BRFSS 2017 Survey EDA-BMI

Amanda Kimball

11/2/2020

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. I initially uploaded the data and pulled out the data related to body mass index and age demographics. My study is focused on individuals age 18-64.

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.

data <- subset(BRFSS2017, select = c(47:49, 122:129, 177))  
data$County <- lapply(data$MMSANAME, as.character)  
  
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': 230875 obs. of 14 variables:  
## $ WEIGHT2 : num 200 140 300 175 160 250 170 155 160 119 ...  
## $ HEIGHT3 : num 507 505 605 504 504 511 502 504 504 506 ...  
## $ PREGNANT : num NA NA NA 2 NA NA NA NA NA NA ...  
## $ X\_AGEG5YR: num 9 10 1 2 13 3 10 13 13 13 ...  
## $ X\_AGE65YR: num 1 2 1 1 2 1 2 2 2 2 ...  
## $ X\_AGE80 : num 61 65 21 28 80 31 67 80 80 80 ...  
## $ X\_AGE\_G : num 5 6 1 2 6 2 6 6 6 6 ...  
## $ WTKG3 : num 9072 6350 13608 7938 7257 ...  
## $ X\_BMI5 : num 3132 2330 3557 3004 2746 ...  
## $ X\_BMI5CAT: num 4 2 4 4 3 4 4 3 3 2 ...  
## $ X\_RFBMI5 : num 2 1 2 2 2 2 2 2 2 1 ...  
## $ MMSANAME : Factor w/ 136 levels "Aberdeen, SD, Micropolitan Statistical Area",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ County : chr "Aberdeen" "Aberdeen" "Aberdeen" "Aberdeen" ...  
## $ State : chr "SD" "SD" "SD" "SD" ...

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

table(data$County)

##   
## Aberdeen   
## 512   
## Albany-Schenectady-Troy   
## 632   
## Albuquerque   
## 1654   
## Allentown-Bethlehem-Easton   
## 866   
## Anchorage   
## 1029   
## Atlanta-Sandy Springs-Roswell   
## 2944   
## Augusta-Richmond County   
## 851   
## Austin-Round Rock   
## 1632   
## Baltimore-Columbia-Towson   
## 5058   
## Baton Rouge   
## 782   
## Billings   
## 621   
## Birmingham-Hoover   
## 1291   
## Bismarck   
## 1113   
## Boise City   
## 1305   
## Boston   
## 1947   
## Buffalo-Cheektowaga-Niagara Falls   
## 517   
## Burlington-South Burlington   
## 1555   
## Cambridge-Newton-Framingham   
## 2349   
## Camden   
## 1633   
## Cedar Rapids   
## 592   
## Charleston   
## 816   
## Charleston-North Charleston   
## 1480   
## Charlotte-Concord-Gastonia   
## 1504   
## Chicago-Naperville-Elgin   
## 5122   
## Cincinnati   
## 1737   
## Claremont-Lebanon   
## 1763   
## Cleveland-Elyria   
## 1133   
## College Station-Bryan   
## 524   
## Colorado Springs   
## 1195   
## Columbia   
## 1229   
## Columbus   
## 2033   
## Corpus Christi   
## 722   
## Crestview-Fort Walton Beach-Destin   
## 582   
## Dallas-Plano-Irving   
## 1200   
## Dayton   
## 587   
## Deltona-Daytona Beach-Ormond Beach   
## 743   
## Denver-Aurora-Lakewood   
## 4677   
## Des Moines-West Des Moines   
## 1380   
## Duluth   
## 1075   
## El Paso   
## 527   
## Evansville   
## 626   
## Fargo   
## 1314   
## Fayetteville-Springdale-Rogers   
## 796   
## Florence   
## 525   
## Fort Wayne   
## 861   
## Fort Worth-Arlington   
## 565   
## Gainesville   
## 526   
## Grand Forks   
## 599   
## Grand Island   
## 789   
## Grand Rapids-Wyoming   
## 987   
## Greenville-Anderson-Mauldin   
## 1489   
## Hagerstown-Martinsburg   
## 781   
## Hartford-West Hartford-East Hartford   
## 3645   
## Hilton Head Island-Bluffton-Beaufort   
## 622   
## Houston-The Woodlands-Sugar Land   
## 2312   
## Huntington-Ashland   
## 1156   
## Indianapolis-Carmel-Anderson   
## 4214   
## Jackson   
## 535   
## Jacksonville   
## 1807   
## Kahului-Wailuku-Lahaina   
## 1366   
## Kansas City   
## 7757   
## Kingsport-Bristol-Bristol   
## 605   
## Knoxville   
## 589   
## Lansing-East Lansing   
## 510   
## Lexington-Fayette   
## 505   
## Lincoln   
## 1522   
## Little Rock-North Little Rock-Conway   
## 1218   
## Los Angeles-Long Beach-Anaheim   
## 2106   
## Louisville/Jefferson County   
## 2027   
## Manhattan   
## 848   
## Memphis   
## 959   
## Miami-Fort Lauderdale-West Palm Beach   
## 1407   
## Milwaukee-Waukesha-West Allis   
## 1115   
## Minneapolis-St. Paul-Bloomington   
## 8139   
## Minot   
## 654   
## Montgomery County-Bucks County-Chester County   
## 615   
## Myrtle Beach-Conway-North Myrtle Beach   
## 918   
## Nashville-Davidson--Murfreesboro--Franklin   
## 925   
## Nassau County-Suffolk County   
## 1232   
## New Orleans-Metairie   
## 886   
## New York-Jersey City-White Plains   
## 8759   
## Newark   
## 3701   
## North Platte   
## 563   
## North Port-Sarasota-Bradenton   
## 769   
## Oakland-Hayward-Berkeley   
## 696   
## Ogden-Clearfield   
## 1802   
## Oklahoma City   
## 2158   
## Omaha-Council Bluffs   
## 3333   
## Orlando-Kissimmee-Sanford   
## 1450   
## Panama City   
## 749   
## Pensacola-Ferry Pass-Brent   
## 689   
## Philadelphia   
## 988   
## Phoenix-Mesa-Scottsdale   
## 9165   
## Pittsburgh   
## 1425   
## Port St. Lucie   
## 666   
## Portland-South Portland   
## 2579   
## Portland-Vancouver-Hillsboro   
## 3140   
## Providence-Warwick   
## 6140   
## Provo-Orem   
## 1795   
## Rapid City   
## 1068   
## Reno   
## 1256   
## Richmond   
## 1530   
## Riverside-San Bernardino-Ontario   
## 1020   
## Rochester   
## 1231   
## Rockingham County-Strafford County   
## 1356   
## Sacramento--Roseville--Arden-Arcade   
## 932   
## Salina   
## 623   
## Salisbury   
## 2140   
## Salt Lake City   
## 3333   
## San Antonio-New Braunfels   
## 635   
## San Juan-Carolina-Caguas   
## 2815   
## Scottsbluff   
## 616   
## Seattle-Bellevue-Everett   
## 4189   
## Silver Spring-Frederick-Rockville   
## 2442   
## Sioux City   
## 1060   
## Sioux Falls   
## 1096   
## South Bend-Mishawaka   
## 568   
## Spartanburg   
## 510   
## Spokane-Spokane Valley   
## 1086   
## Springfield   
## 616   
## St. Cloud   
## 689   
## St. Louis   
## 2236   
## Tallahassee   
## 931   
## Tampa-St. Petersburg-Clearwater   
## 1559   
## Toledo   
## 766   
## Topeka   
## 1657   
## Tulsa   
## 1257   
## Tuscaloosa   
## 538   
## Virginia Beach-Norfolk-Newport News   
## 2100   
## Warren-Troy-Farmington Hills   
## 2542   
## Washington-Arlington-Alexandria   
## 8542   
## Wichita   
## 4054   
## Wichita Falls   
## 516   
## Wilmington   
## 2398   
## Worcester   
## 1437

