Household survey and health

Mohana Priya N

2023-02-18

## Data set Description:

The data set is taken from data.gov.in and it is about the survey and health condition of the household from various states. The average value of a state is given.

**Attributes explanation:**

* Indicators -- Improved sanitation includes flush or pour flush toilet (which is flushed to piped sewer system/septic tank/ pit latrine), pit latrine, twin pit, composting toilet.
* MeanHouseholdSize -- average people in household
* Households\_with\_improved\_sanitation -- Improved sanitation includes flush or pour flush toilet (which is flushed to piped sewer system/septic tank/ pit latrine), pit latrine, twin pit, composting toilet.
* Households\_practicing\_open\_defecation -- Open defecation refers to the practice whereby people go out in fields, bushes, forests, open bodies of water, or other open spaces rather than using the toilet to defecate. This information was collected from eligible household respondent
* Households\_exposed\_to\_indoor\_pollution -- Indoor pollution refers to households using solid and unclean fuel (for cooking or any other purpose) inside the household and cooking in traditional chullah/stove or open fire without any ventilation. Unclean fuel refers to use of kerosene, charcoal, coal, crop residue, wood/shrub and dung cake.
* Currently\_smoking -- Currently smoking tobacco refers to smoking any tobacco product such as cigarette, bidi, cigar, hookah and cheroot etc.
* Currently\_consuming\_tobacco -- Currently consuming tobacco refers to consumption of any smokeless tobacco product such as chewing tobacco, gutka or pan masala etc
* Prevalence\_of\_heavy\_episodic\_drinking -- Heavy episodic drinking refers to those who reported consumption of at least 60 grams or more (10 grams = 1 standard drink) of pure alcohol on at least one occasion in the past 30 days.
* Yoga\_practice\_meditation\_asana\_and\_pranayama -- Refers to yoga/meditation/asanas/pranayama etc. practicing daily or more than once a week.
* Cardiovascular\_diseases -- Cardiovascular diseases include hypertension, heart disease and stroke (any one or more).
* Hypertension\_or\_high\_blood\_pressure -- Having blood pressure measures consistently above normal may result in a diagnosis of high blood pressure (or hypertension).
* Stroke -- An ischemic stroke occurs when the blood supply to part of the brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients.
* Chronic\_lung\_diseases -- Chronic lung diseases include chronic obstructive pulmonary disease (COPD), asthma and chronic bronchitis (any one or more).
* Chronic\_Obstructive\_Pulmonary\_Disease -- Chronic obstructive pulmonary disease (COPD) is a chronic inflammatory lung disease that causes obstructed airflow from the lungs. Symptoms include breathing difficulty, cough, mucus (sputum) production and wheezing.
* Depression -- A group of conditions associated with the elevation or lowering of a person's mood, such as depression or bipolar disorder
* Prevalence\_of\_diagnosed\_Cancer\_or\_Malignant\_Tumor -- Includes oral cancer, breast cancer, cervical cancer, lung cancer or cancer/malignant tumors originating in any other organs/ body parts (any one or more).
* Cataract -- A cataract is a cloudy lens.
* Glaucoma -- Glaucoma is a group of eye conditions that damage the optic nerve. The optic nerve sends visual information from your eye to your brain and is vital for good vision. Damage to the optic nerve is often related to high pressure in your eye. But glaucoma can happen even with normal eye pressure.\
* Typhoid -- Typhoid fever is an infection that spreads through contaminated food and water.
* Jaundice/Hepatitis -- Jaundice is a condition in which the skin, whites of the eyes and mucous membranes turn yellow because of a high level of bilirubin, a yellow-orange bile pigment. Jaundice has many causes, including hepatitis, gallstones and tumors
* Sleep\_problems -- Sleep problems refer to difficulty in falling asleep or staying asleep throughout the night and also feeling unrested during the day (any one or more).Cleaning the dataset

