BRFSS 2017 Survey EDA - Health indicators

Amanda Kimball

11/17/2020

The Smart: City and county survey data put together by the Center for Disease Control and Prevention: Behavioral risk factor surveillance system (Smart data, 2017). Contains a number of survey response data that is equally distributed across a range of areas referred to as Metropolitan and Micropolitan statistical areas. The survey data covers health indicators, diet choices, health insurance availability, demographics, and economic factors. This file will focus on the diabetes heath indicator.

library(foreign)  
BRFSS2017 <- read.xport("MMSA2017.XPT")

I needed to create an additional factor with the county MMSA name and state separated from the full MMSANAME field. I used the gsub function to remove the MSA test and created the 2 new columns.

I located the following variables within the dataset:

These variables where inline with the risk factors listed on the cdc website: The most important factors listed by the CDC as conditions that increase a risk for COVID19 are: cancer, chronic kidney disease, chronic obstructive pulmonary disease, heart disease, immunocompromised organ transplant, obesity, sickle cell disease, smoking, and type 2 diabetes mellitus (CDC, 2020b).

BPHIGH4 - Ever been told you have high blood pressure? Yes - 1, Yes pregnant - 2, No - 3, No borderline - 4, not sure - 7, Refused - 9, blanks TOLDHI2 - Ever been told you have high cholestrol? Yes - 1, No - 2, not sure - 7, Refused - 9, blanks CVDCRHD4 - Ever Diagnosed with Angina or Coronary Heart Disease? Yes - 1, No - 2, not sure - 7, Refused - 9, blanks CHCOCNCR - Ever diagnosed with cancer (non-skin)? Yes - 1, No - 2, Not sure - 7, Refused - 9, blanks CHCKIDNY - Ever diagnosed with kidney disease? Yes - 1, No - 2, Not sure - 7, Refused - 9, blanks DIABETE3 - Ever diagnosed with diabetes? Yes - 1, Yes pregnant - 2, No - 3, No prediabetes - 4, not sure - 7, Refused - 9, blanks RFSMOK3 - Adults who are current smokers? No - 1, Yes - 2, not sure/Refused - 9, blanks

I added the following due to it’s potential to impact covid prevalence: i.e. large household size, lack of healthcare,

GENHLTH - Would you say that in general your health is: Excellent - 1, Very Good - 2, Good - 3, Fair - 4, Poor - 5, not sure -7, Refused - 9, blanks HLTHPLN1 - Do you have any kind of health care coverage? Yes - 1, No - 2, Not sure - 7, Refused - 9, blanks

data <- subset(BRFSS2017, select = c(5, 9, 13, 16, 19, 24, 28:29, 124, 133, 174, 175, 177))  
data$MMSANAME <- lapply(data$MMSANAME, as.character)  
  
data$MMSANAME <- gsub(", Metropolitan Statistical Area", "", data$MMSANAME)  
data$MMSANAME <- gsub(", Micropolitan Statistical Area", "", data$MMSANAME)  
data$MMSANAME <- gsub(", Metropolitan Division", "", data$MMSANAME)  
names(data) <- c("General\_health", "Health\_coverage", "Blood\_pressure\_high", "Cholesterol\_high",  
 "Coronary\_heart\_disease", "cancer", "Kidney\_disease", "Diabetes", "Age", "Smoke",  
 "MMSA\_Number", "WT\_Factor", "MMSA\_NAME")  
str(data)

## 'data.frame': 230875 obs. of 13 variables:  
## $ General\_health : num 3 2 2 4 1 2 2 2 2 3 ...  
## $ Health\_coverage : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Blood\_pressure\_high : num 1 3 3 3 3 3 3 3 1 3 ...  
## $ Cholesterol\_high : num 2 2 2 2 2 2 2 1 2 2 ...  
## $ Coronary\_heart\_disease: num 2 2 2 2 2 2 2 2 2 2 ...  
## $ cancer : num 2 1 2 2 1 2 1 2 1 2 ...  
## $ Kidney\_disease : num 2 2 2 2 2 2 2 2 2 1 ...  
## $ Diabetes : num 3 3 3 1 3 3 3 4 3 3 ...  
## $ Age : num 61 65 21 28 80 31 67 80 80 80 ...  
## $ Smoke : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ MMSA\_Number : num 10100 10100 10100 10100 10100 10100 10100 10100 10100 10100 ...  
## $ WT\_Factor : num 110.2 28.6 115.6 376 20.7 ...  
## $ MMSA\_NAME : chr "Aberdeen, SD" "Aberdeen, SD" "Aberdeen, SD" "Aberdeen, SD" ...

Here are the calculated body mass indicator variables. As shown there are several missing values.

summary(data)

## General\_health Health\_coverage Blood\_pressure\_high Cholesterol\_high  
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000   
## Median :2.000 Median :1.000 Median :3.000 Median :2.000   
## Mean :2.549 Mean :1.104 Mean :2.242 Mean :1.683   
## 3rd Qu.:3.000 3rd Qu.:1.000 3rd Qu.:3.000 3rd Qu.:2.000   
## Max. :9.000 Max. :9.000 Max. :9.000 Max. :9.000   
## NA's :2 NA's :13064   
## Coronary\_heart\_disease cancer Kidney\_disease Diabetes   
## Min. :1.000 Min. :1.000 Min. :1.00 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:2.00 1st Qu.:3.000   
## Median :2.000 Median :2.000 Median :2.00 Median :3.000   
## Mean :1.988 Mean :1.918 Mean :1.98 Mean :2.768   
## 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:2.00 3rd Qu.:3.000   
## Max. :9.000 Max. :9.000 Max. :9.00 Max. :9.000   
## NA's :1 NA's :2   
## Age Smoke MMSA\_Number WT\_Factor   
## Min. :18.00 Min. :1.000 Min. :10100 Min. : 0.15   
## 1st Qu.:40.00 1st Qu.:1.000 1st Qu.:20260 1st Qu.: 116.28   
## Median :57.00 Median :1.000 Median :34820 Median : 292.12   
## Mean :54.34 Mean :1.471 Mean :31568 Mean : 684.76   
## 3rd Qu.:68.00 3rd Qu.:1.000 3rd Qu.:40060 3rd Qu.: 723.15   
## Max. :80.00 Max. :9.000 Max. :49340 Max. :43227.53   
##   
## MMSA\_NAME   
## Length:230875   
## Class :character   
## Mode :character   
##   
##   
##   
##

