Ava Masucci

Statistics and Plots for Capstone project

I am interested in determining which variables are correlated with the number (or percentage) of women diagnosed with chronic illnesses. I used the summary() function to learn more about this variable:

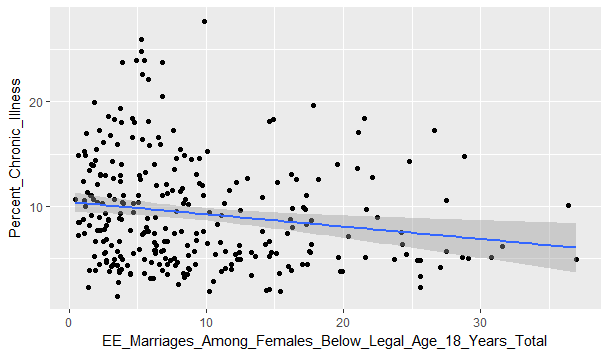
summary(IndiaMean$KK\_Having\_Diagnosed\_For\_Chronic\_Illness\_Per\_100000\_Population\_Any\_Kind\_Of\_Chronic\_Illness\_Female\_Total)  
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 1441 5164 8239 9343 12539 27701

I made several scatterplots to visualize the correlation between independent variables and the dependant variable:

* Are the percentages of women married below the age of 18 and those diagnosed with chronic illnesses correlated?

> MarriagevsChronic <- ggplot(IndiaPercent, aes(EE\_Marriages\_Among\_Females\_Below\_Legal\_Age\_18\_Years\_Total, Percent\_Chronic\_Illness)) + geom\_point() + geom\_smooth(method='lm')

> print(MarriagevsChronic)

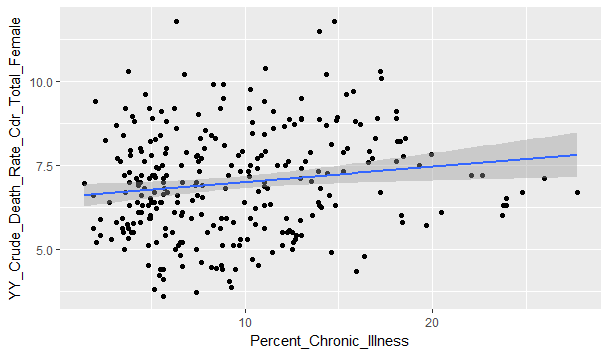


The scatterplot above shows a slight negative correlation between chronic illness and the percent of marriages among females below the legal age of 18 years.

* Is the Crude Death Rate correlated with chronic illnesses?

CDRvsChronic <- ggplot(IndiaPercent, aes(Percent\_Chronic\_Illness, YY\_Crude\_Death\_Rate\_Cdr\_Total\_Female)) + geom\_point() + geom\_smooth(method='lm')

> print(CDRvsChronic)

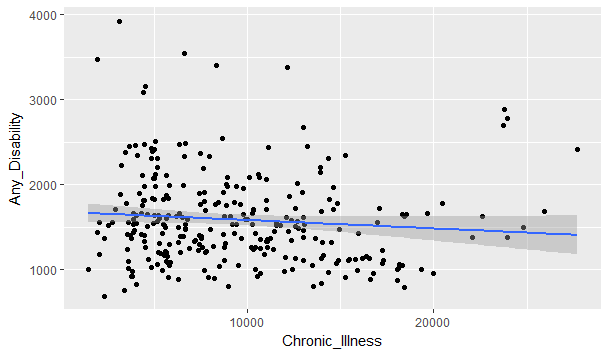


This scatterplot shows a slight positive correlation between the percentage of chronically ill women and the Crude Death Rate (CDR).

* Are disabilities correlated with chronic illness?

DisabilityvsChronic <- ggplot(IndiaMean, aes(Chronic\_Illness, Any\_Disability)) + geom\_point() + geom\_smooth(method='lm')

> print(DisabilityvsChronic)

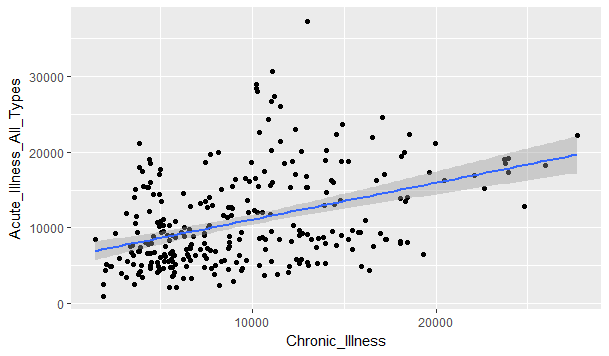


From the above scatterplot, there’s an extremely slight negative correlation between chronic illnesses and having a disability in women.

* Are acute illness cases correlated with chronic illness cases?

AcutevsChronic <- ggplot(IndiaMean, aes(Chronic\_Illness, Acute\_Illness\_All\_Types)) + geom\_point() + geom\_smooth(method='lm')

> print(AcutevsChronic)

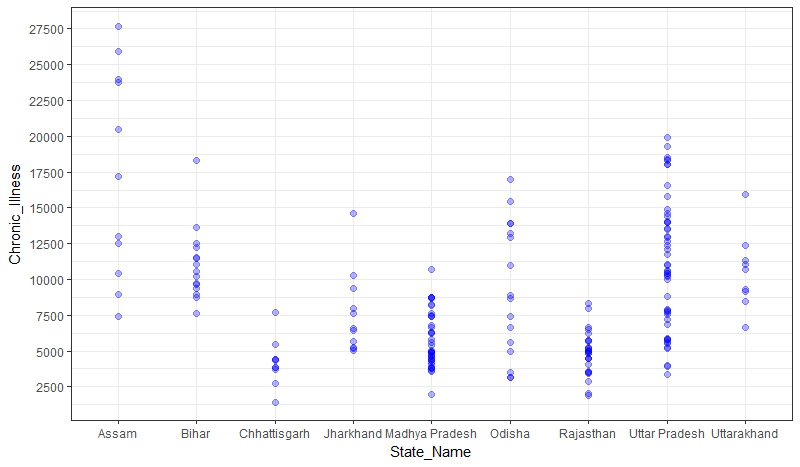


Cases of acute illness are positively correlated with chronic illness. Logically, this makes sense because many times chronic illnesses weaken the immune system.

* How do the number of chronic illnesses vary by state?

StatevsChronic <- ggplot(IndiaHealthnoNas, aes(State\_Name, Chronic\_Illness)) + geom\_point(alpha = 3/10, colour = "blue", size = 2) + theme\_bw() + scale\_y\_continuous(breaks=seq(0,30000,2500))

> print(StatevsChronic)



The above plot illustrates the amount of women diagnosed with chronic illnesses per state. States are divided into districts and health data is provided per district. Districts in the states of Assam and Uttar Pradesh generally have more cases of chronic illness than the other states. Districts in Chhattisgarh and Rajasthan are generally below the median of diagnosed chronic illnesses (8239 per 100,000). Assam, Bihar, and Uttarakhand are generally above the median. State seems to be a reliable predictor of the number of diagnosed chronic illnesses.

Conclusions:

Data visualization tools help frame my problem as a classification problem. Logically, some of the variables I examined should be highly correlated with chronic illnesses. However, the plots reveal that the independent variables are uncorrelated with my dependent variable, with the exception of acute illness cases. This data set is not well suited for a linear or random forest regression models. Logistic regression and random forest classification models will provide more accurate predictions.