(tidyr)  
library(broom)  
library(dslabs)  
library(ggplot2)  
library(ggthemes)  
library(ggrepel)  
# data cleaning for socio economic factors  
soecon <- read\_csv('soecon.csv')

## New names:  
## Rows: 459 Columns: 11  
## ── Column specification  
## ──────────────────────────────────────────────────────── Delimiter: "," chr  
## (1): Location dbl (10): ...1, year, SNAP, individual\_homeless, fam\_homeless,  
## total\_homeles...  
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ  
## Specify the column types or set `show\_col\_types = FALSE` to quiet this message.  
## • `` -> `...1`

fulldat <- read\_csv('FullData.csv')

## New names:  
## Rows: 468 Columns: 42  
## ── Column specification  
## ──────────────────────────────────────────────────────── Delimiter: "," chr  
## (3): LocationAbbr, LocationDesc, State.Name dbl (39): ...1, YearStart,  
## DataValue, LowConfidenceLimit, HighConfidenceLimi...  
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ  
## Specify the column types or set `show\_col\_types = FALSE` to quiet this message.  
## • `` -> `...1`

names(fulldat)[names(fulldat) == 'DataValue'] <- 'Asthmap'  
names(fulldat)[names(fulldat) == 'TOBDataValue'] <- 'TOB'

#### EDA of socioeconomic data

names(soecon)

## [1] "...1" "Location" "year"   
## [4] "SNAP" "individual\_homeless" "fam\_homeless"   
## [7] "total\_homeless" "unemployment\_rate" "Male\_poverty"   
## [10] "Female\_poverty" "Total\_poverty"

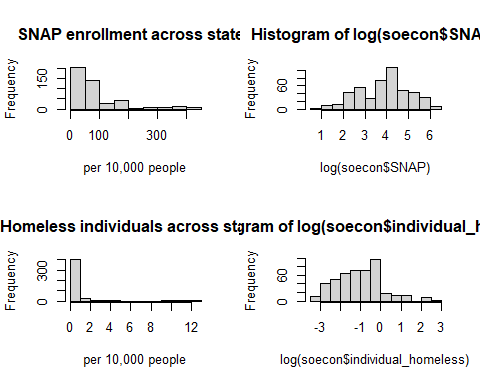
Run a summary of the variables in socioeconomic category

summary(soecon)

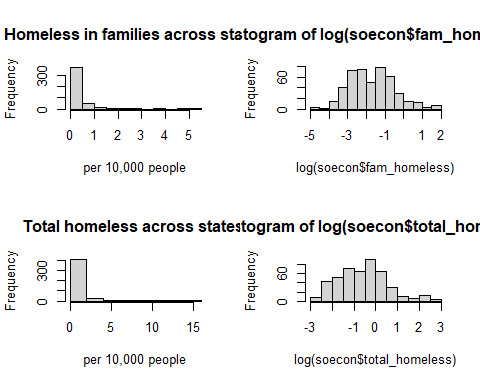
## ...1 Location year SNAP   
## Min. : 1.0 Length:459 Min. :2011 Min. : 2.431   
## 1st Qu.:115.5 Class :character 1st Qu.:2013 1st Qu.: 18.913   
## Median :230.0 Mode :character Median :2015 Median : 60.619   
## Mean :230.0 Mean :2015 Mean : 85.846   
## 3rd Qu.:344.5 3rd Qu.:2017 3rd Qu.: 93.756   
## Max. :459.0 Max. :2019 Max. :441.777   
## NA's :1   
## individual\_homeless fam\_homeless total\_homeless unemployment\_rate  
## Min. : 0.0320 Min. :0.00750 Min. : 0.0542 Min. : 2.100   
## 1st Qu.: 0.1500 1st Qu.:0.07225 1st Qu.: 0.2376 1st Qu.: 3.900   
## Median : 0.3505 Median :0.16880 Median : 0.5785 Median : 5.000   
## Mean : 0.7216 Mean :0.39994 Mean : 1.1215 Mean : 5.471   
## 3rd Qu.: 0.6794 3rd Qu.:0.39065 3rd Qu.: 1.0424 3rd Qu.: 6.700   
## Max. :12.8777 Max. :5.21150 Max. :15.1278 Max. :15.800   
##   
## Male\_poverty Female\_poverty Total\_poverty   
## Min. :0.0570 Min. :0.0780 Min. :0.0700   
## 1st Qu.:0.0980 1st Qu.:0.1250 1st Qu.:0.1120   
## Median :0.1140 Median :0.1460 Median :0.1290   
## Mean :0.1167 Mean :0.1508 Mean :0.1341   
## 3rd Qu.:0.1360 3rd Qu.:0.1740 3rd Qu.:0.1545   
## Max. :0.2020 Max. :0.2610 Max. :0.2230   
##