The 9 value’s are essential N/A’s. Everything else is presents absence characteristics. Even the high cholesterol indication is proceeded by a question asking if the participant had cholesterol in the last 5 years. THe N/As here are mostly participants indicating that they have not had a cholesterol check.

All N/As will be coded as 9.

index <- is.na(data[1])  
data[index, 1] <- 9  
index <- is.na(data[4])  
data[index, 4] <- 9  
index <- is.na(data[7])  
data[index, 7] <- 9  
index <- is.na(data[8])  
data[index, 8] <- 9  
  
summary(data)

## General\_health Health\_coverage Blood\_pressure\_high Cholesterol\_high  
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000   
## Median :2.000 Median :1.000 Median :3.000 Median :2.000   
## Mean :2.549 Mean :1.104 Mean :2.242 Mean :2.097   
## 3rd Qu.:3.000 3rd Qu.:1.000 3rd Qu.:3.000 3rd Qu.:2.000   
## Max. :9.000 Max. :9.000 Max. :9.000 Max. :9.000   
## Coronary\_heart\_disease cancer Kidney\_disease Diabetes   
## Min. :1.000 Min. :1.000 Min. :1.00 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:2.00 1st Qu.:3.000   
## Median :2.000 Median :2.000 Median :2.00 Median :3.000   
## Mean :1.988 Mean :1.918 Mean :1.98 Mean :2.768   
## 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:2.00 3rd Qu.:3.000   
## Max. :9.000 Max. :9.000 Max. :9.00 Max. :9.000   
## Age Smoke MMSA\_Number WT\_Factor   
## Min. :18.00 Min. :1.000 Min. :10100 Min. : 0.15   
## 1st Qu.:40.00 1st Qu.:1.000 1st Qu.:20260 1st Qu.: 116.28   
## Median :57.00 Median :1.000 Median :34820 Median : 292.12   
## Mean :54.34 Mean :1.471 Mean :31568 Mean : 684.76   
## 3rd Qu.:68.00 3rd Qu.:1.000 3rd Qu.:40060 3rd Qu.: 723.15   
## Max. :80.00 Max. :9.000 Max. :49340 Max. :43227.53   
## MMSA\_NAME   
## Length:230875   
## Class :character   
## Mode :character   
##   
##   
##

I am going to create dummy variables for each of the categories individually based off their conditions from above. i retain both yes, no and unknown values in case they are needed in future analysis.

For general health, I will have good, bad and unknown categories. I keep an unknown value so that I can remove those individuals from the regional estimate for percentage of the region with a condition.

General Health - Would you say that in general your health is: Excellent - 1, Very Good - 2, Good - 3, Fair - 4, Poor - 5, not sure -7, Refused - 9, blanks

data$Gen\_Hlth\_good <- ifelse(data$General\_health == 1, 1,   
 ifelse(data$General\_health == 2, 1,   
 ifelse(data$General\_health == 3, 1, 0)))  
data$Gen\_Hlth\_bad <- ifelse(data$General\_health == 4, 1,   
 ifelse(data$General\_health == 5, 1, 0))  
data$Gen\_Hlth\_unkn <- ifelse(data$General\_health == 7, 1,   
 ifelse(data$General\_health == 9, 1, 0))  
  
summary(data[14:16])

## Gen\_Hlth\_good Gen\_Hlth\_bad Gen\_Hlth\_unkn   
## Min. :0.0000 Min. :0.0000 Min. :0.000000   
## 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:0.000000   
## Median :1.0000 Median :0.0000 Median :0.000000   
## Mean :0.8211 Mean :0.1764 Mean :0.002543   
## 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:0.000000   
## Max. :1.0000 Max. :1.0000 Max. :1.000000

Health\_coverage - Do you have any kind of health care coverage? Yes - 1, No - 2, Not sure - 7, Refused - 9, blanks

data$Insured <- ifelse(data$Health\_coverage == 1, 1, 0)   
data$Uninsured <- ifelse(data$Health\_coverage == 2, 1, 0)   
data$Unkn\_insurance <- ifelse(data$Health\_coverage == 7, 1,   
 ifelse(data$Health\_coverage == 9, 1, 0))  
  
summary(data[17:19])

## Insured Uninsured Unkn\_insurance   
## Min. :0.0000 Min. :0.0000 Min. :0.000000   
## 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:0.000000   
## Median :1.0000 Median :0.0000 Median :0.000000   
## Mean :0.9191 Mean :0.0769 Mean :0.004006   
## 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:0.000000   
## Max. :1.0000 Max. :1.0000 Max. :1.000000

