BRFSS 2013-2017 Survey EDA-BMI

Amanda Kimball

11/19/2020

The Smart: City and county survey data put together by the Center for Disease Control and Prevention: Behavioral risk factor surveillance system (Smart data, 2013-2017). 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. I initially uploaded the data and pulled out the data related to body mass index and age demographics. I evaluated using 5 years of data in order to increase the number of regions in my analysis. I ended up using only the 2017 analysis becuase adding the 5 years removes important variables high blood pressure and high cholesterol but am saving the analysis for future reference.

library(foreign)  
BRFSS2017 <- read.xport("MMSA2017.XPT")  
BRFSS2016 <- read.xport("MMSA2016.XPT")  
BRFSS2015 <- read.xport("MMSA2015.XPT")  
BRFSS2014 <- read.xport("MMSA2014.XPT")  
BRFSS2013 <- read.xport("MMSA2013.XPT")

I identified the variables needed for the analysis as indicated ‘vars’ below. Then subset each of the imported files by those factors.

#data <- subset(BRFSS2017, select = c(47:49, 122:129, 174, 175, 177))  
  
vars <- c("PREGNANT", "X\_BMI5", "X\_BMI5CAT", "X\_RFBMI5", "X\_AGE80", "X\_MMSA", "X\_MMSAWT", "MMSANAME")   
  
data17 <- subset(BRFSS2017, select = vars)  
data16 <- subset(BRFSS2016, select = vars)  
data15 <- subset(BRFSS2015, select = vars)  
data14 <- subset(BRFSS2014, select = vars)  
data13 <- subset(BRFSS2013, select = vars)  
  
data <- rbind(data13, data14, data15, data16, data17)  
  
#data$County <- gsub(", Metropolitan Statistical Area", "", data$County)  
#data$County <- gsub(", Micropolitan Statistical Area", "", data$County)  
#data$County <- gsub(", Metropolitan Division", "", data$County)  
#data$State <- gsub(".\*,", "", data$County)  
#data$State <- gsub(" ", "", data$State)  
#data$County <- gsub(",.\*", "", data$County)  
str(data)

## 'data.frame': 1208648 obs. of 8 variables:  
## $ PREGNANT : num 2 NA NA NA NA NA NA NA NA NA ...  
## $ X\_BMI5 : num 2103 3023 3212 3067 2441 ...  
## $ X\_BMI5CAT: num 2 4 4 4 2 3 2 2 2 3 ...  
## $ X\_RFBMI5 : num 1 2 2 2 1 2 1 1 1 2 ...  
## $ X\_AGE80 : num 40 75 63 52 70 38 53 49 74 66 ...  
## $ X\_MMSA : num 10380 10380 10380 10380 10380 ...  
## $ X\_MMSAWT : num 294.3 89.2 251.1 448.6 60.2 ...  
## $ MMSANAME : Factor w/ 170 levels "Aguadilla-Isabela, PR, Metropolitan Statistical Area",..: 1 1 1 1 1 1 1 1 1 1 ...

Here are the counties. I use this listing to compare this data to other data sources for specifics on MMSA name alignment.

I am focusing on adults 18 and older for the analysis and will eliminate all younger participants which are not included in the BRFSS data, so I remove the younger participants.

data <- data[(data$X\_AGE80 >= 18),]  
summary(data$X\_AGE80)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 42.00 57.00 54.45 68.00 80.00

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

summary(data)

## PREGNANT X\_BMI5 X\_BMI5CAT X\_RFBMI5   
## Min. :1 Min. :1202 Min. :1.00 Min. :1.000   
## 1st Qu.:2 1st Qu.:2367 1st Qu.:2.00 1st Qu.:1.000   
## Median :2 Median :2663 Median :3.00 Median :2.000   
## Mean :2 Mean :2781 Mean :2.92 Mean :2.207   
## 3rd Qu.:2 3rd Qu.:3072 3rd Qu.:4.00 3rd Qu.:2.000   
## Max. :9 Max. :9960 Max. :4.00 Max. :9.000   
## NA's :1010619 NA's :91861 NA's :91861   
## X\_AGE80 X\_MMSA X\_MMSAWT   
## Min. :18.00 Min. :10100 Min. : 0.15   
## 1st Qu.:42.00 1st Qu.:19740 1st Qu.: 111.65   
## Median :57.00 Median :34980 Median : 274.50   
## Mean :54.45 Mean :31606 Mean : 648.26   
## 3rd Qu.:68.00 3rd Qu.:40380 3rd Qu.: 690.78   
## Max. :80.00 Max. :49660 Max. :43227.53   
##   
## MMSANAME   
## Minneapolis-St. Paul-Bloomington, MN-WI, Metropolitan Statistical Area: 43290   
## Washington-Arlington-Alexandria, DC-VA-MD-WV, Metropolitan Division : 43019   
## New York-Jersey City-White Plains, NY-NJ, Metropolitan Division : 41963   
## Providence-Warwick, RI-MA, Metropolitan Statistical Area : 35906   
## Kansas City, MO-KS, Metropolitan Statistical Area : 32192   
## Phoenix-Mesa-Scottsdale, AZ, Metropolitan Statistical Area : 30456   
## (Other) :981822