Draw histograms of the variables in the dataset

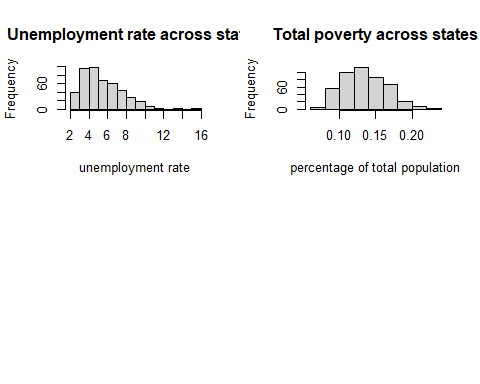
par(mfrow = c(2,2))  
hist(soecon$SNAP, main='SNAP enrollment across states', xlab = 'per 10,000 people' )  
hist(log(soecon$SNAP))  
  
hist(soecon$individual\_homeless, main = 'Homeless individuals across states', xlab = 'per 10,000 people')  
hist(log(soecon$individual\_homeless))



hist(soecon$fam\_homeless, main = 'Homeless in families across states', xlab = 'per 10,000 people')  
hist(log(soecon$fam\_homeless))  
  
hist(soecon$total\_homeless, main = 'Total homeless across states ', xlab = 'per 10,000 people')  
hist(log(soecon$total\_homeless))

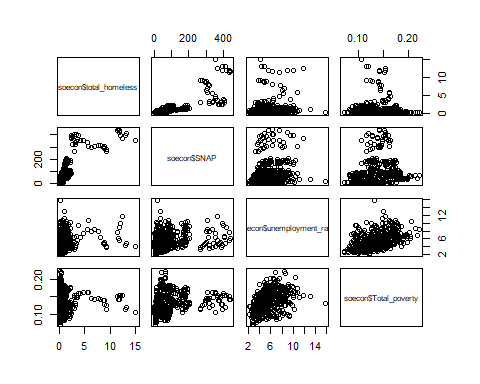


hist(soecon$unemployment\_rate, main='Unemployment rate across states', xlab = 'unemployment rate')  
  
  
hist(soecon$Total\_poverty, main='Total poverty across states', xlab = 'percentage of total population')



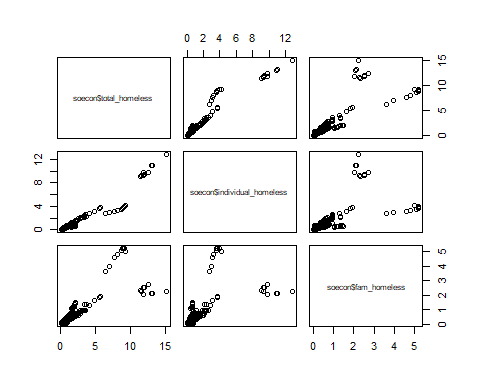
based on the histograms, most of the variables have skewed distribution, but after taking the logarithms of them, they tend to have more normal distribution. This is because most of the variables are population data, and the population values are big, but after taking the logarithms, the values will be more normally distributed.

pairs(soecon$total\_homeless ~ soecon$SNAP + soecon$unemployment\_rate + soecon$Total\_poverty )



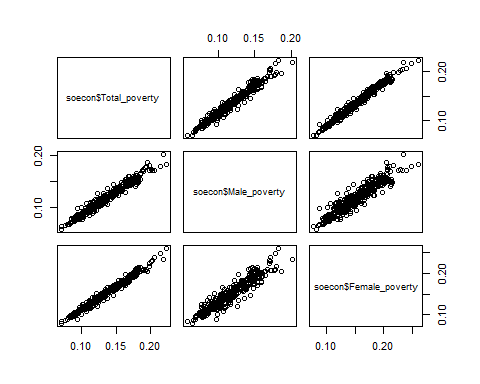
Based on the paired scatter plot, there is no obvious colinearity between the four main socioeconomic variables: SNAP enrollment, total\_poverty rate, unemployment rate and total number of homeless people, which means that these four variables are independent with each other.

