Ava Masucci

Intro to Data Science Course

Here are the most important steps I took in cleaning and wrangling my data:

The original data set had 642 variables (columns) and 284 rows. The column names begin with letters such as “AA”, “BB”, “CC”, etc. I deleted strings of columns with the following code:

drop <- Indiahealth[, -grep("AA", colnames(Indiahealth))]

After narrowing my columns down to less than 100, I used the following code to delete more:

> View(drop)

> India28 <- India[, c(2,3,4,6,8,11,12,13,18,24,25,29,38,40,55,56,57,58,61,76:85)]

I created binary columns for each state:

> IndiaMean$State\_Category\_Assam <- ifelse(IndiaMean$State\_Name == "Assam", 1, 0)

> IndiaMean$State\_Category\_Bihar <- ifelse(IndiaMean$State\_Name == "Bihar", 1, 0)

> IndiaMean$State\_Category\_Chhattisgarh <- ifelse(IndiaMean$State\_Name == "Chhattisgarh", 1, 0)

> IndiaMean$State\_Category\_Jharkhand <- ifelse(IndiaMean$State\_Name == "Jharkhand", 1, 0)

> IndiaMean$State\_Category\_Madhya\_Pradesh <- ifelse(IndiaMean$State\_Name == "Madhya Pradesh", 1, 0)

> IndiaMean$State\_Category\_Odisha <- ifelse(IndiaMean$State\_Name == "Odisha", 1, 0)

> IndiaMean$State\_Category\_Rajasthan <- ifelse(IndiaMean$State\_Name == "Rajasthan", 1, 0)

> IndiaMean$State\_Category\_Uttar\_Pradesh <- ifelse(IndiaMean$State\_Name == "Uttar Pradesh", 1, 0)

> IndiaMean$State\_Category\_Uttarakhand <- ifelse(IndiaMean$State\_Name == "Uttarakhand", 1, 0)

>

Next, I deleted the text categories of state and district names:

> IndiaMean <- IndiaMean[, -c(1,2)]

>

Then I created a binary column, GreaterThanMedianChronicIllness, for the variable I’m interested in analyzing:

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

> IndiaMean$GreaterThanMedianChronicIllness <- ifelse(IndiaMean$KK\_Having\_Diagnosed\_For\_Chronic\_Illness\_Per\_100000\_Population\_Any\_Kind\_Of\_Chronic\_Illness\_Female\_Total > 8239, 1, 0)

>

> summary(IndiaMean$GreaterThanMedianChronicIllness)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0 0.0 0.5 0.5 1.0 1.0

I replaced the NA’s (missing data) with the mean for each column. Deleting the missing data resulted in deleting too much of my data set:

IndiaMean[] <- lapply(IndiaMean, function(x) {

+ x[is.na(x)] <- mean(x, na.rm = TRUE)

+ x

+ })

Warning message:

In mean.default(x, na.rm = TRUE) :

argument is not numeric or logical: returning NA

> View(IndiaMean)

> row.has.na <- apply(IndiaMean, 1, function(x){any(is.na(x))})

> sum(row.has.na)

[1] 0

I saved my data set as the following:

> saveRDS(IndiaMean, file="IndiaMean.Rda")

For further analysis, I converted the column “Having\_Diagnosed\_For\_Chronic\_Illness\_Per\_100000\_Population\_Any\_Kind\_Of\_Chronic\_Illness\_Female\_Total” to a percentage column, Percent\_Chronic\_Illness, with the following code:

> IndiaRegRF$Percent\_Chronic\_Illness <- (IndiaRegRF$KK\_Having\_Diagnosed\_For\_Chronic\_Illness\_Per\_100000\_Population\_Any\_Kind\_Of\_Chronic\_Illness\_Female\_Total/100000)\*100