There are a significant number of NA’s for the BMI index. I am going to focus on one county’s data to understand the nuances in the data set and then apply what I have learned to the other counties. Albuquerque is in all five data sets. I can safely remove the individuals listed as pregnant as outliers with a cause. It appears that the remaining values are missing the weight or height factor either listed as 9999 (Refused) or 7777 (Not sure).

Albuquerque <- data[data$X\_MMSA == 10740,]  
summary(Albuquerque)

## PREGNANT X\_BMI5 X\_BMI5CAT X\_RFBMI5 X\_AGE80   
## Min. :1.000 Min. :1225 Min. :1.000 Min. :1.0 Min. :18.00   
## 1st Qu.:2.000 1st Qu.:2348 1st Qu.:2.000 1st Qu.:1.0 1st Qu.:41.00   
## Median :2.000 Median :2650 Median :3.000 Median :2.0 Median :56.00   
## Mean :2.012 Mean :2735 Mean :2.867 Mean :2.1 Mean :53.96   
## 3rd Qu.:2.000 3rd Qu.:3011 3rd Qu.:4.000 3rd Qu.:2.0 3rd Qu.:67.00   
## Max. :9.000 Max. :8503 Max. :4.000 Max. :9.0 Max. :80.00   
## NA's :6910 NA's :530 NA's :530   
## X\_MMSA X\_MMSAWT   
## Min. :10740 Min. : 3.099   
## 1st Qu.:10740 1st Qu.: 173.541   
## Median :10740 Median : 321.368   
## Mean :10740 Mean : 420.603   
## 3rd Qu.:10740 3rd Qu.: 553.970   
## Max. :10740 Max. :4069.332   
##   
## MMSANAME   
## Albuquerque, NM, Metropolitan Statistical Area :8290   
## Aguadilla-Isabela, PR, Metropolitan Statistical Area : 0   
## Akron, OH, Metropolitan Statistical Area : 0   
## Allentown-Bethlehem-Easton, PA-NJ, Metropolitan Statistical Area: 0   
## Anchorage, AK, Metropolitan Statistical Area : 0   
## Atlanta-Sandy Springs-Roswell, GA, Metropolitan Statistical Area: 0   
## (Other) : 0

I calculated the average for the Albuquerque and compared it with the average for the entire data sample and find that they are within 0.2 kg/m2 difference. The weight factor ‘X\_MMSAWT’ is used to adjust for sampling methodology. Note that the documentation indicates a 2 decimal place value in the 4 digit value, so i divide by 100. Based off this analysis, I am going to impute the average for the population as the NAs.

Albuquerqueavg <- mean((Albuquerque$X\_BMI5\*Albuquerque$X\_MMSAWT),na.rm=TRUE)/  
 mean(Albuquerque$X\_MMSAWT,na.rm=TRUE)  
popavg <- mean((data$X\_BMI5\*data$X\_MMSAWT), na.rm=TRUE)/mean(data$X\_MMSAWT,na.rm=TRUE)  
paste("Average BMI in population is", round(popavg/100, 1), "and average BMI for Albuquerque is",  
 round(Albuquerqueavg/100, 1), ".")

## [1] "Average BMI in population is 27.4 and average BMI for Albuquerque is 27.2 ."

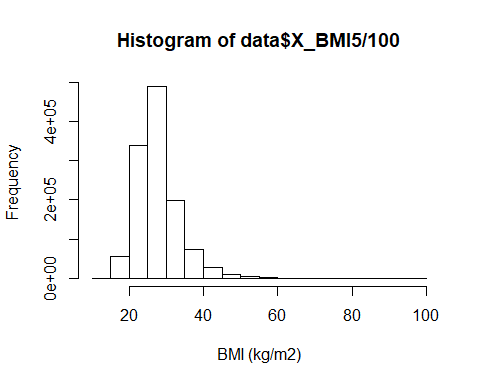
As indicated during the Aberdeen review - pregnant individuals can be removed as outliers outside the range of this analysis. The NAs in that column are the male subjects and will be included in the analysis. The average BMI will be added otherwise. The BMI average is 27.9 so the BMI5CAT variable is over weight = 3 and the RFBMI5 is over 25 = 2.