# for the homeless variable, test for colinearity between its three subvariables   
pairs(soecon$total\_homeless ~ soecon$individual\_homeless + soecon$fam\_homeless)



From the paired scatter plot, we can see that the total number of homeless people is correlated with number os homeless individuals and the number of homeless people in families, so these three variables are not independent with each other.

# for the homeless variable, test for colinearity between its three subvariables   
pairs(soecon$Total\_poverty ~ soecon$Male\_poverty + soecon$Female\_poverty)



By plotting the percentage of males and females that are in poverty in each state against each other, we can see that these variables are linearly correlated with each other and are not independent.

#### EDA for CDI data

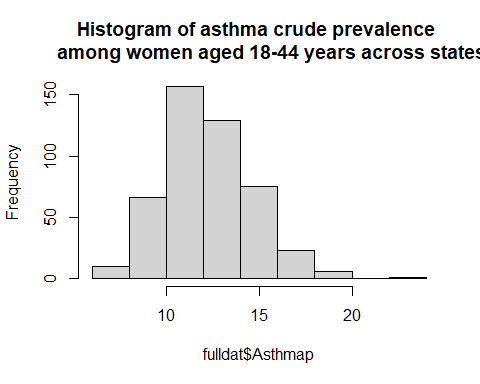
names(fulldat[, 1:25])

## [1] "...1" "YearStart"   
## [3] "LocationAbbr" "LocationDesc"   
## [5] "Asthmap" "LowConfidenceLimit"   
## [7] "HighConfidenceLimit" "LocationID"   
## [9] "TOB" "TOBLowConfidenceLimit"   
## [11] "TOBHighConfidenceLimit" "AlcHeavyDataValue"   
## [13] "AlcHeavyLowConfidenceLimit" "AlcHeavyHighConfidenceLimit"  
## [15] "AlcBingeDataValue" "AlcBingeLowConfidenceLimit"   
## [17] "AlcBingeHighConfidenceLimit" "ObeDataValue"   
## [19] "ObeLowConfidenceLimit" "ObeHighConfidenceLimit"   
## [21] "selfHlthDataValue" "selfHlthLowConfidenceLimit"   
## [23] "selfHlthHighConfidenceLimit" "HlthCareDataValue"   
## [25] "HlthCareLowConfidenceLimit"

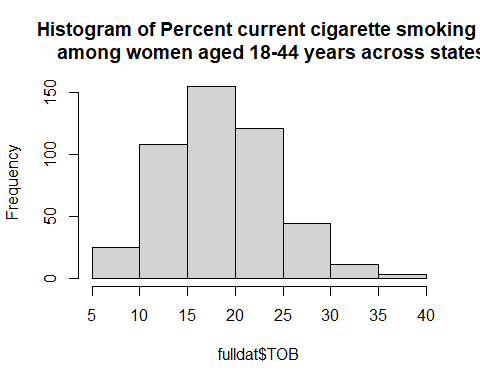
summary(fulldat[, 4:25])

