**Edward Kang**

**912655433**

**STA 141 A**

**The codes and results derived by using these codes constitute my own work. I have consulted the following resources regarding this assignment: classmate: Qiming Jiang.**

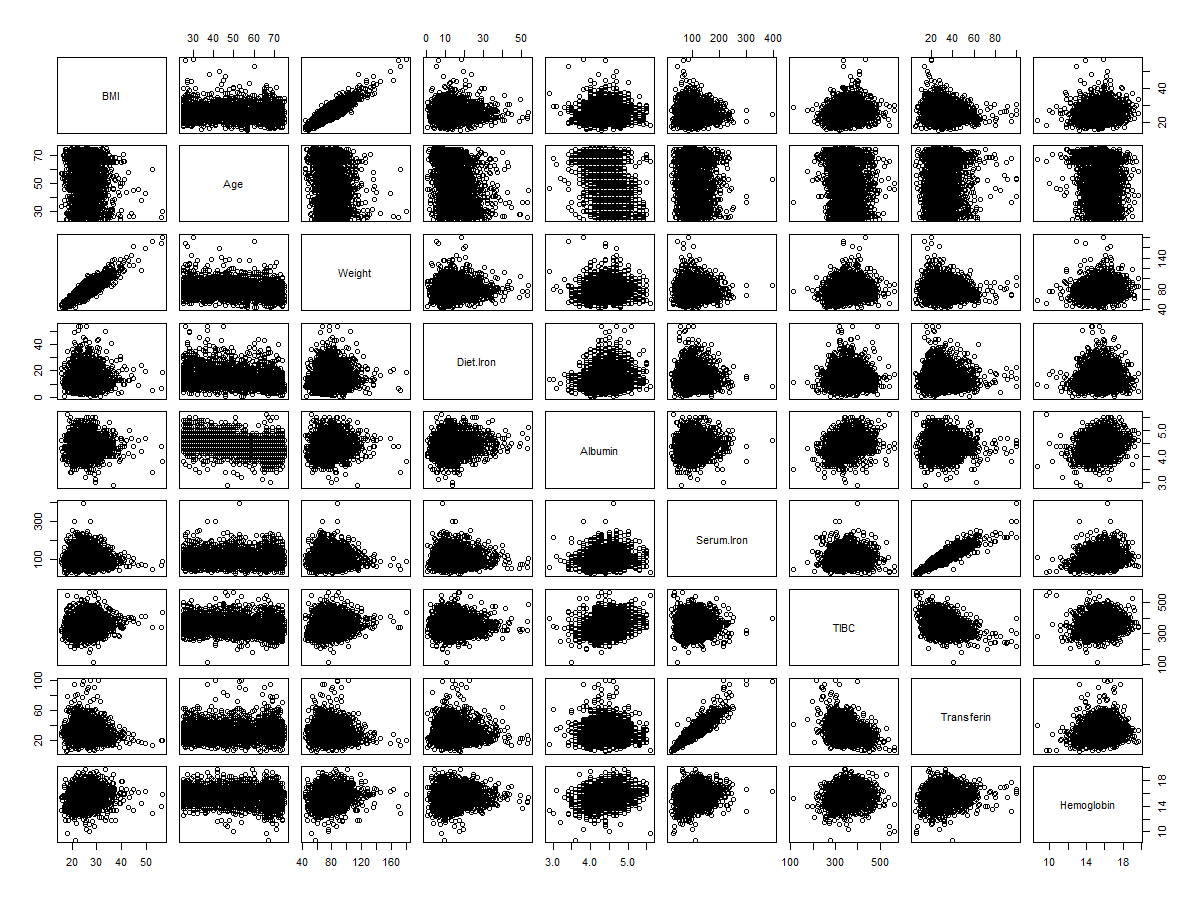
In this document, I am going to use linear regression and scatterplot smoother to find interesting patterns in the data of NHANES: National Health and Nutrition Examination Survey.