Blood\_pressure\_high - Ever been told you have high blood pressure? Yes - 1, Yes pregnant - 2, No - 3, No borderline - 4, not sure - 7, Refused - 9, blanks

data$BP\_High <- ifelse(data$Blood\_pressure\_high == 1, 1,   
 ifelse(data$Blood\_pressure\_high == 2, 1,   
 ifelse(data$Blood\_pressure\_high == 4, 1, 0)))  
data$BP\_normal <- ifelse(data$Blood\_pressure\_high == 3, 1, 0)  
data$BP\_unkn <- ifelse(data$Blood\_pressure\_high == 7, 1,   
 ifelse(data$Blood\_pressure\_high == 9, 1, 0))  
  
summary(data[20:22])

## BP\_High BP\_normal BP\_unkn   
## Min. :0.0000 Min. :0.0000 Min. :0.000000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.000000   
## Median :0.0000 Median :1.0000 Median :0.000000   
## Mean :0.4031 Mean :0.5937 Mean :0.003236   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.000000   
## Max. :1.0000 Max. :1.0000 Max. :1.000000

Cholesterol\_high - Ever been told you have high cholestrol? Yes - 1, No - 2, not sure - 7, Refused - 9, blanks

data$Chol\_High <- ifelse(data$Cholesterol\_high == 1, 1, 0)   
data$Chol\_normal <- ifelse(data$Cholesterol\_high == 2, 1, 0)   
data$Unkn\_chol <- ifelse(data$Cholesterol\_high == 7, 1,   
 ifelse(data$Cholesterol\_high == 9, 1, 0))  
  
summary(data[23:25])

## Chol\_High Chol\_normal Unkn\_chol   
## Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.00000   
## Median :0.0000 Median :1.0000 Median :0.00000   
## Mean :0.3478 Mean :0.5862 Mean :0.06599   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.00000   
## Max. :1.0000 Max. :1.0000 Max. :1.00000

Coronary\_heart\_disease - Ever Diagnosed with Angina or Coronary Heart Disease? Yes - 1, No - 2, not sure - 7, Refused - 9, blanks

data$HD\_yes <- ifelse(data$Coronary\_heart\_disease == 1, 1, 0)   
data$HD\_no <- ifelse(data$Coronary\_heart\_disease == 2, 1, 0)   
data$Unkn\_HD <- ifelse(data$Coronary\_heart\_disease == 7, 1,   
 ifelse(data$Coronary\_heart\_disease == 9, 1, 0))  
  
summary(data[26:28])

## HD\_yes HD\_no Unkn\_HD   
## Min. :0.00000 Min. :0.00 Min. :0.000000   
## 1st Qu.:0.00000 1st Qu.:1.00 1st Qu.:0.000000   
## Median :0.00000 Median :1.00 Median :0.000000   
## Mean :0.05221 Mean :0.94 Mean :0.007783   
## 3rd Qu.:0.00000 3rd Qu.:1.00 3rd Qu.:0.000000   
## Max. :1.00000 Max. :1.00 Max. :1.000000

cancer - Ever diagnosed with cancer (non-skin)? Yes - 1, No - 2, Not sure - 7, Refused - 9, blanks

data$cancer\_yes <- ifelse(data$cancer == 1, 1, 0)   
data$cancer\_no <- ifelse(data$cancer == 2, 1, 0)   
data$Unkn\_cancer <- ifelse(data$cancer == 7, 1,   
 ifelse(data$cancer == 9, 1, 0))  
  
summary(data[29:31])

## cancer\_yes cancer\_no Unkn\_cancer   
## Min. :0.0000 Min. :0.0000 Min. :0.000000   
## 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:0.000000   
## Median :0.0000 Median :1.0000 Median :0.000000   
## Mean :0.0953 Mean :0.9022 Mean :0.002508   
## 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:0.000000   
## Max. :1.0000 Max. :1.0000 Max. :1.000000

Kidney\_disease - Ever diagnosed with kidney disease? Yes - 1, No - 2, Not sure - 7, Refused - 9, blanks

data$KD\_yes <- ifelse(data$Kidney\_disease == 1, 1, 0)   
data$KD\_no <- ifelse(data$Kidney\_disease == 2, 1, 0)   
data$Unkn\_KD <- ifelse(data$Kidney\_disease == 7, 1,   
 ifelse(data$Kidney\_disease == 9, 1, 0))  
  
summary(data[32:34])

## KD\_yes KD\_no Unkn\_KD   
## Min. :0.00000 Min. :0.0000 Min. :0.000000   
## 1st Qu.:0.00000 1st Qu.:1.0000 1st Qu.:0.000000   
## Median :0.00000 Median :1.0000 Median :0.000000   
## Mean :0.03724 Mean :0.9594 Mean :0.003318   
## 3rd Qu.:0.00000 3rd Qu.:1.0000 3rd Qu.:0.000000   
## Max. :1.00000 Max. :1.0000 Max. :1.000000

Diabetes - Ever diagnosed with diabetes? Yes - 1, Yes pregnant - 2, No - 3, No prediabetes - 4, not sure - 7, Refused - 9, blanks

data$Diabetes\_yes <- ifelse(data$Diabetes == 1, 1,   
 ifelse(data$Diabetes == 2, 1,   
 ifelse(data$Diabetes == 4, 1, 0)))   
data$Diabetes\_no <- ifelse(data$Diabetes == 3, 1, 0)   
data$Unkn\_Diabetes <- ifelse(data$Diabetes == 7, 1,   
 ifelse(data$Diabetes == 9, 1, 0))  
  
summary(data[35:37])

## Diabetes\_yes Diabetes\_no Unkn\_Diabetes   
## Min. :0.0000 Min. :0.0000 Min. :0.000000   
## 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:0.000000   
## Median :0.0000 Median :1.0000 Median :0.000000   
## Mean :0.1516 Mean :0.8466 Mean :0.001863   
## 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:0.000000   
## Max. :1.0000 Max. :1.0000 Max. :1.000000

