Data Description

Zayed Shahjahan

12/8/2020

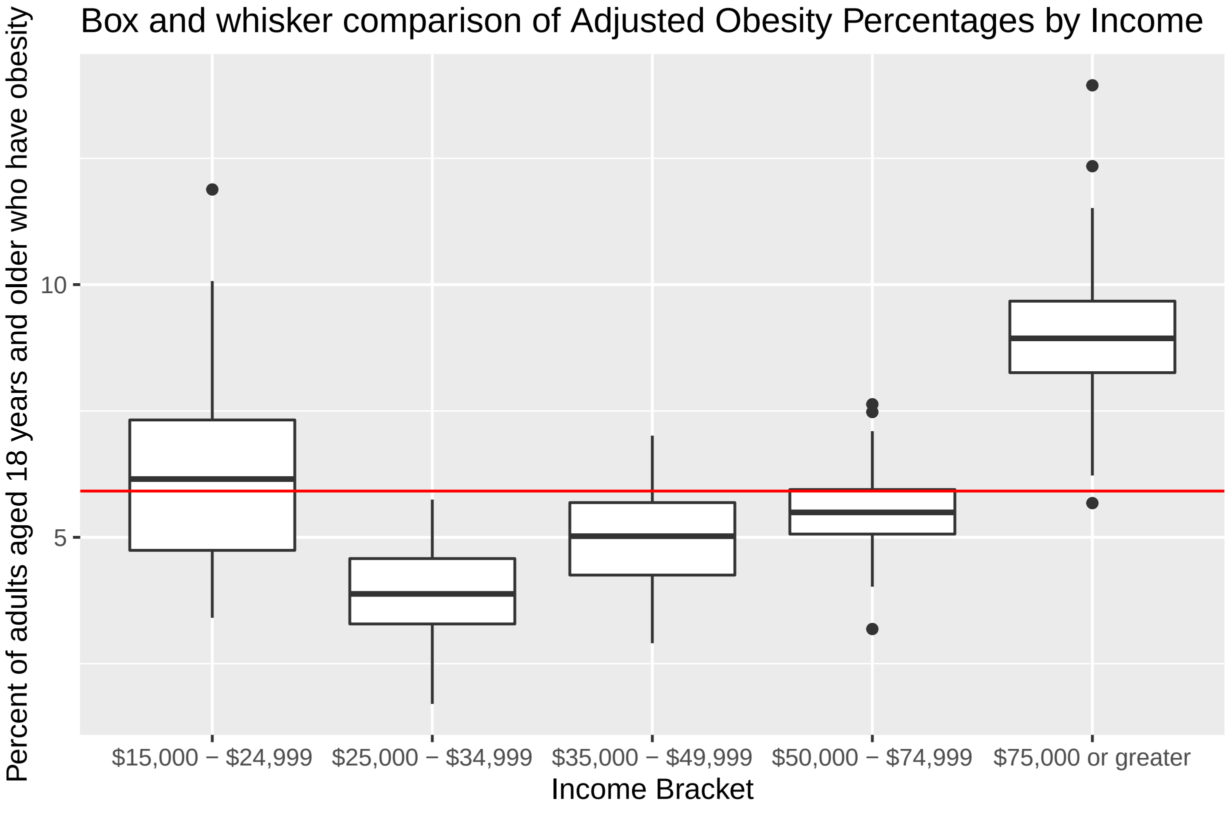
Variable descriptions:

LocationAbbr: This is the State code. Obtained from the original dataset. It is used in the creation of the ‘region’ variable. Consists of all 50 States and the District of Columbia