The data has many missing values, so I use na.omit() to omit the N/A values, Moreover, in this data, Age, Weight, BMI, Diet.Iron, Albumin, Serum.Iron, TIBC, Transferin, Hemoglobin are numeric variables while Cancer.Incidence, Cancer.Death, Smoke, Ed, Race and Sex are categorical variables. I will use the factor() function to tell R to treat those six as categorical variables.

|  |
| --- |
| data <- na.omit(NHANES)  data$Cancer.Incidence <- factor (data$Cancer.Incidence)  data$Cancer.Death <- factor (data$Cancer.Death)  data$Smoke <- factor(data$Smoke)  data$Ed <- factor(data$Ed)  data$Race <- factor(data$Race)  data$Sex <- factor(data$Sex) |

We first look at the data to see if we can find obvious relationships. I use pairs() function to create a set of scatterplots between every variable in the data

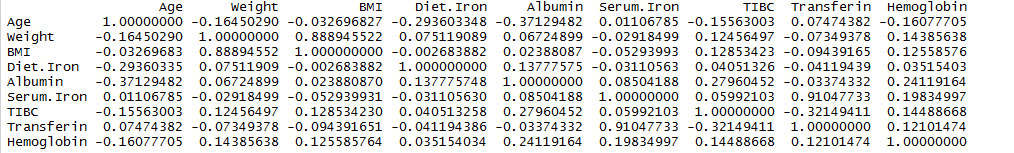
|  |
| --- |
| pairs(BMI ~ Age+ Weight + Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin + Hemoglobin, data = data) |



From the graph, we can see that BMI with weight and Serum.Iron and Transferin have a positive correaltion.

Next, we’ll look at the correlation coefficients between all the variables.

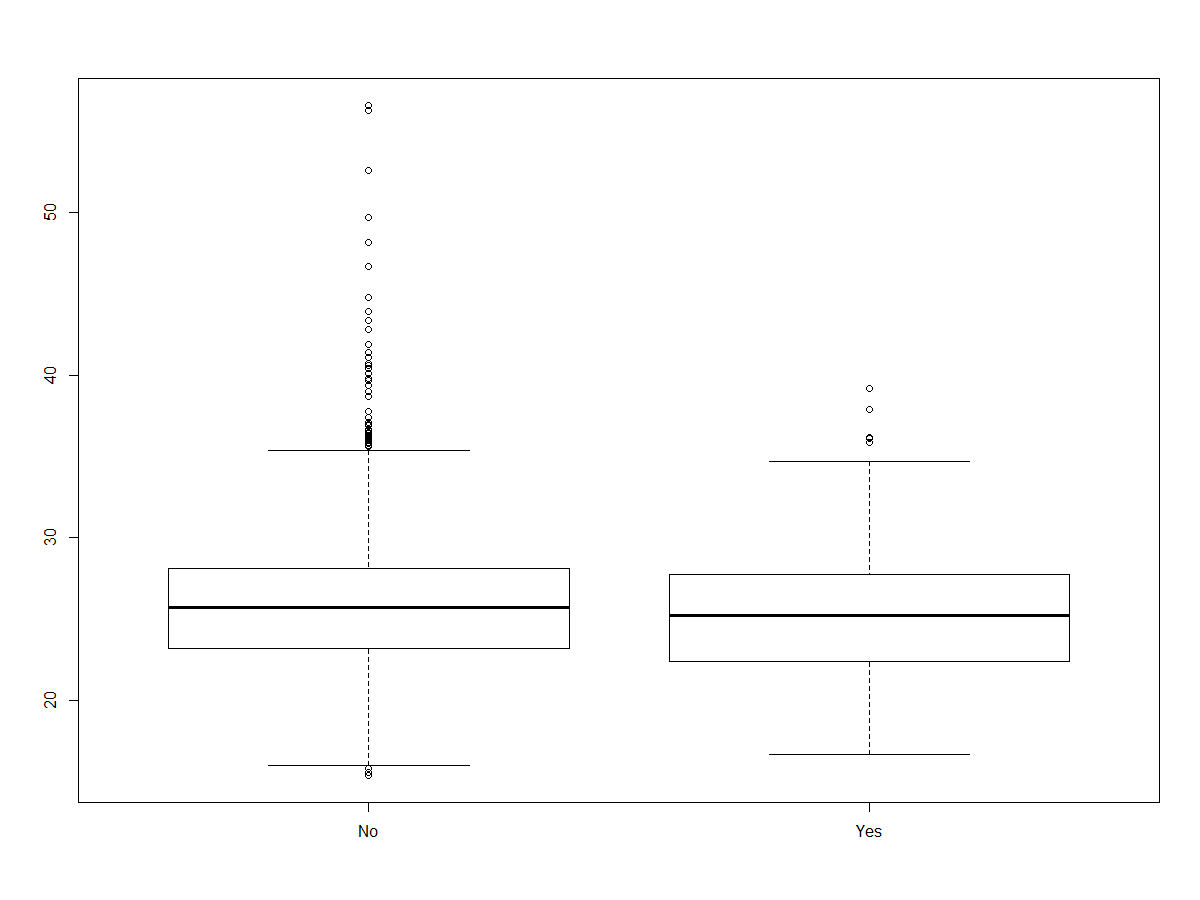
|  |
| --- |
| cor(data[,c(3,7,8,9,10,11,12,13, 14)], use = "everything") |