Smoke - Adults who are current smokers? No - 1, Yes - 2, not sure/Refused - 9, blanks

data$Smoke\_yes <- ifelse(data$Smoke == 1, 1, 0)   
data$Smoke\_no <- ifelse(data$Smoke == 2, 1, 0)   
data$Unkn\_Smoke <- ifelse(data$Smoke == 9, 1, 0)  
  
summary(data[38:40])

## Smoke\_yes Smoke\_no Unkn\_Smoke   
## Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:0.00000   
## Median :1.0000 Median :0.0000 Median :0.00000   
## Mean :0.8289 Mean :0.1283 Mean :0.04279   
## 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:0.00000   
## Max. :1.0000 Max. :1.0000 Max. :1.00000

Age - 18-80 with all values over 80 collapsed to 80.  
For the age category, I breakdown each group in to 10 year groupings from 30 to 70. Note that the wieght factors in BRFSS are established to maintain the population profile for age vs US census data.

data$age\_30minus <- ifelse(data$Age < 30, 1, 0)   
data$age\_30\_40 <- ifelse(data$Age < 30, 0, ifelse(data$Age >= 40, 0, 1))   
data$age\_40\_50 <- ifelse(data$Age < 40, 0, ifelse(data$Age >= 50, 0, 1))   
data$age\_50\_60 <- ifelse(data$Age < 50, 0, ifelse(data$Age >= 60, 0, 1))   
data$age\_60\_70 <- ifelse(data$Age < 60, 0, ifelse(data$Age >= 70, 0, 1))   
data$age\_70plus <- ifelse(data$Age < 70, 0, 1)   
summary(data[41:46])

## age\_30minus age\_30\_40 age\_40\_50 age\_50\_60   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.0000 Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.1164 Mean :0.1204 Mean :0.1321 Mean :0.1847   
## 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## age\_60\_70 age\_70plus   
## Min. :0.0000 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:0.000   
## Median :0.0000 Median :0.000   
## Mean :0.2184 Mean :0.228   
## 3rd Qu.:0.0000 3rd Qu.:0.000   
## Max. :1.0000 Max. :1.000

I will aggregate the data as a proportion of population for each value by county (MMSA Region).

#Each observation dummy value times it's weight factor.  
MMSA\_Health <- data[14:46]\*data$WT\_Factor  
#Keep MMSA by each observation  
MMSA\_Health$MMSA\_Name <- data$MMSA\_Number  
#Create total wt\_factor for each MMSA  
MMSAWT <- aggregate(data$WT\_Factor~data$MMSA\_Number, FUN = sum)  
names(MMSAWT) <- c("MMSA\_NAME", "WT\_Factor")  
#Aggregate the health indicators by MMSA  
MMSAHealth <- aggregate(.~MMSA\_Name, data = MMSA\_Health, FUN = sum)  
#Divide out each of the total wt factors.  
MMSAHealth[2:34] <- MMSAHealth[2:34]/MMSAWT$WT\_Factor  
summary(MMSAHealth)