data <- subset(data, (PREGNANT != 1) | is.na(PREGNANT))  
index <- is.na(data$X\_BMI5)  
data$X\_BMI5[index] <- popavg  
data$X\_BMI5CAT[index] <- 3  
data$X\_RFBMI5[index] <- 2  
summary(data)

## PREGNANT X\_BMI5 X\_BMI5CAT X\_RFBMI5   
## Min. :2 Min. :1202 Min. :1.000 Min. :1.000   
## 1st Qu.:2 1st Qu.:2389 1st Qu.:2.000 1st Qu.:1.000   
## Median :2 Median :2734 Median :3.000 Median :2.000   
## Mean :2 Mean :2778 Mean :2.924 Mean :1.673   
## 3rd Qu.:2 3rd Qu.:3030 3rd Qu.:4.000 3rd Qu.:2.000   
## Max. :9 Max. :9960 Max. :4.000 Max. :2.000   
## NA's :1010619   
## X\_AGE80 X\_MMSA X\_MMSAWT   
## Min. :18.0 Min. :10100 Min. : 0.15   
## 1st Qu.:42.0 1st Qu.:19740 1st Qu.: 111.24   
## Median :57.0 Median :34980 Median : 273.50   
## Mean :54.6 Mean :31604 Mean : 645.85   
## 3rd Qu.:68.0 3rd Qu.:40380 3rd Qu.: 687.73   
## Max. :80.0 Max. :49660 Max. :43227.53   
##   
## MMSANAME   
## Minneapolis-St. Paul-Bloomington, MN-WI, Metropolitan Statistical Area: 42987   
## Washington-Arlington-Alexandria, DC-VA-MD-WV, Metropolitan Division : 42774   
## New York-Jersey City-White Plains, NY-NJ, Metropolitan Division : 41583   
## Providence-Warwick, RI-MA, Metropolitan Statistical Area : 35752   
## Kansas City, MO-KS, Metropolitan Statistical Area : 32005   
## Phoenix-Mesa-Scottsdale, AZ, Metropolitan Statistical Area : 30317   
## (Other) :976031

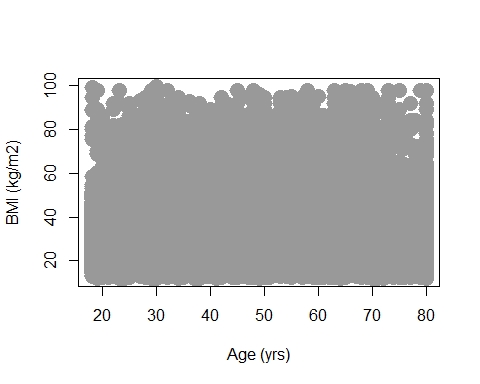
I plot a histogram of the data and find that most of the data sits in the 25-30 kg/m2 range. The ~90k that I added did not significantly change the data structure.

hist(data$X\_BMI5/100, xlab = "BMI (kg/m2)", breaks = 20)



The distribution of BMI does not correlate with age.

plot(data$X\_AGE80, data$X\_BMI5/100, ylab = "BMI (kg/m2)", xlab = "Age (yrs)", lwd = 10, col = "grey60")



Extreme Obesity is defined as over 40 kg/m2 but was not segregated in the original data set. I will add that level to the data for the visualizations and gives another level of detail towards poor health conditions.

data$X\_BMI5CAT[data$X\_BMI5 > 4000] <- 5  
summary(data)

## PREGNANT X\_BMI5 X\_BMI5CAT X\_RFBMI5   
## Min. :2 Min. :1202 Min. :1.000 Min. :1.000   
## 1st Qu.:2 1st Qu.:2389 1st Qu.:2.000 1st Qu.:1.000   
## Median :2 Median :2734 Median :3.000 Median :2.000   
## Mean :2 Mean :2778 Mean :2.963 Mean :1.673   
## 3rd Qu.:2 3rd Qu.:3030 3rd Qu.:4.000 3rd Qu.:2.000   
## Max. :9 Max. :9960 Max. :5.000 Max. :2.000   
## NA's :1010619   
## X\_AGE80 X\_MMSA X\_MMSAWT   
## Min. :18.0 Min. :10100 Min. : 0.15   
## 1st Qu.:42.0 1st Qu.:19740 1st Qu.: 111.24   
## Median :57.0 Median :34980 Median : 273.50   
## Mean :54.6 Mean :31604 Mean : 645.85   
## 3rd Qu.:68.0 3rd Qu.:40380 3rd Qu.: 687.73   
## Max. :80.0 Max. :49660 Max. :43227.53   
##   
## MMSANAME   
## Minneapolis-St. Paul-Bloomington, MN-WI, Metropolitan Statistical Area: 42987   
## Washington-Arlington-Alexandria, DC-VA-MD-WV, Metropolitan Division : 42774   
## New York-Jersey City-White Plains, NY-NJ, Metropolitan Division : 41583   
## Providence-Warwick, RI-MA, Metropolitan Statistical Area : 35752   
## Kansas City, MO-KS, Metropolitan Statistical Area : 32005   
## Phoenix-Mesa-Scottsdale, AZ, Metropolitan Statistical Area : 30317   
## (Other) :976031