## LocationDesc Asthmap LowConfidenceLimit HighConfidenceLimit  
## Length:468 Min. : 6.10 Min. : 4.200 Min. : 8.80   
## Class :character 1st Qu.:10.70 1st Qu.: 8.500 1st Qu.:13.15   
## Mode :character Median :12.10 Median : 9.800 Median :14.90   
## Mean :12.31 Mean : 9.972 Mean :15.15   
## 3rd Qu.:13.80 3rd Qu.:11.350 3rd Qu.:17.00   
## Max. :22.20 Max. :18.000 Max. :27.00   
## NA's :1 NA's :1 NA's :1   
## LocationID TOB TOBLowConfidenceLimit TOBHighConfidenceLimit  
## Min. : 1.00 Min. : 5.80 Min. : 4.80 Min. : 7.00   
## 1st Qu.:16.75 1st Qu.:14.60 1st Qu.:11.95 1st Qu.:17.50   
## Median :29.50 Median :18.50 Median :15.50 Median :21.80   
## Mean :29.79 Mean :18.61 Mean :15.75 Mean :21.88   
## 3rd Qu.:42.50 3rd Qu.:22.40 3rd Qu.:19.30 3rd Qu.:25.95   
## Max. :72.00 Max. :38.30 Max. :34.50 Max. :42.20   
## NA's :1 NA's :1 NA's :1   
## AlcHeavyDataValue AlcHeavyLowConfidenceLimit AlcHeavyHighConfidenceLimit  
## Min. : 2.500 Min. : 1.500 Min. : 3.300   
## 1st Qu.: 5.200 1st Qu.: 3.700 1st Qu.: 7.200   
## Median : 6.300 Median : 4.700 Median : 8.300   
## Mean : 6.383 Mean : 4.722 Mean : 8.619   
## 3rd Qu.: 7.300 3rd Qu.: 5.500 3rd Qu.: 9.700   
## Max. :14.300 Max. :11.400 Max. :19.500   
## NA's :3 NA's :3 NA's :3   
## AlcBingeDataValue AlcBingeLowConfidenceLimit AlcBingeHighConfidenceLimit  
## Min. : 8.40 Min. : 5.40 Min. :10.70   
## 1st Qu.:15.20 1st Qu.:12.55 1st Qu.:18.20   
## Median :17.90 Median :15.20 Median :21.30   
## Mean :18.12 Mean :15.26 Mean :21.41   
## 3rd Qu.:20.80 3rd Qu.:17.70 3rd Qu.:24.30   
## Max. :32.80 Max. :27.30 Max. :39.70   
## NA's :1 NA's :1 NA's :1   
## ObeDataValue ObeLowConfidenceLimit ObeHighConfidenceLimit selfHlthDataValue  
## Min. :39.40 Min. :34.60 Min. :43.00 Min. :78.50   
## 1st Qu.:49.50 1st Qu.:45.20 1st Qu.:53.25 1st Qu.:85.85   
## Median :53.40 Median :49.20 Median :57.60 Median :87.80   
## Mean :53.41 Mean :49.28 Mean :57.48 Mean :87.57   
## 3rd Qu.:57.70 3rd Qu.:53.60 3rd Qu.:61.55 3rd Qu.:89.55   
## Max. :69.40 Max. :64.70 Max. :74.10 Max. :95.30   
## NA's :1 NA's :1 NA's :1 NA's :1   
## selfHlthLowConfidenceLimit selfHlthHighConfidenceLimit HlthCareDataValue  
## Min. :72.70 Min. :81.5 Min. :59.60   
## 1st Qu.:82.90 1st Qu.:88.3 1st Qu.:79.30   
## Median :84.90 Median :90.0 Median :84.50   
## Mean :84.73 Mean :89.9 Mean :83.61   
## 3rd Qu.:87.20 3rd Qu.:91.6 3rd Qu.:88.90   
## Max. :92.90 Max. :97.0 Max. :96.50   
## NA's :1 NA's :1 NA's :1   
## HlthCareLowConfidenceLimit  
## Min. :56.40   
## 1st Qu.:75.75   
## Median :81.50   
## Mean :80.47   
## 3rd Qu.:85.90   
## Max. :94.80   
## NA's :1

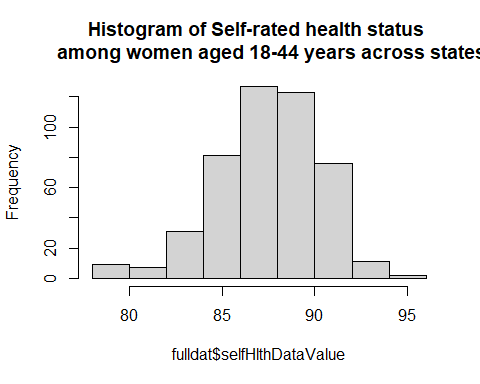
hist(fulldat$Asthmap, main='Histogram of asthma crude prevalence   
 among women aged 18-44 years across states' )



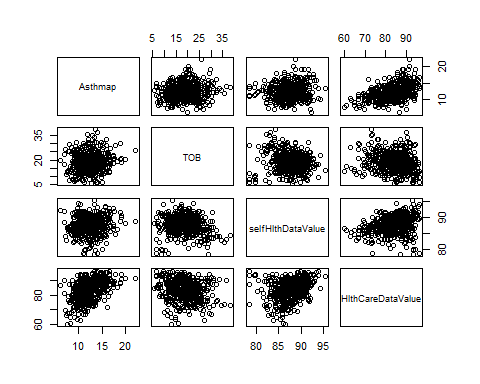
hist(fulldat$TOB, main='Histogram of Percent current cigarette smoking   
 among women aged 18-44 years across states')



hist(fulldat$selfHlthDataValue, main = 'Histogram of Self-rated health status   
 among women aged 18-44 years across states')