# Assumption

* I assume that state which have high rate open defecation has also high rate in cholera, typhoid and hepatitis since practicing open defecation causes these illnesses.
* States which are high in exposed to indoor pollution has also have high lung diseases rate, stroke, COPD and chronic lung diseases.
* Those states with high rate in yoga should have low rate in depression, mental health problems, hypertension, and also heart diseases.
* Smoking and tobacco not only affect single person but indirectly affect family too. It leads to lung disease, cancer, stroke, heart problems.
* Tobacco affects eye and causes cataracts and glaucoma.
* Episodic drinking causes cancer, cardiovascular disorder.
* I assume cholesterol level may be reason for the heart diseases.

## Required packages and loading dataset

library(lattice)  
library(rmarkdown)  
library(dplyr)

##   
## 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(readxl)  
df <-read\_excel("e last nivi.xls")  
head(df)

## # A tibble: 6 × 28  
## Indicators MeanH…¹ House…² House…³ House…⁴ House…⁵ Curre…⁶ Curre…⁷ Preva…⁸  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 INDIA 5.7 72.3 27.5 52.4 14 12.2 19.8 2.9   
## 2 Andaman & Nic… 5.3 90.6 9.4 64.9 10.5 5.1 23 6.9   
## 3 Andhra Pradesh 4.6 84.7 15.2 85 7.5 12.8 6 2   
## 4 Arunachal Pra… 5.4 96 3.9 53.1 17.4 5.3 5.6 9.2   
## 5 Assam 5.3 95.9 4 34.1 12.7 6.6 42.3 8.2   
## 6 Bihar 6.6 43.2 56.5 31.5 23.1 8.7 27.5 0.34  
## # … with 19 more variables: Yoga\_practice\_meditation\_asana\_and\_pranayama <dbl>,  
## # Cardiovascular\_diseases <dbl>, Hypertension\_or\_high\_blood\_pressure <dbl>,  
## # Chronic\_heart\_diseases <dbl>, Stroke <dbl>,  
## # Diabetes\_or\_high\_blood\_sugar <dbl>, High\_Cholesterol <dbl>, Anaemia <dbl>,  
## # Chronic\_lung\_diseases <dbl>, Chronic\_Obstructive\_Pulmonary\_Disease <dbl>,  
## # Prevalence\_of\_diagnosed\_Cancer\_or\_Malignant\_Tumor <dbl>, Cataract <dbl>, …

## Exploring the dataset

head(df)

## # A tibble: 6 × 28  
## Indicators MeanH…¹ House…² House…³ House…⁴ House…⁵ Curre…⁶ Curre…⁷ Preva…⁸  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 INDIA 5.7 72.3 27.5 52.4 14 12.2 19.8 2.9   
## 2 Andaman & Nic… 5.3 90.6 9.4 64.9 10.5 5.1 23 6.9   
## 3 Andhra Pradesh 4.6 84.7 15.2 85 7.5 12.8 6 2   
## 4 Arunachal Pra… 5.4 96 3.9 53.1 17.4 5.3 5.6 9.2   
## 5 Assam 5.3 95.9 4 34.1 12.7 6.6 42.3 8.2   
## 6 Bihar 6.6 43.2 56.5 31.5 23.1 8.7 27.5 0.34  
## # … with 19 more variables: Yoga\_practice\_meditation\_asana\_and\_pranayama <dbl>,  
## # Cardiovascular\_diseases <dbl>, Hypertension\_or\_high\_blood\_pressure <dbl>,  
## # Chronic\_heart\_diseases <dbl>, Stroke <dbl>,  
## # Diabetes\_or\_high\_blood\_sugar <dbl>, High\_Cholesterol <dbl>, Anaemia <dbl>,  
## # Chronic\_lung\_diseases <dbl>, Chronic\_Obstructive\_Pulmonary\_Disease <dbl>,  
## # Prevalence\_of\_diagnosed\_Cancer\_or\_Malignant\_Tumor <dbl>, Cataract <dbl>, …

names(df)