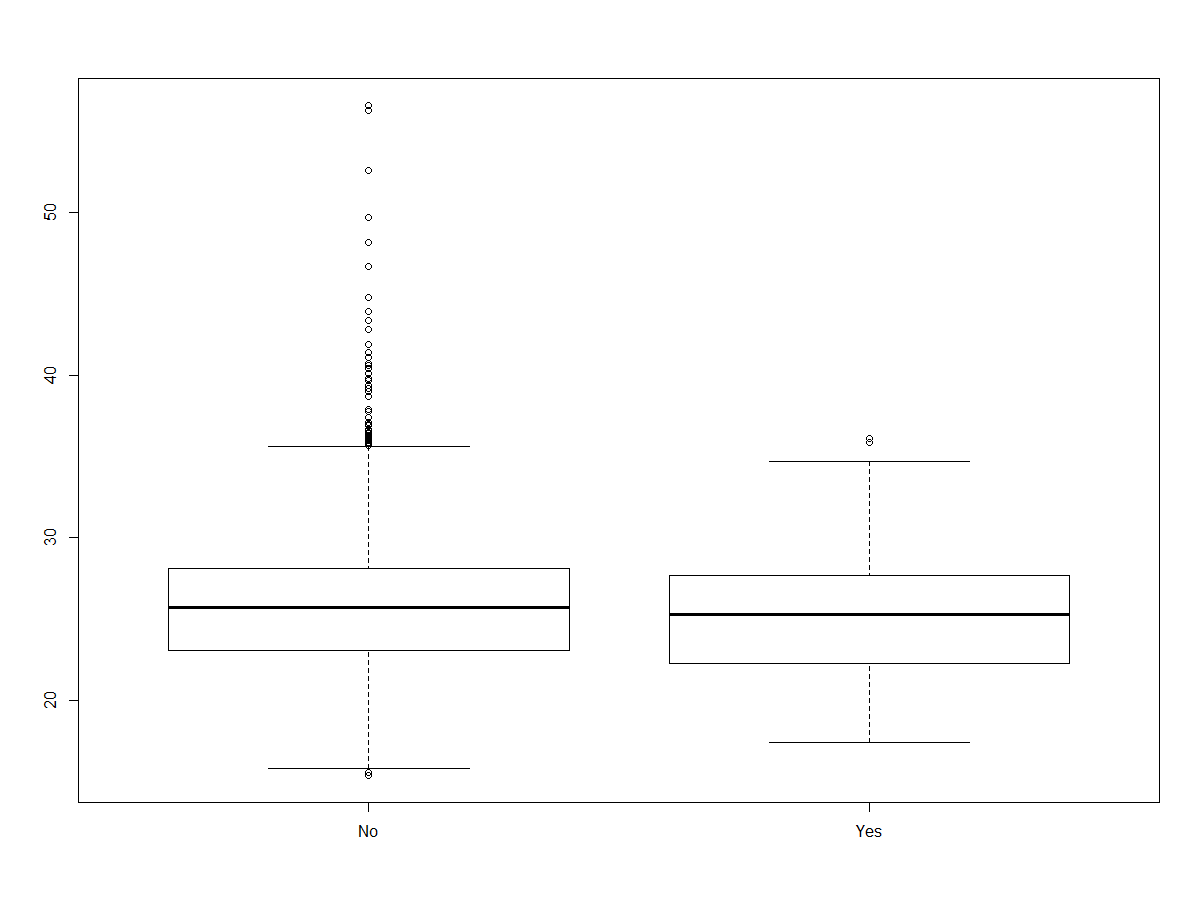
We can see that the correlation of between Weight and BMI and Serum.Iron and Transferin are particularly higher than others. And Serum.lron and Transferice have the highest positive coefficients of 0.9104773. It means that the most important linear variables are Serum.Iron and Transferin as well as Weight and BMI.