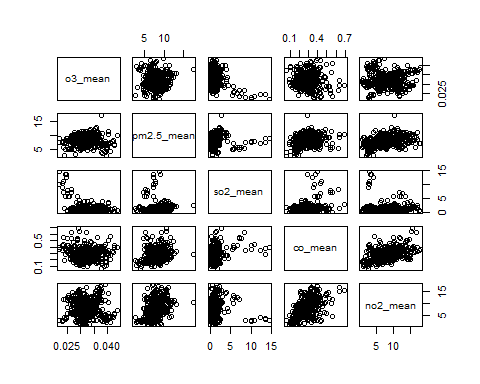
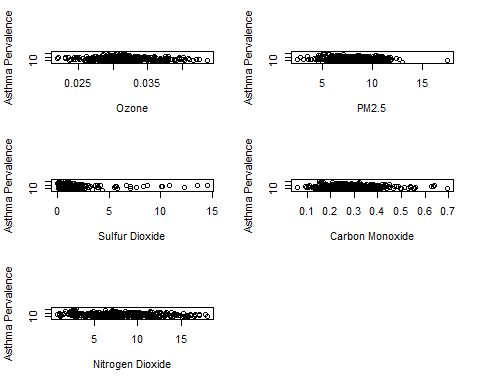
 Based on these histograms, we can see that most of the covariates here are normally distributed with some skewness since they’re data from the realworld.

# paired scatter plots of Asthma Prevalence and CDI variables   
pairs(Asthmap ~ TOB + selfHlthDataValue + HlthCareDataValue, dat = fulldat)

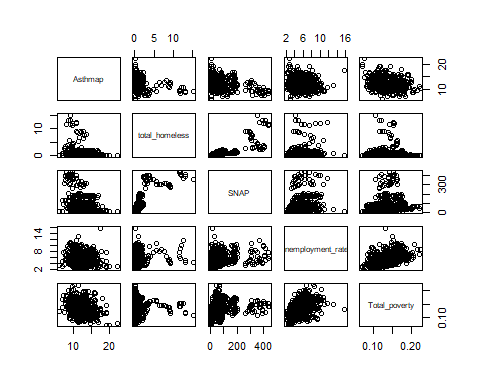
 Based on these paired comparisons, we can see that the asthma prevalence has a relatively strong correlation with health care coverage data compared to the other two covariates. It will likely be a confounder that needs to be addressed.

#### EDA for Air Pollution Data

#air pollutant exploratory  
  
par(mfrow = c(3,2))  
plot(fulldat$Asthmap ~ fulldat$o3\_mean, ylab = "Asthma Pervalence", xlab = "Ozone")  
plot(fulldat$Asthmap ~ fulldat$pm2.5\_mean,ylab = "Asthma Pervalence", xlab = "PM2.5")  
plot(fulldat$Asthmap ~ fulldat$so2\_mean,ylab = "Asthma Pervalence", xlab = "Sulfur Dioxide")  
plot(fulldat$Asthmap ~ fulldat$co\_mean, ylab = "Asthma Pervalence", xlab = "Carbon Monoxide")  
plot(fulldat$Asthmap ~ fulldat$no2\_mean, ylab = "Asthma Pervalence", xlab = "Nitrogen Dioxide")  
  
pairs(o3\_mean ~ pm2.5\_mean + so2\_mean + co\_mean + no2\_mean, data = fulldat)



# paired scatter plots of Asthma Pervalence and socioeconomic variables   
pairs(Asthmap ~ total\_homeless + SNAP + unemployment\_rate + Total\_poverty, dat = fulldat)

 from the paired scatterplot, the pervalence of Asthma is independent of the four main variables in the socioeconomic category.