## [1] "Indicators"   
## [2] "MeanHouseholdSize"   
## [3] "Households\_with\_improved\_sanitation"   
## [4] "Households\_practicing\_open\_defecation"   
## [6] "Households\_exposed\_to\_indoor\_pollution"   
## [7] "Currently\_smoking"   
## [8] "Currently\_consuming\_tobacco"   
## [9] "Prevalence\_of\_heavy\_episodic\_drinking"   
## [10] "Yoga\_practice\_meditation\_asana\_and\_pranayama"   
## [11] "Cardiovascular\_diseases"   
## [12] "Hypertension\_or\_high\_blood\_pressure"   
## [13] "Chronic\_heart\_diseases"   
## [14] "Stroke"   
## [15] "Diabetes\_or\_high\_blood\_sugar"   
## [16] "High\_Cholesterol"   
## [18] "Chronic\_lung\_diseases"   
## [19] "Chronic\_Obstructive\_Pulmonary\_Disease"   
## [21] "Depression"   
## [22] "Prevalence\_of\_diagnosed\_Cancer\_or\_Malignant\_Tumor"  
## [23] "Cataract"   
## [24] "Glaucoma"   
## [26] "Typhoid"   
## [28] "Sleep\_problems"

glimpse(df)

## Rows: 37  
## Columns: 28  
## $ Indicators <chr> "INDIA", "Andaman & …  
## $ MeanHouseholdSize <dbl> 5.7, 5.3, 4.6, 5.4, …  
## $ Households\_with\_improved\_sanitation <dbl> 72.3, 90.6, 84.7, 96…  
## $ Households\_practicing\_open\_defecation <dbl> 27.5, 9.4, 15.2, 3.9…  
## $ Households\_exposed\_to\_indoor\_pollution <dbl> 14.0, 10.5, 7.5, 17.…  
## $ Currently\_smoking <dbl> 12.2, 5.1, 12.8, 5.3…  
## $ Currently\_consuming\_tobacco <dbl> 19.8, 23.0, 6.0, 5.6…  
## $ Prevalence\_of\_heavy\_episodic\_drinking <dbl> 2.90, 6.90, 2.00, 9.…  
## $ Yoga\_practice\_meditation\_asana\_and\_pranayama <dbl> 10.80, 8.80, 3.60, 2…  
## $ Cardiovascular\_diseases <dbl> 27.7, 39.7, 35.3, 20…  
## $ Hypertension\_or\_high\_blood\_pressure <dbl> 25.8, 38.5, 33.1, 19…  
## $ Chronic\_heart\_diseases <dbl> 3.60, 2.80, 3.50, 0.…  
## $ Stroke <dbl> 1.80, 2.10, 1.80, 1.…  
## $ Diabetes\_or\_high\_blood\_sugar <dbl> 11.5, 19.4, 17.9, 6.…  
## $ High\_Cholesterol <dbl> 2.10, 3.90, 0.80, 0.…  
## $ Chronic\_lung\_diseases <dbl> 6.20, 3.90, 6.80, 2.…  
## $ Chronic\_Obstructive\_Pulmonary\_Disease <dbl> 2.10, 1.10, 2.20, 1.…  
## $ Depression <dbl> 0.60, 1.10, 0.63, 0.…  
## $ Prevalence\_of\_diagnosed\_Cancer\_or\_Malignant\_Tumor <dbl> 0.60, 1.20, 0.36, 0.…  
## $ Cataract <dbl> 13.1, 4.7, 11.2, 2.5…  
## $ Glaucoma <dbl> 1.80, 2.90, 3.70, 1.…  
## $ Typhoid <dbl> 5.80, 3.20, 6.90, 5.…  
## $ Sleep\_problems <dbl> 12.7, 15.6, 10.1, 5.…