Now, let’s look at the categorical variables. Since my particular interested is the correlation of BMI between other variables, I do boxplots to see the relationship between different levels of the categorical variables and BMI

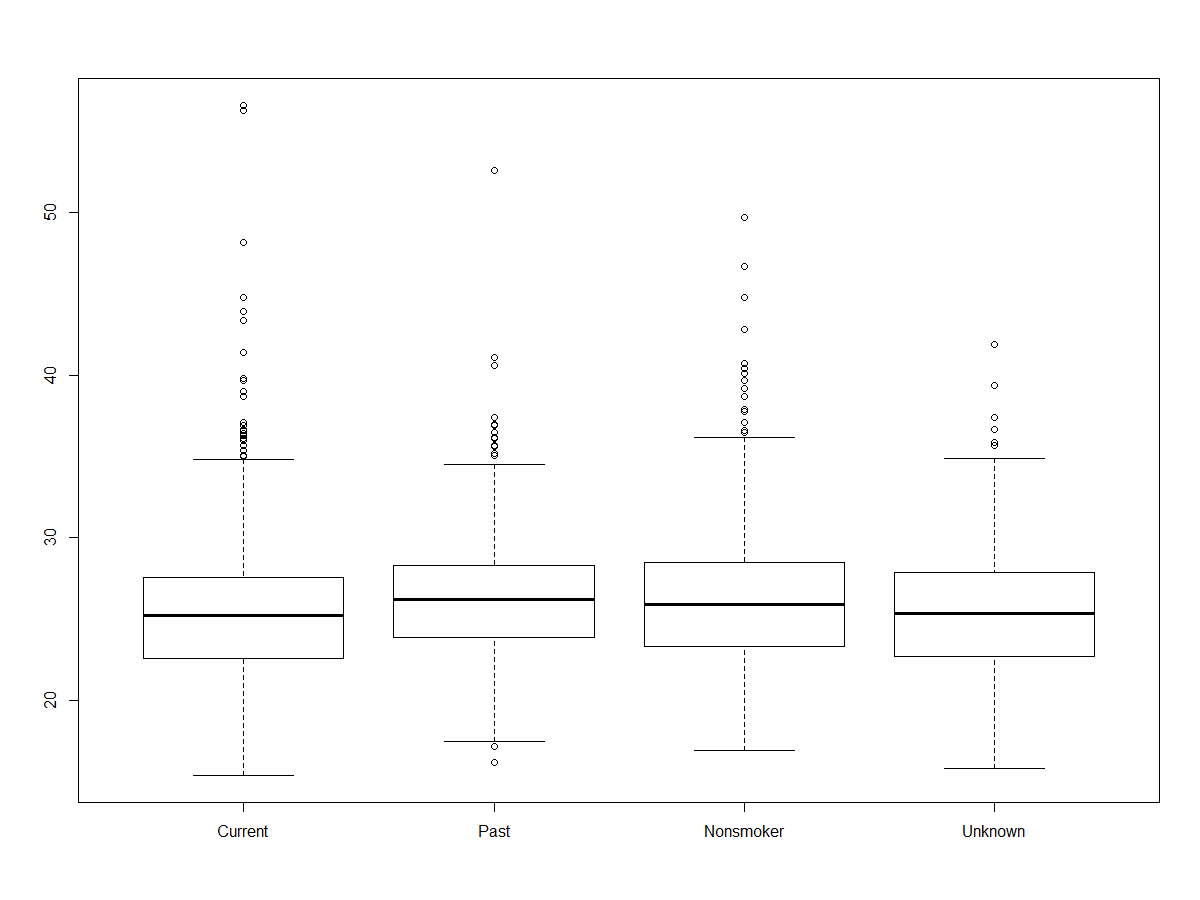
|  |
| --- |
| boxplot(BMI ~ Cancer.Incidence, data = data) |