#### Forward Selection

# now we try to determine which variable is more significant in predicting the Asthma prevalence  
# forward selection  
coln <- colnames(fulldat)  
  
coln

## [1] "...1" "YearStart"   
## [3] "LocationAbbr" "LocationDesc"   
## [5] "Asthmap" "LowConfidenceLimit"   
## [7] "HighConfidenceLimit" "LocationID"   
## [9] "TOB" "TOBLowConfidenceLimit"   
## [11] "TOBHighConfidenceLimit" "AlcHeavyDataValue"   
## [13] "AlcHeavyLowConfidenceLimit" "AlcHeavyHighConfidenceLimit"  
## [15] "AlcBingeDataValue" "AlcBingeLowConfidenceLimit"   
## [17] "AlcBingeHighConfidenceLimit" "ObeDataValue"   
## [19] "ObeLowConfidenceLimit" "ObeHighConfidenceLimit"   
## [21] "selfHlthDataValue" "selfHlthLowConfidenceLimit"   
## [23] "selfHlthHighConfidenceLimit" "HlthCareDataValue"   
## [25] "HlthCareLowConfidenceLimit" "HlthCareHighConfidenceLimit"  
## [27] "X.x" "SNAP"   
## [29] "individual\_homeless" "fam\_homeless"   
## [31] "total\_homeless" "unemployment\_rate"   
## [33] "Male\_poverty" "Female\_poverty"   
## [35] "Total\_poverty" "X.y"   
## [37] "o3\_mean" "pm2.5\_mean"   
## [39] "so2\_mean" "co\_mean"   
## [41] "no2\_mean" "State.Name"

fulldat <- fulldat |> drop\_na()  
  
require(broom)  
#forward selection procedure using AIC values   
lm1 <- lm(Asthmap ~ 1, data=fulldat)  
stepModel <- step(lm1, direction="forward",  
scope=(~ selfHlthDataValue + HlthCareDataValue + SNAP+ individual\_homeless+fam\_homeless+total\_homeless+unemployment\_rate+Male\_poverty+Female\_poverty+Total\_poverty+o3\_mean+pm2.5\_mean+so2\_mean+co\_mean+no2\_mean), data=fulldat)