**Missing data:**

sapply(df, function(x) sum(is.na(x)))

## Indicators   
## 0   
## MeanHouseholdSize   
## 0   
## Households\_with\_improved\_sanitation   
## 0   
## Households\_practicing\_open\_defecation   
## 0   
## Households\_exposed\_to\_indoor\_pollution   
## 0   
## Currently\_smoking   
## 0   
## Currently\_consuming\_tobacco   
## 0   
## Prevalence\_of\_heavy\_episodic\_drinking   
## 0   
## Yoga\_practice\_meditation\_asana\_and\_pranayama   
## 0   
## Cardiovascular\_diseases   
## 0   
## Hypertension\_or\_high\_blood\_pressure   
## 0   
## Chronic\_heart\_diseases   
## 0   
## Stroke   
## 0   
## High\_Cholesterol   
## 0   
## Chronic\_lung\_diseases   
## 0   
## Chronic\_Obstructive\_Pulmonary\_Disease   
## 0   
## Depression   
## 0   
## Prevalence\_of\_diagnosed\_Cancer\_or\_Malignant\_Tumor   
## 0   
## Cataract   
## 0   
## Glaucoma   
## 0   
## Typhoid   
## 0   
## Sleep\_problems   
## 0

Since some of the datatype of numrical variable is in character. We changed that into numeric type.

class(df$Households\_practicing\_open\_defecation)

## [1] "numeric"

df$Households\_practicing\_open\_defecation <- as.numeric(df$Households\_practicing\_open\_defecation)  
  
df$Chronic\_Obstructive\_Pulmonary\_Disease <- as.numeric(df$Chronic\_Obstructive\_Pulmonary\_Disease)  
  
df$Depression <- as.numeric(df$Depression)  
  
df$Prevalence\_of\_diagnosed\_Cancer\_or\_Malignant\_Tumor <- as.numeric(df$Prevalence\_of\_diagnosed\_Cancer\_or\_Malignant\_Tumor)  
  
df$Glaucoma <- as.numeric(df$Glaucoma)

## Basic summary of the dataset:

The average persons in each house is 5 members. The current smoking has max of 34 but the min, mean, Q3 value median value are 3, 10 and 16.So the max value look like outlier.

summary(df$MeanHouseholdSize)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.200 5.100 5.600 5.557 6.100 7.100

summary(df$Households\_with\_improved\_sanitation)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 43.20 74.30 87.40 84.39 96.50 100.00

summary(df$Households\_practicing\_open\_defecation)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 3.40 12.60 15.33 25.70 56.50

summary(df$Currently\_smoking)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.4 6.4 10.0 12.6 16.5 34.0

## 

## Histogram analysis:

df\_numeric <- select\_if(df, is.numeric)  
n <- ncol(df\_numeric)

for (col in colnames(df\_numeric)) {  
 hist(df[[col]], main = paste("Histogram of", col), xlab = "Values", col = "pink")}

* Figure 1.1 : The indoor pollution is positively skewed but the more states has values ranging from 0 to 15. But few states has values greater than 15. They are most northern states of India, dadra and nagar haveli is the only union territory.
* Figure 1.2: The currently smoking is positively skewed and Mizoram is the only states which has value greater than 30.
* Figure 1.3: The currently consuming tobacco is positively skewed means many states has low rate of consuming tobacco. There is only few states which are high in 40. Assam and Odisha has high rate in tobacco consumption it may due to availability and culture that chewing tobacco is common there.