I am focusing on adults between 18-64 for the analysis and will eliminate all younger and older participants, so I remove the older and younger participants.

data <- data[(data$X\_AGE65YR == 1),]  
summary(data$X\_AGE80)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 33.00 47.00 44.69 57.00 64.00

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

summary(data[9:11])

## X\_BMI5 X\_BMI5CAT X\_RFBMI5   
## Min. :1207 Min. :1.000 Min. :1.000   
## 1st Qu.:2371 1st Qu.:2.000 1st Qu.:1.000   
## Median :2712 Median :3.000 Median :2.000   
## Mean :2816 Mean :2.953 Mean :2.285   
## 3rd Qu.:3125 3rd Qu.:4.000 3rd Qu.:2.000   
## Max. :9931 Max. :4.000 Max. :9.000   
## NA's :12801 NA's :12801

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. Aberdeen has 25 NA values of 261 entries for the BMI indexes (~9% of the entries which is consistent with the population). There are 2 values listed as pregnant (rows 311 & 453). 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).

Aberdeen <- data[data$County == "Aberdeen",]

I calculated the average for the Aberdeen and compared it with the average for the data and find that they are 28.5 and 27.6 kg/m2. Note that the documentation indicates a 2 decimal place value. Based off this analysis, I am going to impute the average for the population as the NAs.