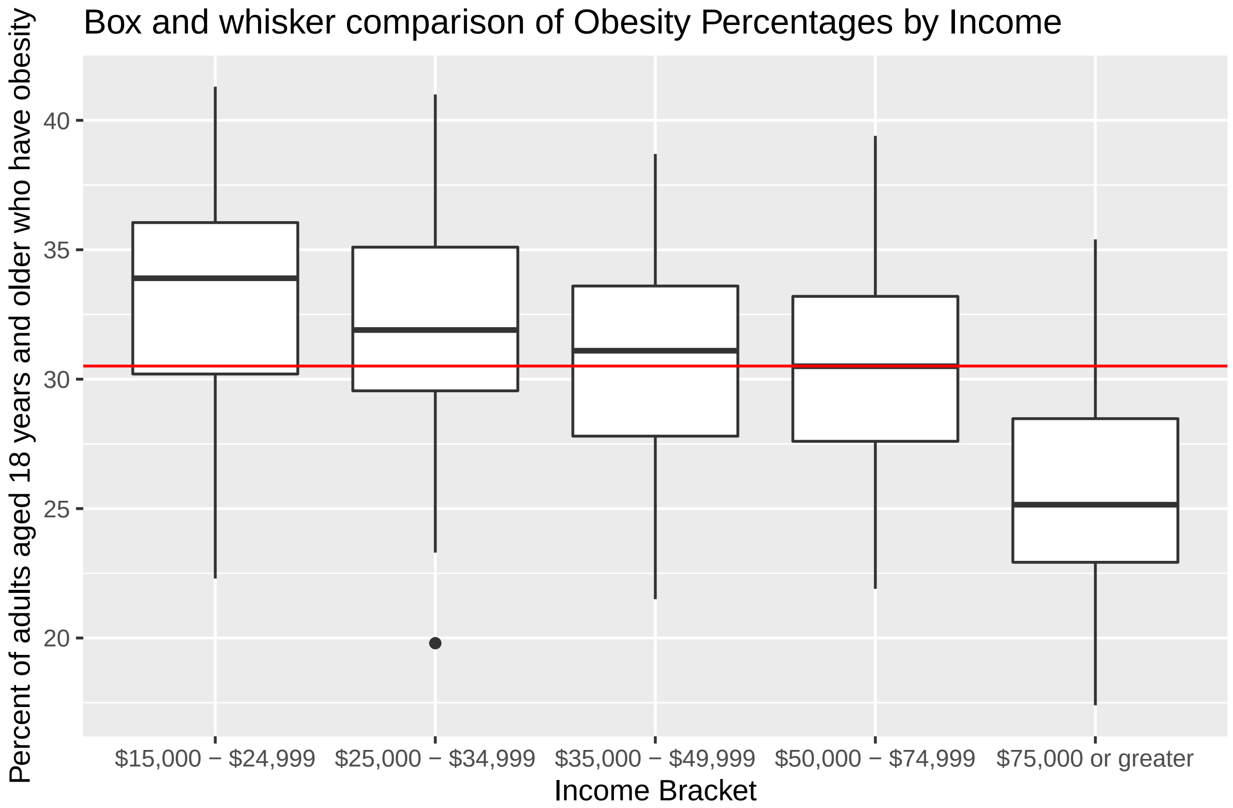
```
##  [1] "AK" "AL" "AR" "AZ" "CA" "CO" "CT" "DC" "DE" "FL" "GA" "HI" "IA" "ID" "IL"
## [16] "IN" "KS" "KY" "LA" "MA" "MD" "ME" "MI" "MN" "MO" "MS" "MT" "NC" "ND" "NE"
## [31] "NH" "NJ" "NM" "NV" "NY" "OH" "OK" "OR" "PA" "RI" "SC" "SD" "TN" "TX" "UT"
## [46] "VA" "VT" "WA" "WI" "WV" "WY"
```

Income: Factor consisting of 5 levels corresponding to income brackets. This is one of our variables of interest. This is obtained from the original dataset.

```
## [1] "$15,000 - $24,999"  "$25,000 - $34,999"  "$35,000 - $49,999"
## [4] "$50,000 - $74,999"  "$75,000 or greater"
```



red line represents dataset average



red line represents dataset average

Upon population adjustments, we can see that the relationship between obesity and income is not as linear as what we would have initially assumed.

perc_: Percentages for each of the questions of interest (see Introduction).

perc_strength: Percent of adults who engage in muscle-strengthening activities on 2 or more days a week

perc_cardio: Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination)

perc_excer: Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic physical activity and engage in muscle-strengthening activities on 2 or more days a week