## MMSA\_Name Gen\_Hlth\_good Gen\_Hlth\_bad Gen\_Hlth\_unkn   
## Min. :10100 Min. :0.6433 Min. :0.08294 Min. :0.0000000   
## 1st Qu.:19316 1st Qu.:0.8015 1st Qu.:0.14265 1st Qu.:0.0005339   
## Median :31110 Median :0.8263 Median :0.17220 Median :0.0016162   
## Mean :30108 Mean :0.8239 Mean :0.17385 Mean :0.0022506   
## 3rd Qu.:40080 3rd Qu.:0.8558 3rd Qu.:0.19548 3rd Qu.:0.0030704   
## Max. :49340 Max. :0.9159 Max. :0.35126 Max. :0.0164154   
## Insured Uninsured Unkn\_insurance BP\_High   
## Min. :0.7024 Min. :0.04156 Min. :0.000000 Min. :0.2161   
## 1st Qu.:0.8536 1st Qu.:0.07760 1st Qu.:0.002799 1st Qu.:0.3050   
## Median :0.8888 Median :0.10289 Median :0.004539 Median :0.3369   
## Mean :0.8820 Mean :0.11261 Mean :0.005410 Mean :0.3412   
## 3rd Qu.:0.9175 3rd Qu.:0.14191 3rd Qu.:0.007163 3rd Qu.:0.3701   
## Max. :0.9536 Max. :0.29517 Max. :0.023744 Max. :0.4744   
## BP\_normal BP\_unkn Chol\_High Chol\_normal   
## Min. :0.5223 Min. :0.000000 Min. :0.1560 Min. :0.4861   
## 1st Qu.:0.6272 1st Qu.:0.001023 1st Qu.:0.2604 1st Qu.:0.6036   
## Median :0.6607 Median :0.002564 Median :0.2804 Median :0.6187   
## Mean :0.6559 Mean :0.002856 Mean :0.2826 Mean :0.6191   
## 3rd Qu.:0.6918 3rd Qu.:0.004204 3rd Qu.:0.3066 3rd Qu.:0.6380   
## Max. :0.7807 Max. :0.009679 Max. :0.3887 Max. :0.7081   
## Unkn\_chol HD\_yes HD\_no Unkn\_HD   
## Min. :0.05095 Min. :0.01498 Min. :0.8534 Min. :0.000000   
## 1st Qu.:0.07861 1st Qu.:0.03162 1st Qu.:0.9430 1st Qu.:0.004265   
## Median :0.09264 Median :0.04006 Median :0.9530 Median :0.006223   
## Mean :0.09828 Mean :0.04090 Mean :0.9520 Mean :0.007084   
## 3rd Qu.:0.11174 3rd Qu.:0.04891 3rd Qu.:0.9622 3rd Qu.:0.008417   
## Max. :0.20129 Max. :0.10576 Max. :0.9805 Max. :0.040834   
## cancer\_yes cancer\_no Unkn\_cancer KD\_yes   
## Min. :0.02447 Min. :0.8830 Min. :0.0000000 Min. :0.01502   
## 1st Qu.:0.06320 1st Qu.:0.9196 1st Qu.:0.0005683 1st Qu.:0.02389   
## Median :0.06960 Median :0.9280 Median :0.0017418 Median :0.02938   
## Mean :0.07010 Mean :0.9277 Mean :0.0021513 Mean :0.03012   
## 3rd Qu.:0.07778 3rd Qu.:0.9357 3rd Qu.:0.0032246 3rd Qu.:0.03467   
## Max. :0.11598 Max. :0.9755 Max. :0.0072558 Max. :0.05434   
## KD\_no Unkn\_KD Diabetes\_yes Diabetes\_no   
## Min. :0.9391 Min. :0.0000000 Min. :0.07741 Min. :0.7728   
## 1st Qu.:0.9618 1st Qu.:0.0009183 1st Qu.:0.11173 1st Qu.:0.8536   
## Median :0.9685 Median :0.0021697 Median :0.12908 Median :0.8695   
## Mean :0.9673 Mean :0.0025822 Mean :0.13072 Mean :0.8673   
## 3rd Qu.:0.9735 3rd Qu.:0.0037011 3rd Qu.:0.14345 3rd Qu.:0.8877   
## Max. :0.9844 Max. :0.0098713 Max. :0.22705 Max. :0.9218   
## Unkn\_Diabetes Smoke\_yes Smoke\_no Unkn\_Smoke   
## Min. :0.0000000 Min. :0.6527 Min. :0.04079 Min. :0.008302   
## 1st Qu.:0.0001036 1st Qu.:0.7691 1st Qu.:0.13178 1st Qu.:0.034957   
## Median :0.0010573 Median :0.7946 Median :0.16014 Median :0.043693   
## Mean :0.0020291 Mean :0.7944 Mean :0.16034 Mean :0.045308   
## 3rd Qu.:0.0029759 3rd Qu.:0.8209 3rd Qu.:0.18855 3rd Qu.:0.054639   
## Max. :0.0285909 Max. :0.9314 Max. :0.30243 Max. :0.089837   
## age\_30minus age\_30\_40 age\_40\_50 age\_50\_60   
## Min. :0.1411 Min. :0.09025 Min. :0.09347 Min. :0.1055   
## 1st Qu.:0.1897 1st Qu.:0.15901 1st Qu.:0.14030 1st Qu.:0.1598   
## Median :0.2051 Median :0.17671 Median :0.15483 Median :0.1694   
## Mean :0.2109 Mean :0.17569 Mean :0.15420 Mean :0.1693   
## 3rd Qu.:0.2205 3rd Qu.:0.19557 3rd Qu.:0.16657 3rd Qu.:0.1812   
## Max. :0.4388 Max. :0.23855 Max. :0.20856 Max. :0.2108   
## age\_60\_70 age\_70plus   
## Min. :0.08162 Min. :0.08167   
## 1st Qu.:0.14177 1st Qu.:0.11917   
## Median :0.15362 Median :0.13381   
## Mean :0.15374 Mean :0.13616   
## 3rd Qu.:0.16697 3rd Qu.:0.14608   
## Max. :0.22131 Max. :0.28016

Assuming that I did that all correctly, then the sum of the individual category columns for each MMSA should be 1 (i.e. 100%). This was a good review, because I found that I missed some values (3 and 4) in the Diabetes categories.

MMSA <- "10100"  
  
paste("General Health has", sum(MMSAHealth[2:4][MMSAHealth$MMSA\_Name == MMSA,])\*100, "% of data.")

## [1] "General Health has 100 % of data."

paste("Health coverage has", sum(MMSAHealth[5:7][MMSAHealth$MMSA\_Name == MMSA,])\*100, "% of data.")

## [1] "Health coverage has 100 % of data."

paste("Blood Pressure has", sum(MMSAHealth[8:10][MMSAHealth$MMSA\_Name == MMSA,])\*100, "% of data.")

## [1] "Blood Pressure has 100 % of data."

paste("Cholestrol has", sum(MMSAHealth[11:13][MMSAHealth$MMSA\_Name == MMSA,])\*100, "% of data.")

## [1] "Cholestrol has 100 % of data."

paste("Heart Disease has", sum(MMSAHealth[14:16][MMSAHealth$MMSA\_Name == MMSA,])\*100, "% of data.")

## [1] "Heart Disease has 100 % of data."

paste("Cancer has", sum(MMSAHealth[17:19][MMSAHealth$MMSA\_Name == MMSA,])\*100, "% of data.")

## [1] "Cancer has 100 % of data."

paste("Kidney Disease has", sum(MMSAHealth[20:22][MMSAHealth$MMSA\_Name == MMSA,])\*100, "% of data.")

## [1] "Kidney Disease has 100 % of data."

paste("Diabetes has", sum(MMSAHealth[23:25][MMSAHealth$MMSA\_Name == MMSA,])\*100, "% of data.")

## [1] "Diabetes has 100 % of data."

paste("Smoke has", sum(MMSAHealth[26:28][MMSAHealth$MMSA\_Name == MMSA,])\*100, "% of data.")

## [1] "Smoke has 100 % of data."

paste("Age has", sum(MMSAHealth[29:34][MMSAHealth$MMSA\_Name == MMSA,])\*100, "% of data.")