I will aggregate the data as an average BMI by county (MMSA Region).

MMSAWT <- aggregate(data$X\_MMSAWT~data$X\_MMSA, FUN = mean)  
MMSABMI <- aggregate(data$X\_BMI5\*data$X\_MMSAWT/100~data$X\_MMSA, FUN = mean)  
names(MMSABMI) <- c("MMSA", "BMI\_(kg/m2)")  
MMSABMI$'BMI\_(kg/m2)' <- MMSABMI$'BMI\_(kg/m2)'/MMSAWT$`data$X\_MMSAWT`  
summary(MMSABMI)

## MMSA BMI\_(kg/m2)   
## Min. :10100 Min. :25.94   
## 1st Qu.:19450 1st Qu.:27.49   
## Median :30970 Median :27.93   
## Mean :30098 Mean :27.88   
## 3rd Qu.:40200 3rd Qu.:28.24   
## Max. :49660 Max. :29.61

Then I will create 4 proportion (percentage) values for portion of population underweight, Normal weight, Overweight, Obese, and extreme obesity by region.

data$underweight <- ifelse(data$X\_BMI5CAT == 1, 1, 0)  
data$normalweight <- ifelse(data$X\_BMI5CAT == 2, 1, 0)  
data$overweight <- ifelse(data$X\_BMI5CAT == 3, 1, 0)  
data$obese <- ifelse(data$X\_BMI5CAT == 4, 1, 0)  
data$extremeobesity <- ifelse(data$X\_BMI5CAT == 5, 1, 0)  
  
#Each observation dummy value times it's weight factor.  
MMSA\_BMI <- data[c("underweight", "normalweight", "overweight", "obese", "extremeobesity")]\*data$X\_MMSAWT  
#Keep MMSA by each observation  
MMSA\_BMI$MMSA\_Name <- data$X\_MMSA  
#Create total wt\_factor for each MMSA  
MMSAWT <- aggregate(data$X\_MMSAWT~data$X\_MMSA, FUN = sum)  
names(MMSAWT) <- c("MMSA\_NAME", "WT\_Factor")  
#Aggregate the health indicators by MMSA  
MMSA\_BMI <- aggregate(.~MMSA\_Name, data = MMSA\_BMI, FUN = sum)  
#Divide out each of the total wt factors.  
MMSABMI[3:7] <- MMSA\_BMI[2:6]/MMSAWT$WT\_Factor  
summary(MMSABMI)

## MMSA BMI\_(kg/m2) underweight normalweight   
## Min. :10100 Min. :25.94 Min. :0.007477 Min. :0.2342   
## 1st Qu.:19450 1st Qu.:27.49 1st Qu.:0.014803 1st Qu.:0.2892   
## Median :30970 Median :27.93 Median :0.016730 Median :0.3065   
## Mean :30098 Mean :27.88 Mean :0.017439 Mean :0.3095   
## 3rd Qu.:40200 3rd Qu.:28.24 3rd Qu.:0.019679 3rd Qu.:0.3285   
## Max. :49660 Max. :29.61 Max. :0.036068 Max. :0.4393   
## overweight obese extremeobesity   
## Min. :0.3384 Min. :0.1313 Min. :0.01793   
## 1st Qu.:0.3844 1st Qu.:0.2132 1st Qu.:0.03537   
## Median :0.3964 Median :0.2342 Median :0.04298   
## Mean :0.3963 Mean :0.2336 Mean :0.04324   
## 3rd Qu.:0.4094 3rd Qu.:0.2509 3rd Qu.:0.05107   
## Max. :0.4556 Max. :0.3593 Max. :0.07540

Check to see if they add up to 100% of values for 1 region: Albuquerque, NM was in all 5 years.

paste("BMI has", sum(MMSABMI[3:7][MMSABMI == 10740,])\*100, "% of data.")

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

I export this data so that I can create GEO Map data visualizations of the health indicators in Tableau.

write.csv(MMSABMI, 'healthdata4.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>