BRFSS 2017 Survey EDA-BMI

Amanda Kimball

11/22/2020

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.

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, 174, 175, 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 16 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 ...  
## $ X\_MMSA : num 10100 10100 10100 10100 10100 10100 10100 10100 10100 10100 ...  
## $ X\_MMSAWT : num 110.2 28.6 115.6 376 20.7 ...  
## $ 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 calculated body mass indicator variables. As shown there are several missing values.

summary(data[9:11])

## X\_BMI5 X\_BMI5CAT X\_RFBMI5   
## Min. :1205 Min. :1.00 Min. :1.000   
## 1st Qu.:2373 1st Qu.:2.00 1st Qu.:1.000   
## Median :2695 Median :3.00 Median :2.000   
## Mean :2797 Mean :2.94 Mean :2.281   
## 3rd Qu.:3096 3rd Qu.:4.00 3rd Qu.:2.000   
## Max. :9931 Max. :4.00 Max. :9.000   
## NA's :19570 NA's :19570

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$X\_MMSA == 10100,]

I calculated the average for the Aberdeen and compared it with the average for the entire data sample and find that they are 27.9 and 27.5 kg/m2. The weight factor ‘X\_MMSAWT’ is used to adjust for sampling methodology. 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$X\_BMI5\*Aberdeen$X\_MMSAWT),na.rm=TRUE)/mean(Aberdeen$X\_MMSAWT,na.rm=TRUE)  
popavg <- mean((data$X\_BMI5\*data$X\_MMSAWT), na.rm=TRUE)/mean(data$X\_MMSAWT,na.rm=TRUE)

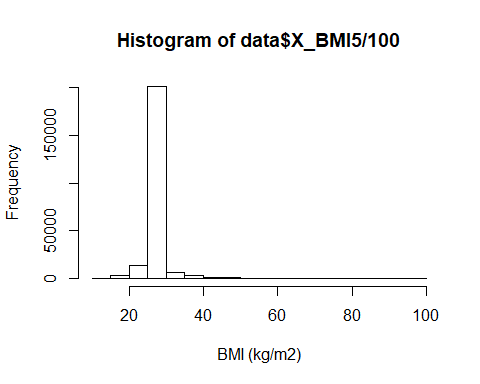
As indicated during the Aberdeen review - pregnant individuals can be removed as outliers outside the range of this analysis. The average BMI will be added otherwise. The BMI average is 27.9 so the BMI5CAT variable is over weight = 3 and the RFBMI5 is over 25 = 2.

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. :1.00 Min. :1.000   
## 1st Qu.:2750 1st Qu.:3.00 1st Qu.:2.000   
## Median :2750 Median :3.00 Median :2.000   
## Mean :2751 Mean :2.97 Mean :1.926   
## 3rd Qu.:2750 3rd Qu.:3.00 3rd Qu.:2.000   
## Max. :9931 Max. :4.00 Max. :2.000

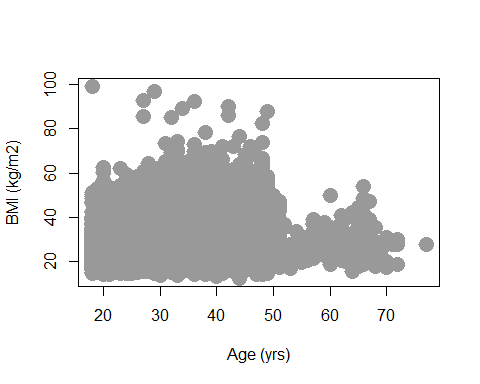
I plot a histogram of the data and find that most of the data sits in the 25-30 kg/m2 range. The ~13k that I added did not significantly change the data structure.

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[9:11])

## X\_BMI5 X\_BMI5CAT X\_RFBMI5   
## Min. :1221 Min. :1.00 Min. :1.000   
## 1st Qu.:2750 1st Qu.:3.00 1st Qu.:2.000   
## Median :2750 Median :3.00 Median :2.000   
## Mean :2751 Mean :2.98 Mean :1.926   
## 3rd Qu.:2750 3rd Qu.:3.00 3rd Qu.:2.000   
## Max. :9931 Max. :5.00 Max. :2.000

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. :25.03   
## 1st Qu.:19316 1st Qu.:26.75   
## Median :31110 Median :27.46   
## Mean :30108 Mean :27.55   
## 3rd Qu.:40080 3rd Qu.:28.27   
## Max. :49340 Max. :30.29

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

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[17:21]\*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. :25.03 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   
## Min. :0.2074 Min. :0.05835 Min. :0.006852   
## 1st Qu.:0.3004 1st Qu.:0.16969 1st Qu.:0.036991   
## Median :0.3297 Median :0.20359 Median :0.054013   
## Mean :0.3330 Mean :0.20401 Mean :0.057282   
## 3rd Qu.:0.3655 3rd Qu.:0.23704 3rd Qu.:0.073672   
## Max. :0.4705 Max. :0.33312 Max. :0.190972

Check to see if they add up to 100% of values for 1 region: Aberdeen, SD.

paste("BMI has", sum(MMSABMI[3:7][MMSABMI == 10100,])\*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, 'healthdata3.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>