## [1] "Age has 100 % of data."

I reload the body mass index data and merge it with the other health indicators by MMSA number.

BMIdata <- read.csv('healthdata3.csv')  
MMSAHealth <- merge(x = BMIdata, y = MMSAHealth, by.x ="MMSA", by.y = "MMSA\_Name")  
summary(MMSAHealth)

## MMSA BMI\_.kg.m2. underweight normalweight   
## Min. :10100 Min. :25.03 Min. :0.00000 Min. :0.1854   
## 1st Qu.:19316 1st Qu.:26.75 1st Qu.:0.01315 1st Qu.:0.3291   
## Median :31110 Median :27.46 Median :0.02446 Median :0.3887   
## Mean :30108 Mean :27.55 Mean :0.02598 Mean :0.3798   
## 3rd Qu.:40080 3rd Qu.:28.27 3rd Qu.:0.03249 3rd Qu.:0.4172   
## Max. :49340 Max. :30.29 Max. :0.12482 Max. :0.5624   
## overweight obese extremeobesity Gen\_Hlth\_good   
## Min. :0.2074 Min. :0.05835 Min. :0.006852 Min. :0.6433   
## 1st Qu.:0.3004 1st Qu.:0.16969 1st Qu.:0.036991 1st Qu.:0.8015   
## Median :0.3297 Median :0.20359 Median :0.054013 Median :0.8263   
## Mean :0.3330 Mean :0.20401 Mean :0.057282 Mean :0.8239   
## 3rd Qu.:0.3655 3rd Qu.:0.23704 3rd Qu.:0.073672 3rd Qu.:0.8558   
## Max. :0.4705 Max. :0.33312 Max. :0.190972 Max. :0.9159   
## Gen\_Hlth\_bad Gen\_Hlth\_unkn Insured Uninsured   
## Min. :0.08294 Min. :0.0000000 Min. :0.7024 Min. :0.04156   
## 1st Qu.:0.14265 1st Qu.:0.0005339 1st Qu.:0.8536 1st Qu.:0.07760   
## Median :0.17220 Median :0.0016162 Median :0.8888 Median :0.10289   
## Mean :0.17385 Mean :0.0022506 Mean :0.8820 Mean :0.11261   
## 3rd Qu.:0.19548 3rd Qu.:0.0030704 3rd Qu.:0.9175 3rd Qu.:0.14191   
## Max. :0.35126 Max. :0.0164154 Max. :0.9536 Max. :0.29517   
## Unkn\_insurance BP\_High BP\_normal BP\_unkn   
## Min. :0.000000 Min. :0.2161 Min. :0.5223 Min. :0.000000   
## 1st Qu.:0.002799 1st Qu.:0.3050 1st Qu.:0.6272 1st Qu.:0.001023   
## Median :0.004539 Median :0.3369 Median :0.6607 Median :0.002564   
## Mean :0.005410 Mean :0.3412 Mean :0.6559 Mean :0.002856   
## 3rd Qu.:0.007163 3rd Qu.:0.3701 3rd Qu.:0.6918 3rd Qu.:0.004204   
## Max. :0.023744 Max. :0.4744 Max. :0.7807 Max. :0.009679   
## Chol\_High Chol\_normal Unkn\_chol HD\_yes   
## Min. :0.1560 Min. :0.4861 Min. :0.05095 Min. :0.01498   
## 1st Qu.:0.2604 1st Qu.:0.6036 1st Qu.:0.07861 1st Qu.:0.03162   
## Median :0.2804 Median :0.6187 Median :0.09264 Median :0.04006   
## Mean :0.2826 Mean :0.6191 Mean :0.09828 Mean :0.04090   
## 3rd Qu.:0.3066 3rd Qu.:0.6380 3rd Qu.:0.11174 3rd Qu.:0.04891   
## Max. :0.3887 Max. :0.7081 Max. :0.20129 Max. :0.10576   
## HD\_no Unkn\_HD cancer\_yes cancer\_no   
## Min. :0.8534 Min. :0.000000 Min. :0.02447 Min. :0.8830   
## 1st Qu.:0.9430 1st Qu.:0.004265 1st Qu.:0.06320 1st Qu.:0.9196   
## Median :0.9530 Median :0.006223 Median :0.06960 Median :0.9280   
## Mean :0.9520 Mean :0.007084 Mean :0.07010 Mean :0.9277   
## 3rd Qu.:0.9622 3rd Qu.:0.008417 3rd Qu.:0.07778 3rd Qu.:0.9357   
## Max. :0.9805 Max. :0.040834 Max. :0.11598 Max. :0.9755   
## Unkn\_cancer KD\_yes KD\_no Unkn\_KD   
## Min. :0.0000000 Min. :0.01502 Min. :0.9391 Min. :0.0000000   
## 1st Qu.:0.0005683 1st Qu.:0.02389 1st Qu.:0.9618 1st Qu.:0.0009183   
## Median :0.0017418 Median :0.02938 Median :0.9685 Median :0.0021697   
## Mean :0.0021513 Mean :0.03012 Mean :0.9673 Mean :0.0025822   
## 3rd Qu.:0.0032246 3rd Qu.:0.03467 3rd Qu.:0.9735 3rd Qu.:0.0037011   
## Max. :0.0072558 Max. :0.05434 Max. :0.9844 Max. :0.0098713   
## Diabetes\_yes Diabetes\_no Unkn\_Diabetes Smoke\_yes   
## Min. :0.07741 Min. :0.7728 Min. :0.0000000 Min. :0.6527   
## 1st Qu.:0.11173 1st Qu.:0.8536 1st Qu.:0.0001036 1st Qu.:0.7691   
## Median :0.12908 Median :0.8695 Median :0.0010573 Median :0.7946   
## Mean :0.13072 Mean :0.8673 Mean :0.0020291 Mean :0.7944   
## 3rd Qu.:0.14345 3rd Qu.:0.8877 3rd Qu.:0.0029759 3rd Qu.:0.8209   
## Max. :0.22705 Max. :0.9218 Max. :0.0285909 Max. :0.9314   
## Smoke\_no Unkn\_Smoke age\_30minus age\_30\_40   
## Min. :0.04079 Min. :0.008302 Min. :0.1411 Min. :0.09025   
## 1st Qu.:0.13178 1st Qu.:0.034957 1st Qu.:0.1897 1st Qu.:0.15901   
## Median :0.16014 Median :0.043693 Median :0.2051 Median :0.17671   
## Mean :0.16034 Mean :0.045308 Mean :0.2109 Mean :0.17569   
## 3rd Qu.:0.18855 3rd Qu.:0.054639 3rd Qu.:0.2205 3rd Qu.:0.19557   
## Max. :0.30243 Max. :0.089837 Max. :0.4388 Max. :0.23855   
## age\_40\_50 age\_50\_60 age\_60\_70 age\_70plus   
## Min. :0.09347 Min. :0.1055 Min. :0.08162 Min. :0.08167   
## 1st Qu.:0.14030 1st Qu.:0.1598 1st Qu.:0.14177 1st Qu.:0.11917   
## Median :0.15483 Median :0.1694 Median :0.15362 Median :0.13381   
## Mean :0.15420 Mean :0.1693 Mean :0.15374 Mean :0.13616   
## 3rd Qu.:0.16657 3rd Qu.:0.1812 3rd Qu.:0.16697 3rd Qu.:0.14608   
## Max. :0.20856 Max. :0.2108 Max. :0.22131 Max. :0.28016