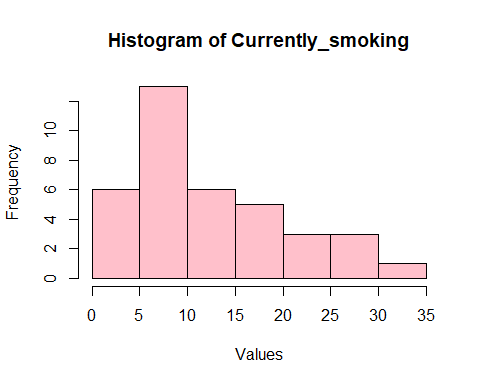
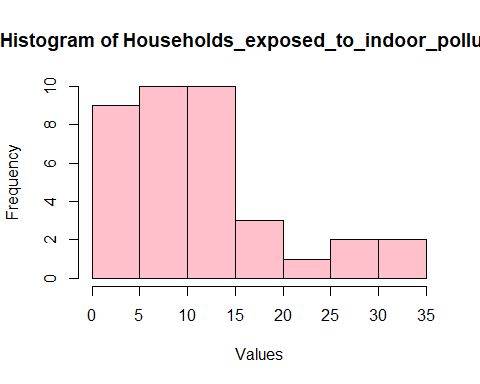


Fig 1.1 and 1.2

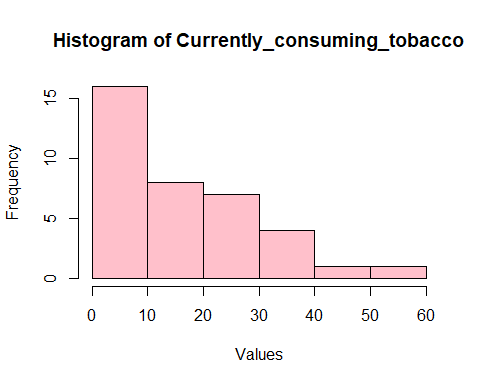


Figure 1.3

df %>%   
 select(Indicators,Households\_exposed\_to\_indoor\_pollution) %>%  
 filter(Households\_exposed\_to\_indoor\_pollution > 15)

## # A tibble: 8 × 2  
## Indicators Households\_exposed\_to\_indoor\_pollution  
## <chr> <dbl>  
## 1 Arunachal Pradesh 17.4  
## 2 Bihar 23.1  
## 3 Dadra & Nagar Haveli 15.2  
## 4 Haryana 27.3  
## 5 Jharkhand 32.1  
## 6 Punjab 28.4  
## 7 Uttar Pradesh 30.9  
## 8 West Bengal 15.9

df %>%   
 select(Indicators,Currently\_smoking) %>%  
 filter(Currently\_smoking > 30)

## # A tibble: 1 × 2  
## Indicators Currently\_smoking  
## <chr> <dbl>  
## 1 Mizoram 34

df %>%   
 select(Indicators,Currently\_consuming\_tobacco) %>%  
 filter(Currently\_consuming\_tobacco > 40)

## # A tibble: 2 × 2  
## Indicators Currently\_consuming\_tobacco  
## <chr> <dbl>  
## 1 Assam 42.3  
## 2 Odisha 55.4

#hist(df$MeanHouseholdSize, col = "pink")

* Fig: 2.1 The household with proved sanitation is negatively skewed.The sanitation condition is improved over the frequency which is good that most of the states has sanitation condition high over the average. 40 % of the total states has values ranges over 90 which is great. Here the mean value is 84 but it is affected by the extreme values.
* Bihar, Odisha and UttarPradesh has very lower sanitation rate.

histogram(df$Households\_with\_improved\_sanitation,col='darkgreen')

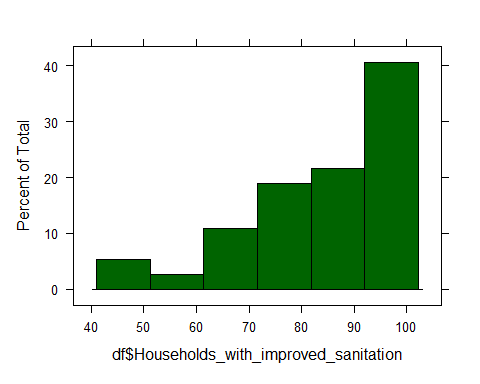


Fig: 2.1

df %>%   
 select(Indicators,Households\_with\_improved\_sanitation) %>%  
 filter(Households\_with\_improved\_sanitation < 62)