Aberdeenavg <- mean(Aberdeen$X\_BMI5,na.rm=TRUE)  
popavg <- mean(data$X\_BMI5, na.rm=TRUE)

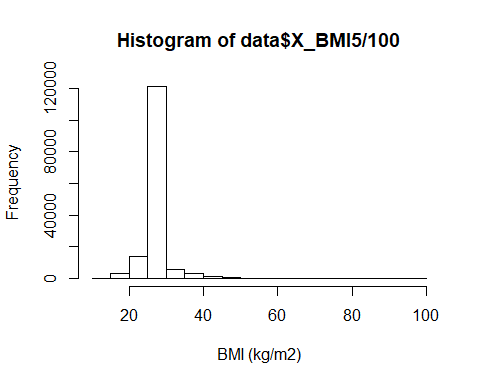
As indicated during the Aberdeen review - pregnant individuals can be removed as outliers outside the range of this analysis. The average BMI will be added otherwise. The BMI average is 28.5 so the BMI5CAT variable is over wieght = 3 and the RFBMI5 is over 25 = 2.

data <- data[!(data$PREGNANT == 1),]  
index <- is.na(data$X\_BMI5)  
data$X\_BMI5[index] <- popavg  
data$X\_BMI5CAT[index] <- 3  
data$X\_RFBMI5[index] <- 2  
summary(data[9:11])

## X\_BMI5 X\_BMI5CAT X\_RFBMI5   
## Min. :1221 Min. :1.000 Min. :1.000   
## 1st Qu.:2816 1st Qu.:3.000 1st Qu.:2.000   
## Median :2816 Median :3.000 Median :2.000   
## Mean :2802 Mean :2.956 Mean :1.889   
## 3rd Qu.:2816 3rd Qu.:3.000 3rd Qu.:2.000   
## Max. :9931 Max. :4.000 Max. :2.000

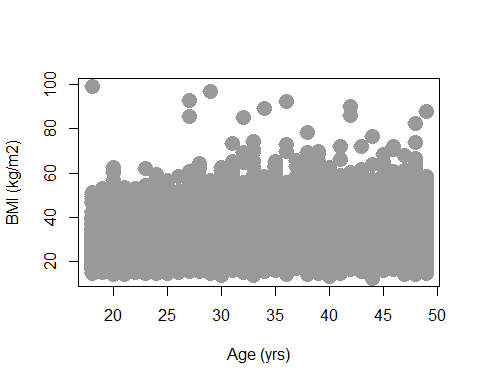
I plot a histogram of the data and find that most of the data sits in the 25-30 kg/m2 range. The ~10,000 that I added did not significantly change the data structure.

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 dataset. I will add that level to the data for the visualizations. This gives another level of detail towards poor health conditions.

data$X\_BMI5CAT[data$X\_BMI5 > 4000] <- 5  
summary(data[9:11])

## X\_BMI5 X\_BMI5CAT X\_RFBMI5   
## Min. :1221 Min. :1.000 Min. :1.000   
## 1st Qu.:2816 1st Qu.:3.000 1st Qu.:2.000   
## Median :2816 Median :3.000 Median :2.000   
## Mean :2802 Mean :2.971 Mean :1.889   
## 3rd Qu.:2816 3rd Qu.:3.000 3rd Qu.:2.000   
## Max. :9931 Max. :5.000 Max. :2.000

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

MMSABMI <- aggregate(data$X\_BMI5/100~data$County, FUN = mean)  
names(MMSABMI) <- c("County", "BMI\_(kg/m2)")  
str(MMSABMI)

## 'data.frame': 135 obs. of 2 variables:  
## $ County : chr "Aberdeen" "Albany-Schenectady-Troy" "Albuquerque" "Allentown-Bethlehem-Easton" ...  
## $ BMI\_(kg/m2): num 28.5 28.9 27.3 26.9 28.2 ...

Then I will create 4 proportion (percentage) values for % underwieght, % Normal wieght, % Overwieght, % Obese, and % extreme obesity by MMSA region.

state <- data[13:14]  
MMSABMI[3:7] <- as.data.frame.matrix(prop.table(table(data$County, data$X\_BMI5CAT), 1))\*100  
MMSABMI <- merge(x = MMSABMI, y = state[!duplicated(data$County),], by = "County")  
names(MMSABMI) <- c("MMSANAME", "BMI\_(kg/m2)", "%Underweight", "%Normal\_weight", "%Overwieght", "%Obese", "%Extreme\_Obesity", "State" )  
head(MMSABMI)

## MMSANAME BMI\_(kg/m2) %Underweight %Normal\_weight  
## 1 Aberdeen 28.48861 3.9215686 33.33333  
## 2 Albany-Schenectady-Troy 28.88532 0.0000000 32.67327  
## 3 Albuquerque 27.34366 2.1052632 38.59649  
## 4 Allentown-Bethlehem-Easton 26.85158 0.7874016 40.94488  
## 5 Anchorage 28.15143 1.5789474 38.94737  
## 6 Atlanta-Sandy Springs-Roswell 27.49161 2.0733652 40.19139  
## %Overwieght %Obese %Extreme\_Obesity State  
## 1 29.41176 23.52941 9.803922 SD  
## 2 35.64356 23.76238 7.920792 NY  
## 3 37.19298 17.54386 4.561404 NM  
## 4 37.00787 19.68504 1.574803 PA-NJ  
## 5 26.84211 25.26316 7.368421 AK  
## 6 31.25997 21.37161 5.103668 GA

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

write.csv(MMSABMI, 'healthdata.csv')

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>