## Start: AIC=730.7  
## Asthmap ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + HlthCareDataValue 1 517.58 1843.0 627.51  
## + SNAP 1 234.38 2126.2 688.26  
## + Female\_poverty 1 168.16 2192.5 701.30  
## + Total\_poverty 1 156.58 2204.1 703.54  
## + individual\_homeless 1 141.75 2218.9 706.39  
## + Male\_poverty 1 132.16 2228.5 708.22  
## + total\_homeless 1 109.11 2251.5 712.59  
## + pm2.5\_mean 1 48.19 2312.4 723.94  
## + selfHlthDataValue 1 44.27 2316.4 724.66  
## + fam\_homeless 1 25.34 2335.3 728.12  
## + unemployment\_rate 1 25.17 2335.5 728.15  
## + o3\_mean 1 22.64 2338.0 728.61  
## + so2\_mean 1 18.38 2342.3 729.38  
## + co\_mean 1 17.73 2342.9 729.50  
## + no2\_mean 1 12.69 2347.9 730.41  
## <none> 2360.6 730.70  
##   
## Step: AIC=627.51  
## Asthmap ~ HlthCareDataValue  
##   
## Df Sum of Sq RSS AIC  
## + individual\_homeless 1 89.073 1754.0 608.46  
## + total\_homeless 1 74.204 1768.8 612.05  
## + SNAP 1 73.805 1769.2 612.14  
## + unemployment\_rate 1 27.231 1815.8 623.19  
## + fam\_homeless 1 24.571 1818.5 623.81  
## + so2\_mean 1 9.664 1833.4 627.28  
## <none> 1843.0 627.51  
## + no2\_mean 1 8.485 1834.6 627.55  
## + Female\_poverty 1 7.001 1836.0 627.89  
## + Total\_poverty 1 5.343 1837.7 628.28  
## + selfHlthDataValue 1 4.820 1838.2 628.40  
## + Male\_poverty 1 3.225 1839.8 628.77  
## + co\_mean 1 1.851 1841.2 629.09  
## + pm2.5\_mean 1 0.235 1842.8 629.46  
## + o3\_mean 1 0.067 1843.0 629.50  
##   
## Step: AIC=608.46  
## Asthmap ~ HlthCareDataValue + individual\_homeless  
##   
## Df Sum of Sq RSS AIC  
## + unemployment\_rate 1 41.986 1712.0 600.16  
## <none> 1754.0 608.46  
## + selfHlthDataValue 1 7.988 1746.0 608.52  
## + Female\_poverty 1 7.629 1746.3 608.61  
## + so2\_mean 1 5.785 1748.2 609.06  
## + Total\_poverty 1 5.276 1748.7 609.18  
## + SNAP 1 4.842 1749.1 609.28  
## + pm2.5\_mean 1 3.445 1750.5 609.62  
## + fam\_homeless 1 3.108 1750.9 609.71  
## + total\_homeless 1 3.108 1750.9 609.71  
## + Male\_poverty 1 2.576 1751.4 609.83  
## + no2\_mean 1 0.738 1753.2 610.28  
## + co\_mean 1 0.552 1753.4 610.33  
## + o3\_mean 1 0.031 1754.0 610.45  
##   
## Step: AIC=600.16  
## Asthmap ~ HlthCareDataValue + individual\_homeless + unemployment\_rate  
##   
## Df Sum of Sq RSS AIC  
## + Female\_poverty 1 23.6828 1688.3 596.24  
## + Total\_poverty 1 21.3343 1690.7 596.83  
## + Male\_poverty 1 16.2212 1695.8 598.12  
## + selfHlthDataValue 1 13.7980 1698.2 598.72  
## + no2\_mean 1 8.0745 1703.9 600.15  
## <none> 1712.0 600.16  
## + co\_mean 1 6.0010 1706.0 600.67  
## + SNAP 1 5.6554 1706.3 600.76  
## + so2\_mean 1 3.3985 1708.6 601.32  
## + fam\_homeless 1 1.0034 1711.0 601.91  
## + total\_homeless 1 1.0034 1711.0 601.91  
## + pm2.5\_mean 1 0.2104 1711.8 602.11  
## + o3\_mean 1 0.0011 1712.0 602.16  
##   
## Step: AIC=596.24  
## Asthmap ~ HlthCareDataValue + individual\_homeless + unemployment\_rate +   
## Female\_poverty  
##   
## Df Sum of Sq RSS AIC  
## + selfHlthDataValue 1 40.696 1647.6 587.87  
## <none> 1688.3 596.24  
## + no2\_mean 1 7.661 1680.7 596.31  
## + co\_mean 1 6.069 1682.2 596.71  
## + SNAP 1 3.006 1685.3 597.48  
## + so2\_mean 1 1.563 1686.8 597.85  
## + pm2.5\_mean 1 1.072 1687.2 597.97  
## + Male\_poverty 1 0.909 1687.4 598.01  
## + Total\_poverty 1 0.834 1687.5 598.03  
## + fam\_homeless 1 0.705 1687.6 598.06  
## + total\_homeless 1 0.705 1687.6 598.06  
## + o3\_mean 1 0.344 1688.0 598.16  
##   
## Step: AIC=587.87  
## Asthmap ~ HlthCareDataValue + individual\_homeless + unemployment\_rate +   
## Female\_poverty + selfHlthDataValue  
##   
## Df Sum of Sq RSS AIC  
## + no2\_mean 1 23.7578 1623.8 583.70  
## + co\_mean 1 10.8609 1636.8 587.06  
## <none> 1647.6 587.87  
## + SNAP 1 5.9157 1641.7 588.34  
## + so2\_mean 1 1.9988 1645.6 589.36  
## + Male\_poverty 1 1.9390 1645.7 589.37  
## + Total\_poverty 1 1.6508 1646.0 589.45  
## + fam\_homeless 1 1.5058 1646.1 589.48  
## + total\_homeless 1 1.5058 1646.1 589.48  
## + pm2.5\_mean 1 0.6644 1647.0 589.70  
## + o3\_mean 1 0.3250 1647.3 589.79  
##   
## Step: AIC=583.7  
## Asthmap ~ HlthCareDataValue + individual\_homeless + unemployment\_rate +   
## Female\_poverty + selfHlthDataValue + no2\_mean  
##   
## Df Sum of Sq RSS AIC  
## <none> 1623.8 583.70  
## + pm2.5\_mean 1 6.9724 1616.9 583.87  
## + SNAP 1 3.9647 1619.9 584.66  
## + Male\_poverty 1 2.1987 1621.7 585.12  
## + Total\_poverty 1 2.1634 1621.7 585.13  
## + so2\_mean 1 1.4531 1622.4 585.32  
## + co\_mean 1 1.0612 1622.8 585.42  
## + fam\_homeless 1 0.9777 1622.9 585.44  
## + total\_homeless 1 0.9777 1622.9 585.44  
## + o3\_mean 1 0.1881 1623.7 585.65

From the results of forward selection, we can roughly have an idea on which covariates may have a stronger effect as a confounder. We can see that the variables that we may need to include in our final model are: health care coverage data, homeless percentage, unemployment rate, poverty rate, self-rated health status and level.