## # A tibble: 3 × 2  
## Indicators Households\_with\_improved\_sanitation  
## <chr> <dbl>  
## 1 Bihar 43.2  
## 2 Odisha 56.2  
## 3 Uttar Pradesh 49.6

* Figure 2.2: The household practicing open defaction is postively skewed. It indicates that most of the states are less in practicing open defaction. The mean value is 15 and median is 12.The 40 to 50 percent for open defecation is only for less than 10. Bihar, Odisha and UttarPradesh has the high percentge of practicing open defaction but this states are also having high sanitation percentage.
* I thought states which has poor sanitation condition also has poor condition in open defecation but my assumption is wrong that Bihar, Odisha and UttarPradesh is states with good sanitation percentage along with states which practicing open defecation.
* The current smoking is postiively distributed and 5 of the states has percentage more than 12.

histogram(df$Households\_practicing\_open\_defecation,col='red')

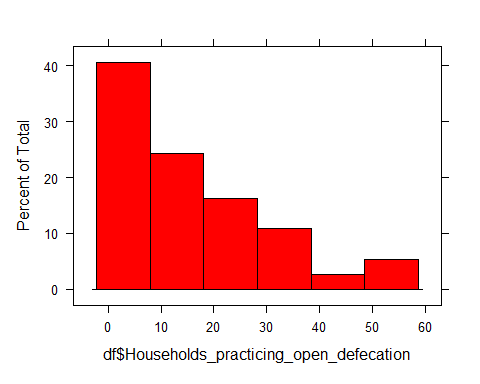


Figure 2.2

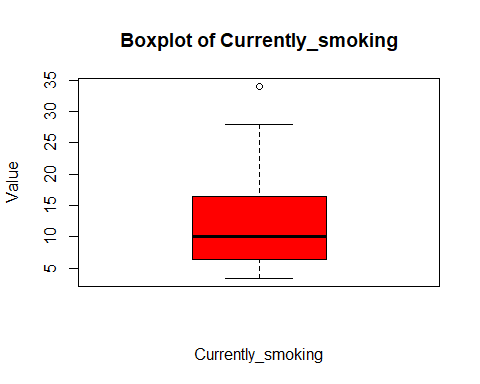
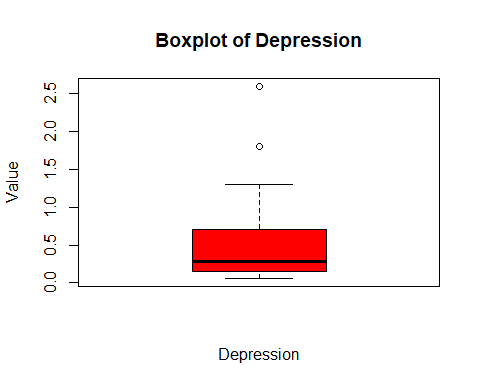
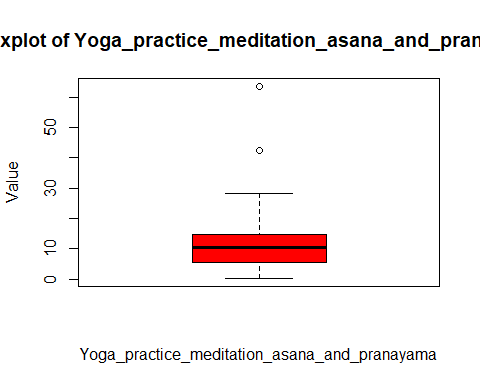
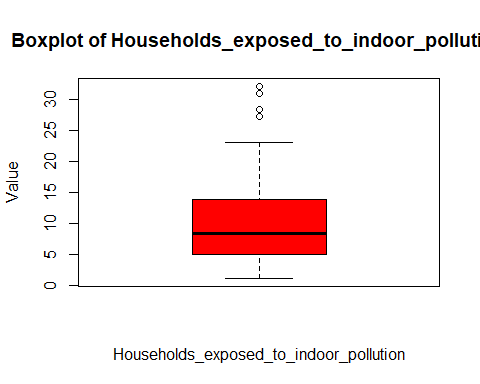
df %>%   
 select(Indicators,Households\_practicing\_open\_defecation) %>%  
 filter(Households\_practicing\_open\_defecation > 40)