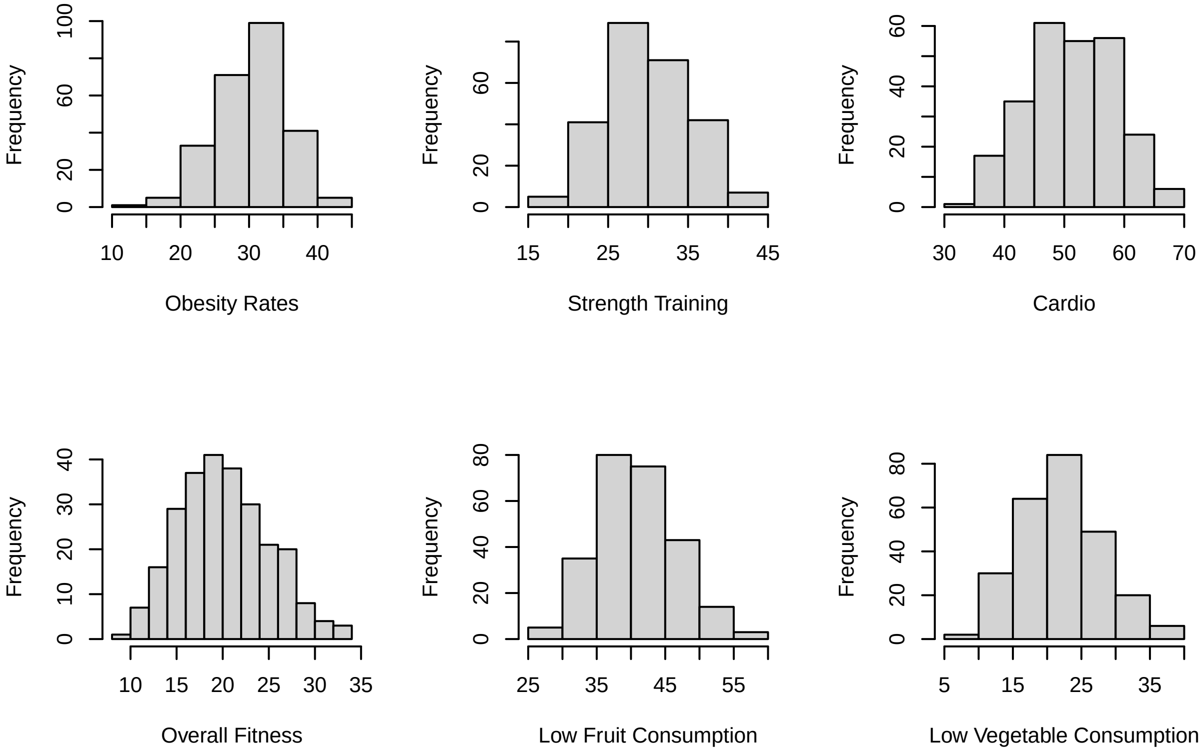
perc_fruit: Percent of adults who report consuming fruit less than one time daily

perc_veg: Percent of adults who report consuming vegetables less than one time daily

perc_obese: Percent of adults aged 18 years and older who have obesity

Obtained from the original dataset. Stratified on the basis of Income brackets for each state.

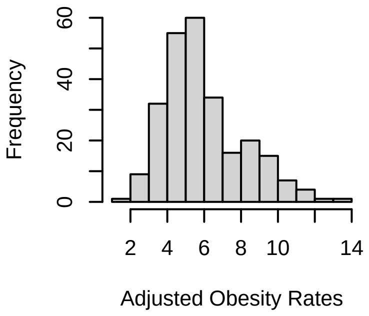
```
##  perc_strength    perc_cardio      perc_excer      perc_fruit
##  Min.   :16.70   Min.   :34.10   Min.   : 9.90   Min.   :27.40
##  1st Qu.:25.90   1st Qu.:46.50   1st Qu.:16.50   1st Qu.:36.45
##  Median :29.90   Median :51.00   Median :19.90   Median :40.60
##  Mean   :29.96   Mean   :51.42   Mean   :20.26   Mean   :40.82
##  3rd Qu.:33.80   3rd Qu.:56.60   3rd Qu.:23.60   3rd Qu.:44.80
##  Max.   :43.90   Max.   :70.00   Max.   :33.60   Max.   :57.50
##     perc_veg       perc_obese
##  Min.   : 9.30   Min.   :12.50
##  1st Qu.:17.40   1st Qu.:27.40  
##  Median :21.60   Median :30.70  
##  Mean   :21.99   Mean   :30.51  
##  3rd Qu.:25.70   3rd Qu.:33.85  
##  Max.   :39.60   Max.   :41.30  
```