I reload the covid data. I created a MMSA number vs name listing and merge that with the health data so that I can merge with the name listing of the covid data.

coviddata <- read.csv('covid\_MMSA.csv')  
MMSA <- subset(data, select = c("MMSA\_NAME","MMSA\_Number"))  
#summary(MMSA)  
MMSA\_Health <- merge(x = MMSA[!duplicated(MMSA$MMSA\_Number),], y = MMSAHealth, by.x ="MMSA\_Number", by.y = "MMSA")  
MMSA\_Health <- merge(x = coviddata, y = MMSA\_Health, by ="MMSA\_NAME")  
summary(MMSA\_Health)

## MMSA\_NAME CASES DEATHS   
## Aberdeen, SD : 1 Min. : 391 Min. : 2.00   
## Albany-Schenectady-Troy, NY : 1 1st Qu.: 6086 1st Qu.: 84.75   
## Albuquerque, NM : 1 Median : 11359 Median : 237.50   
## Allentown-Bethlehem-Easton, PA-NJ: 1 Mean : 26230 Mean : 610.47   
## Anchorage, AK : 1 3rd Qu.: 26707 3rd Qu.: 619.75   
## Augusta-Richmond County, GA-SC : 1 Max. :324050 Max. :7844.00   
## (Other) :98   
## death\_rate MMSA\_Number BMI\_.kg.m2. underweight   
## Min. :0.299 Min. :10100 Min. :25.22 Min. :0.00000   
## 1st Qu.:1.320 1st Qu.:19465 1st Qu.:26.79 1st Qu.:0.01315   
## Median :1.886 Median :31110 Median :27.54 Median :0.02546   
## Mean :2.170 Mean :30329 Mean :27.70 Mean :0.02631   
## 3rd Qu.:2.644 3rd Qu.:40190 3rd Qu.:28.55 3rd Qu.:0.03235   
## Max. :9.556 Max. :49340 Max. :30.29 Max. :0.12482   
##   
## normalweight overweight obese extremeobesity   
## Min. :0.1854 Min. :0.2074 Min. :0.05835 Min. :0.006852   
## 1st Qu.:0.3230 1st Qu.:0.3004 1st Qu.:0.18049 1st Qu.:0.038419   
## Median :0.3810 Median :0.3274 Median :0.21264 Median :0.054013   
## Mean :0.3698 Mean :0.3327 Mean :0.21176 Mean :0.059356   
## 3rd Qu.:0.4110 3rd Qu.:0.3636 3rd Qu.:0.24652 3rd Qu.:0.080298   
## Max. :0.5624 Max. :0.4705 Max. :0.33312 Max. :0.190972   
##   
## Gen\_Hlth\_good Gen\_Hlth\_bad Gen\_Hlth\_unkn Insured   
## Min. :0.6482 Min. :0.08294 Min. :0.0000000 Min. :0.7024   
## 1st Qu.:0.7997 1st Qu.:0.14366 1st Qu.:0.0005354 1st Qu.:0.8536   
## Median :0.8218 Median :0.17523 Median :0.0017286 Median :0.8846   
## Mean :0.8228 Mean :0.17490 Mean :0.0022890 Mean :0.8781   
## 3rd Qu.:0.8548 3rd Qu.:0.19704 3rd Qu.:0.0030322 3rd Qu.:0.9124   
## Max. :0.9159 Max. :0.34998 Max. :0.0164154 Max. :0.9463   
##   
## Uninsured Unkn\_insurance BP\_High BP\_normal   
## Min. :0.04627 Min. :0.0001147 Min. :0.2161 Min. :0.5223   
## 1st Qu.:0.08156 1st Qu.:0.0028803 1st Qu.:0.3096 1st Qu.:0.6199   
## Median :0.10886 Median :0.0046255 Median :0.3453 Median :0.6520   
## Mean :0.11632 Mean :0.0055736 Mean :0.3446 Mean :0.6527   
## 3rd Qu.:0.14258 3rd Qu.:0.0072089 3rd Qu.:0.3740 3rd Qu.:0.6884   
## Max. :0.29517 Max. :0.0237437 Max. :0.4744 Max. :0.7807   
##   
## BP\_unkn Chol\_High Chol\_normal Unkn\_chol   
## Min. :0.0000000 Min. :0.1560 Min. :0.4861 Min. :0.05095   
## 1st Qu.:0.0009135 1st Qu.:0.2546 1st Qu.:0.5989 1st Qu.:0.08353   
## Median :0.0020296 Median :0.2777 Median :0.6167 Median :0.09565   
## Mean :0.0026530 Mean :0.2815 Mean :0.6156 Mean :0.10289   
## 3rd Qu.:0.0035453 3rd Qu.:0.3066 3rd Qu.:0.6372 3rd Qu.:0.11476   
## Max. :0.0096794 Max. :0.3887 Max. :0.7081 Max. :0.20129   
##   