## # A tibble: 3 × 2  
## Indicators Households\_practicing\_open\_defecation  
## <chr> <dbl>  
## 1 Bihar 56.5  
## 2 Odisha 43.8  
## 3 Uttar Pradesh 50.3

## Boxplot

df\_boxplot <- df %>%  
 select(Households\_practicing\_open\_defecation,Households\_exposed\_to\_indoor\_pollution,Yoga\_practice\_meditation\_asana\_and\_pranayama,Depression,Currently\_smoking)  
my\_colors <- c("red", "blue")  
for (col in colnames(df\_boxplot)) {  
 boxplot(df[,col], main = paste("Boxplot of", col), xlab = col, ylab = "Value",col = my\_colors)  
}

* Fig 3.1: From the figure, There is some outliers in the household exposed to indoor pollution which is values higher than 25.
* Fig 3.2: There is two states which has high percentage in practicing yoga. Punjab and Nagaland are the two states which practice yoga high than other states.
* Fig 3.3: The depression figure show that states have low depression rates from 0 to 1 but few with 2 and 2.5
* Fig 3.4: There is only one states which has rate of 35 and mostly there more states which are higher than the median value that is 9.



## Scatterplot

* Figure 1: This figure shows that household practicing open defecation and typhoid has weak positive corelation and has some outliers.
* This shows that there is a possibility of increasing in percentage of typhoid if we follow open defecation.

typ\_def <- df %>%  
 select(Households\_practicing\_open\_defecation,Typhoid,Indicators)   
  
xyplot(Typhoid~Households\_practicing\_open\_defecation,typ\_def)

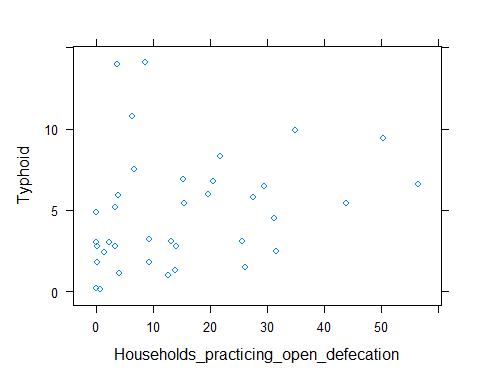


Figure 1

* Figure 2: I assumed indoor pollution causes COPD and stroke but there is no strong corelation between them. The data is spreaded randomly.
* There must be some other factors which is reason for COPD and stroke in households.

#indoor <- df %>%   
 #dplyr::filter(Households\_exposed\_to\_indoor\_pollution > 5 & Households\_exposed\_to\_indoor\_pollution < 14)   
  
xyplot(Households\_exposed\_to\_indoor\_pollution ~ Chronic\_Obstructive\_Pulmonary\_Disease + Stroke,df)

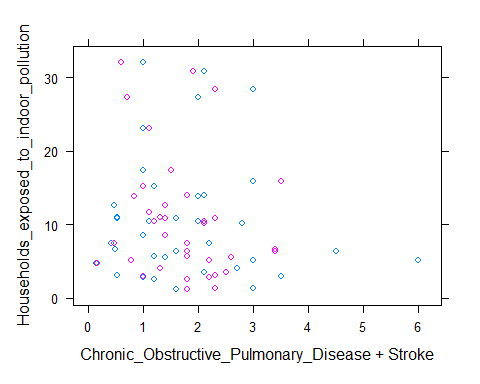


Figure 2

Figure 3: Depression has positive corelation with practicing yoga but sleep problem has neutral corelation with procaticing yoga and most of the points are lie below 20 with resepct to practice. I am not going to remove outlier for practicing yoga since it omits two states which are important.