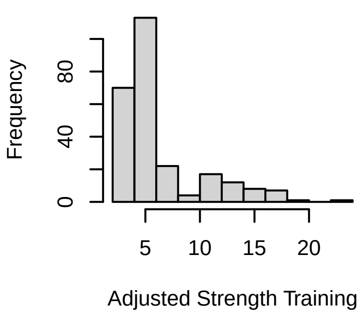


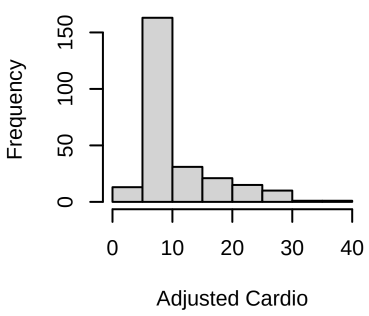
prop*: These variables were generated from the sample_size variable in the original dataset. This is calculated as the sample size in each income bracket divided by the total sample size across all income bracket in each state. This is then used as a proxy for state populations and the percentages for each of the questions are weighted using these prop-variables. 'prop' being short for proportions. The six prop variables correspond to the six questions of interest

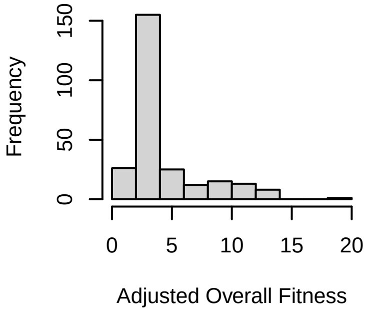
These variables are not of primary interest but are used to create the following adjusted variables.

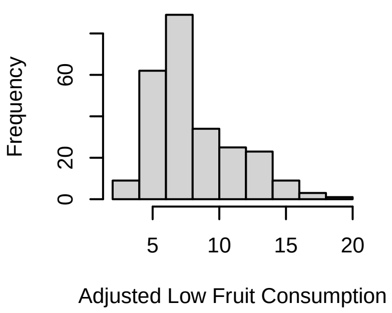
adj* : variables with the adjusted prefix correspond to those that have been justed to reflect population variation across states and across income brackets. This adjustment is intended to gain a more nuanced understanding of obesity in the American heartland.