## HD\_yes HD\_no Unkn\_HD cancer\_yes   
## Min. :0.01498 Min. :0.8534 Min. :0.0005109 Min. :0.02447   
## 1st Qu.:0.03193 1st Qu.:0.9430 1st Qu.:0.0043027 1st Qu.:0.06368   
## Median :0.04077 Median :0.9527 Median :0.0062942 Median :0.07013   
## Mean :0.04169 Mean :0.9509 Mean :0.0074341 Mean :0.07112   
## 3rd Qu.:0.04891 3rd Qu.:0.9614 3rd Qu.:0.0088037 3rd Qu.:0.07905   
## Max. :0.10576 Max. :0.9780 Max. :0.0408336 Max. :0.11598   
##   
## cancer\_no Unkn\_cancer KD\_yes KD\_no   
## Min. :0.8830 Min. :0.0000000 Min. :0.01636 Min. :0.9391   
## 1st Qu.:0.9193 1st Qu.:0.0005076 1st Qu.:0.02438 1st Qu.:0.9613   
## Median :0.9262 Median :0.0017269 Median :0.02910 Median :0.9688   
## Mean :0.9269 Mean :0.0019458 Mean :0.03044 Mean :0.9669   
## 3rd Qu.:0.9352 3rd Qu.:0.0029150 3rd Qu.:0.03496 3rd Qu.:0.9733   
## Max. :0.9755 Max. :0.0072333 Max. :0.05434 Max. :0.9832   
##   
## Unkn\_KD Diabetes\_yes Diabetes\_no Unkn\_Diabetes   
## Min. :0.0000000 Min. :0.07741 Min. :0.7728 Min. :0.000e+00   
## 1st Qu.:0.0009514 1st Qu.:0.10942 1st Qu.:0.8519 1st Qu.:1.771e-05   
## Median :0.0022113 Median :0.13070 Median :0.8682 Median :8.679e-04   
## Mean :0.0026409 Mean :0.13132 Mean :0.8670 Mean :1.662e-03   
## 3rd Qu.:0.0038747 3rd Qu.:0.14742 3rd Qu.:0.8891 3rd Qu.:2.725e-03   
## Max. :0.0098713 Max. :0.22705 Max. :0.9218 Max. :1.337e-02   
##   
## Smoke\_yes Smoke\_no Unkn\_Smoke age\_30minus   
## Min. :0.6897 Min. :0.04079 Min. :0.009265 Min. :0.1411   
## 1st Qu.:0.7619 1st Qu.:0.14209 1st Qu.:0.034754 1st Qu.:0.1886   
## Median :0.7876 Median :0.16952 Median :0.042974 Median :0.2056   
## Mean :0.7877 Mean :0.16711 Mean :0.045221 Mean :0.2133   
## 3rd Qu.:0.8118 3rd Qu.:0.19261 3rd Qu.:0.054906 3rd Qu.:0.2242   
## Max. :0.9314 Max. :0.25851 Max. :0.089837 Max. :0.4388   
##   
## age\_30\_40 age\_40\_50 age\_50\_60 age\_60\_70   
## Min. :0.09025 Min. :0.09347 Min. :0.1055 Min. :0.08162   
## 1st Qu.:0.15994 1st Qu.:0.13860 1st Qu.:0.1584 1st Qu.:0.14330   
## Median :0.17814 Median :0.15336 Median :0.1684 Median :0.15580   
## Mean :0.17680 Mean :0.15273 Mean :0.1670 Mean :0.15455   
## 3rd Qu.:0.19655 3rd Qu.:0.16611 3rd Qu.:0.1800 3rd Qu.:0.16802   
## Max. :0.23855 Max. :0.20027 Max. :0.2108 Max. :0.21279   
##   
## age\_70plus   
## Min. :0.08167   
## 1st Qu.:0.11917   
## Median :0.13140   
## Mean :0.13567   
## 3rd Qu.:0.14505   
## Max. :0.28016   
##

This is the complete data set for analysis and modeling with and without covid data.

write.csv(MMSA\_Health, 'healthcoviddata.csv', row.names = FALSE)  
write.csv(MMSAHealth, 'healthindicatordata.csv', row.names = FALSE)

References:

Unknown. (2017) Smart: City and county survey data: 2017 Data. Center for Disease Control and Prevention: Behavioral risk factor surveillance system. Retrieved November 1, 2020 from: <https://www.cdc.gov/brfss/smart/Smart_data.htm>

Unknown. (2020b). Coronavirus disease 2019 (COVID-19): People with certain medical conditions. Centers for Disease Control and Prevention. Retrieved from: <https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fneed-extra-precautions%2Fgroups-at-higher-risk.html>