xyplot(Yoga\_practice\_meditation\_asana\_and\_pranayama ~ Depression + Sleep\_problems ,df)

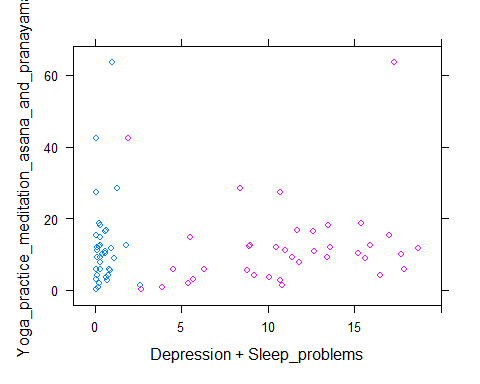


Figure 3

Fig 4: Both has positive corelation with COPD. Smoking and tobacoo is well known for COPD and in few case only states that has low smoking and tobacco rate has COPD.

xyplot(Chronic\_Obstructive\_Pulmonary\_Disease ~ Currently\_smoking + Currently\_consuming\_tobacco ,df)

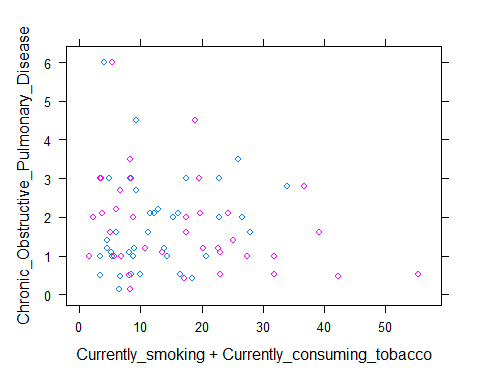


Fig 4

Fig 5: There is weak correlation between chronic heart disease and other two. So there me some other factor which may reason for heart diseases.

xyplot(Chronic\_heart\_diseases ~ Currently\_smoking + Currently\_consuming\_tobacco ,df)

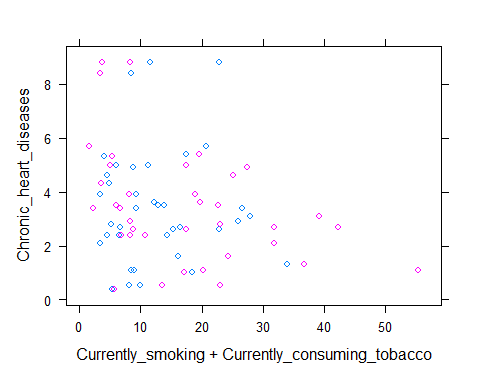


Fig 5

Fig 6: The cardiovascular diseases and high cholesterol has high positive corelation. Most of the data points lies between o to 5 range.

xyplot(Cardiovascular\_diseases~High\_Cholesterol,df)

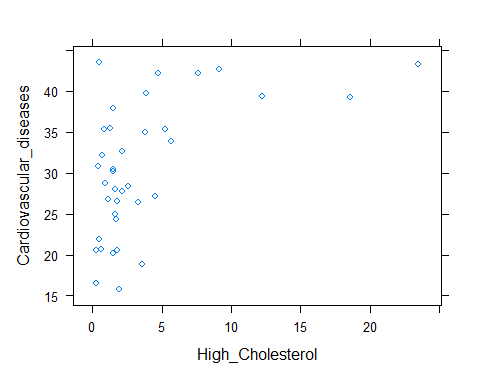


Fig 6