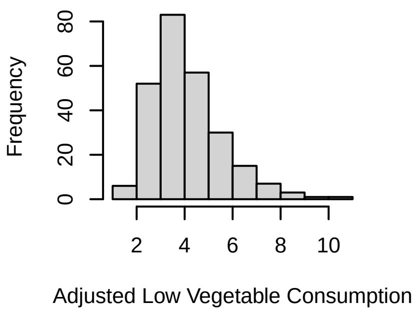






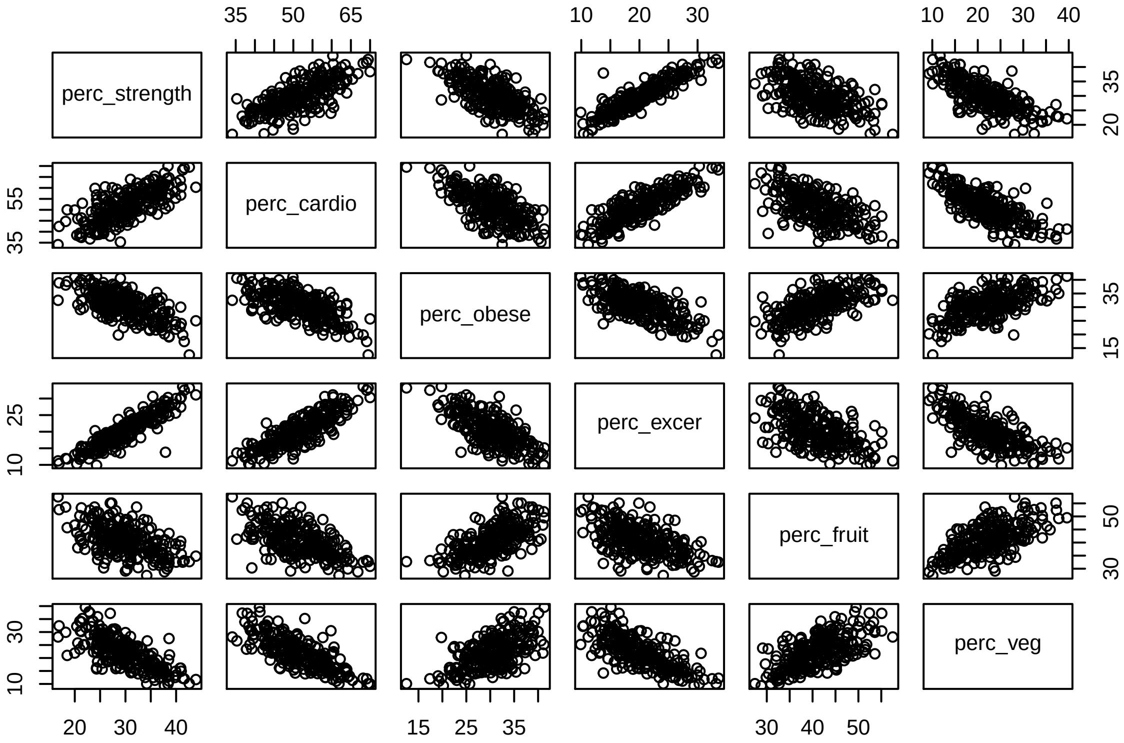




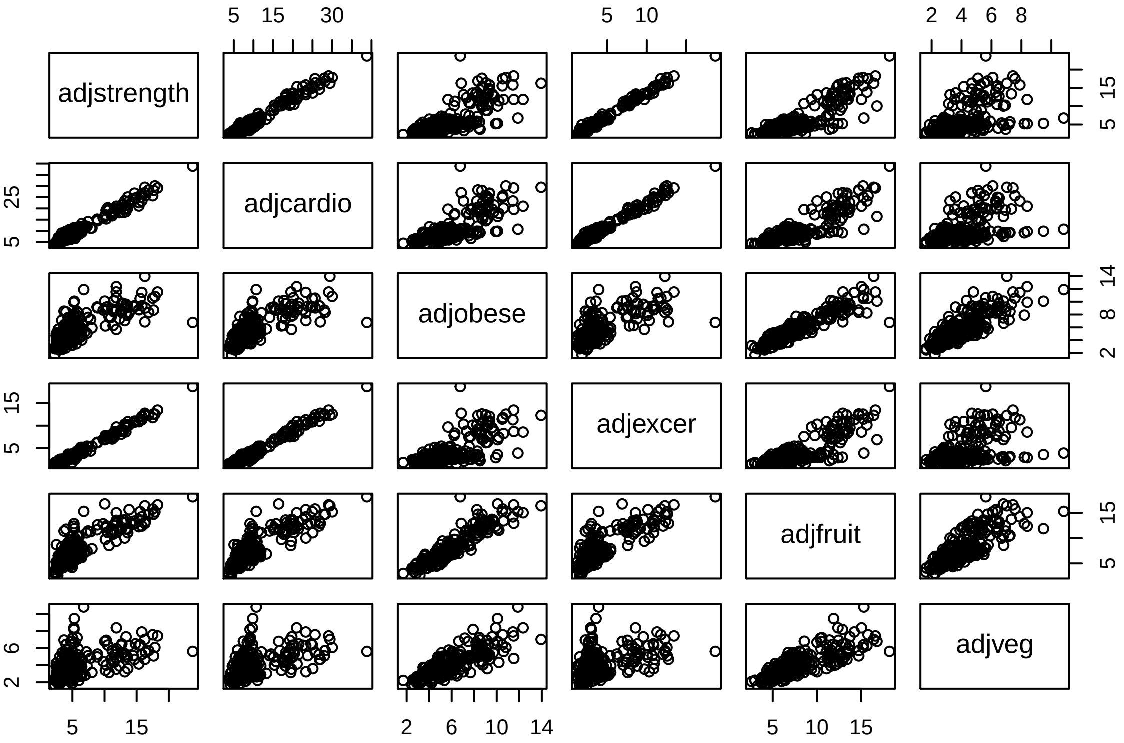


Notice that after taking population variation into account, there exists a strong left-skew in all the variables. The adjusted Obesity variable is what we will primarily be working with.

The following plots will also help highlight the effect of the adjustment

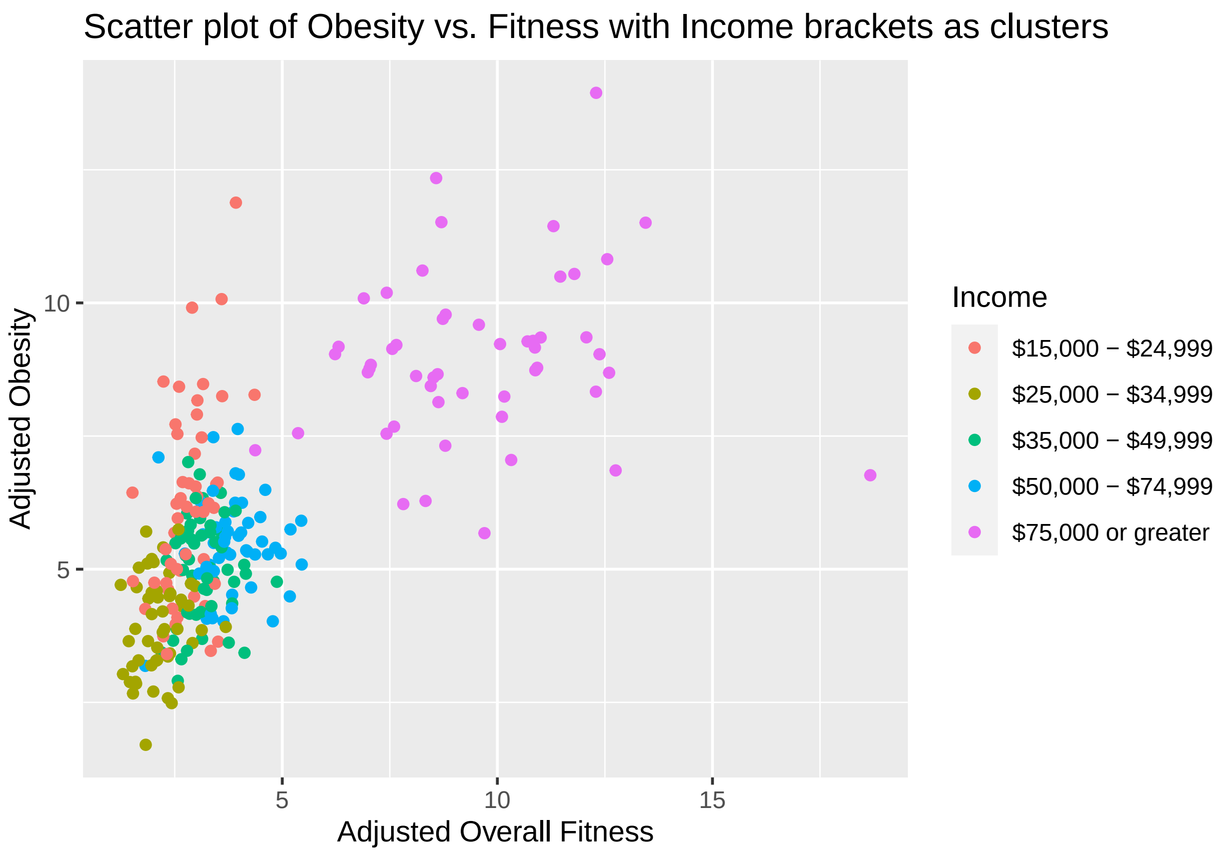


Compare this with:



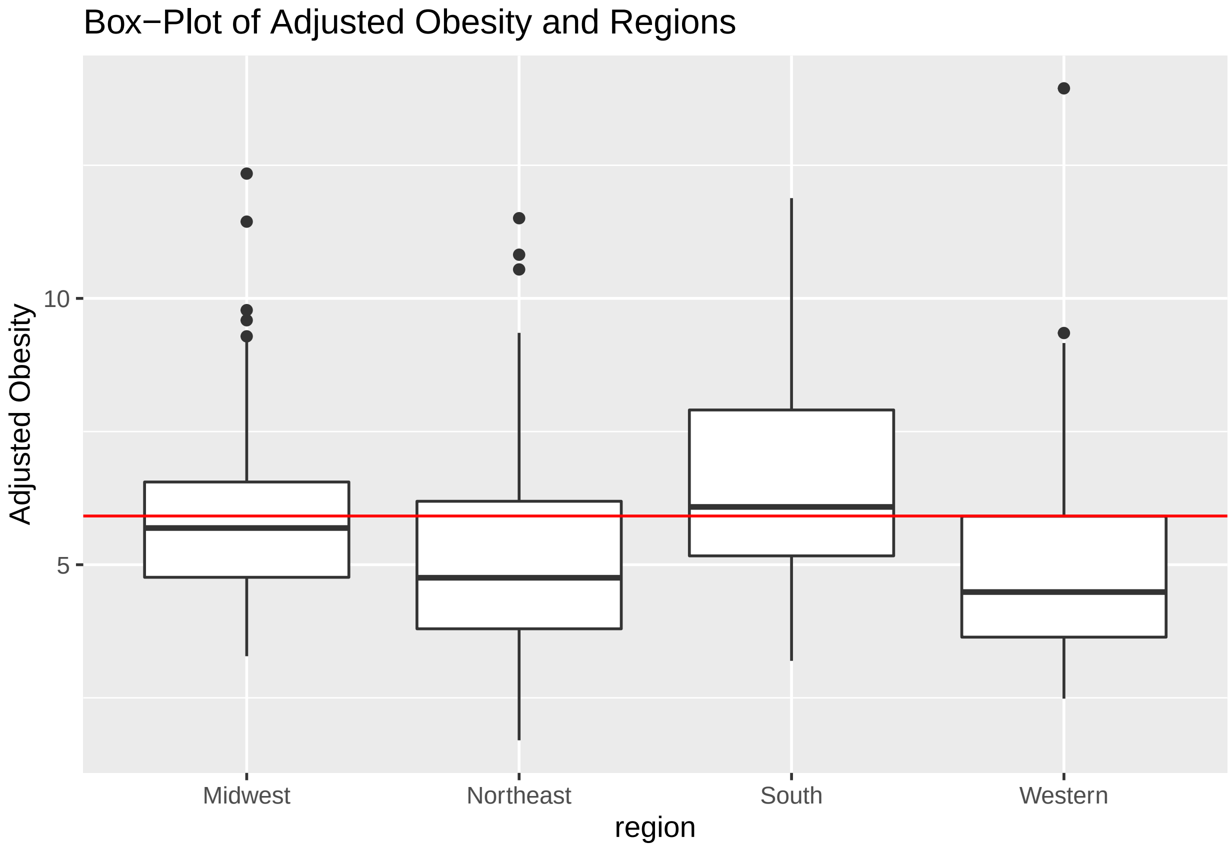
The presence of clusters is made more apparent in the second pairplot. Also the links between physical activity become more tenuous while the links between nutrition become more pronounced.

vrich: This is a dummy variable corresponding to the highest income bracket. The following figure will help illustrate why this variable was created



Clearly the data cloud can be partitioned into 2 very distinct clusters and this justifies the need for the vrich dummy variable.

region: Four geographical regions of the US. Created in order to investigate the effect of regional variations in obesity rates across states



Given the nature of this dataset, I've decided to add a random datapoint. There is not much of an explanation except for the fact that I've been asked to do so. I personally think it is not the best idea given how the original data were arranged and how I wrangled and feature-engineered the final dataset.

```
randpoint <- data.frame("LA",
                        "$75,000 or greater",
                        21.3,
                        0.19680851,
                        2.295946,
                        52.8,
                        0.43691149,
                        11.822823,
                        40.6,
                        0.13894812,
                        6.222623,
                        25.6,
                        0.15335388,
                        10.161504,
                        36.9,
                        0.48839675,
                        6.916326,
                        10.1,
                        0.17822432,
                        3.874917,
                        1,
                        "South")
```

```r
names(randpoint) <- c(colnames(dfFinal2))

dfFinal2 <- rbind(dfFinal2,randpoint)
```

Two new variables were introduced after we detected the presence of multicollinearity in our predictor variables. These are the following:

phys: average of all three physical activity variables

nutri: average of both nutrition related variables.