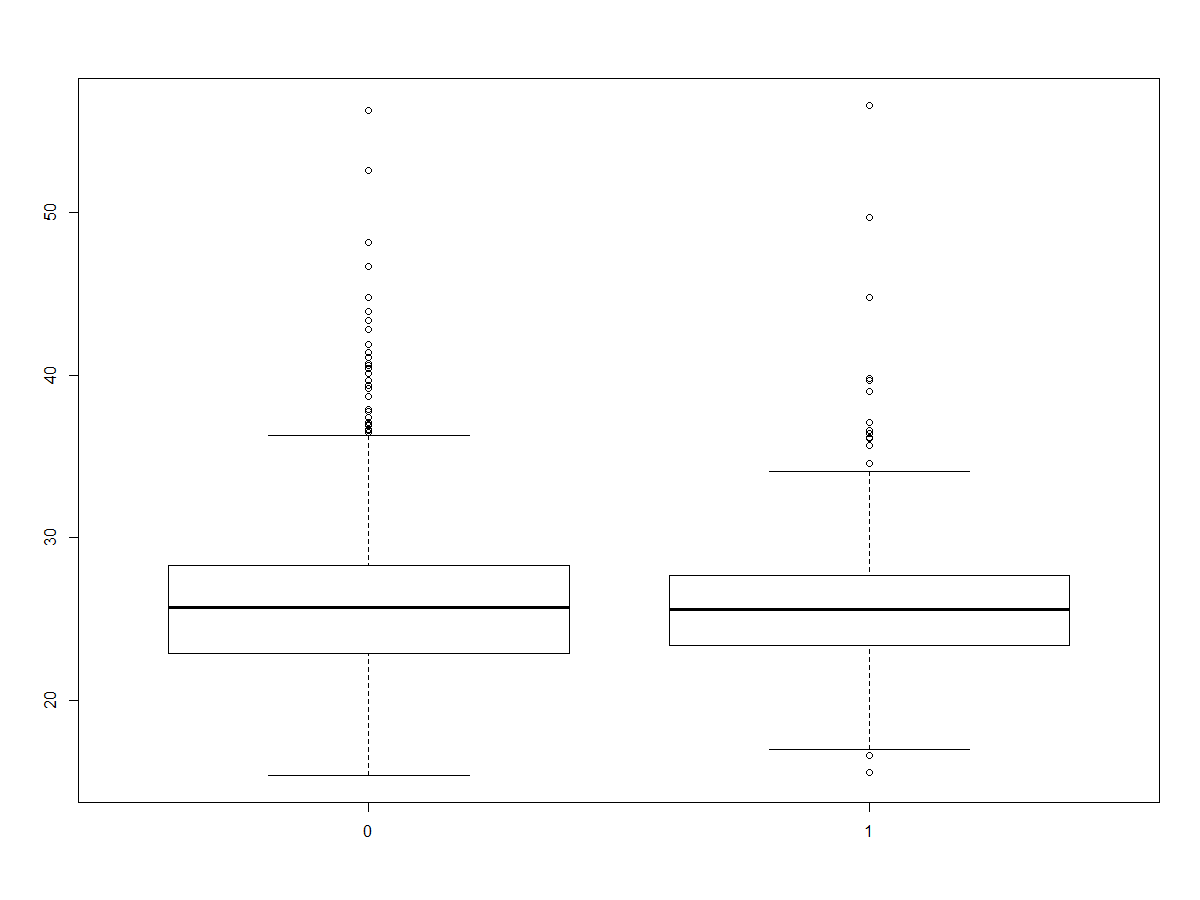
|  |
| --- |
| boxplot(BMI ~ Cancer.Death, data = data) |



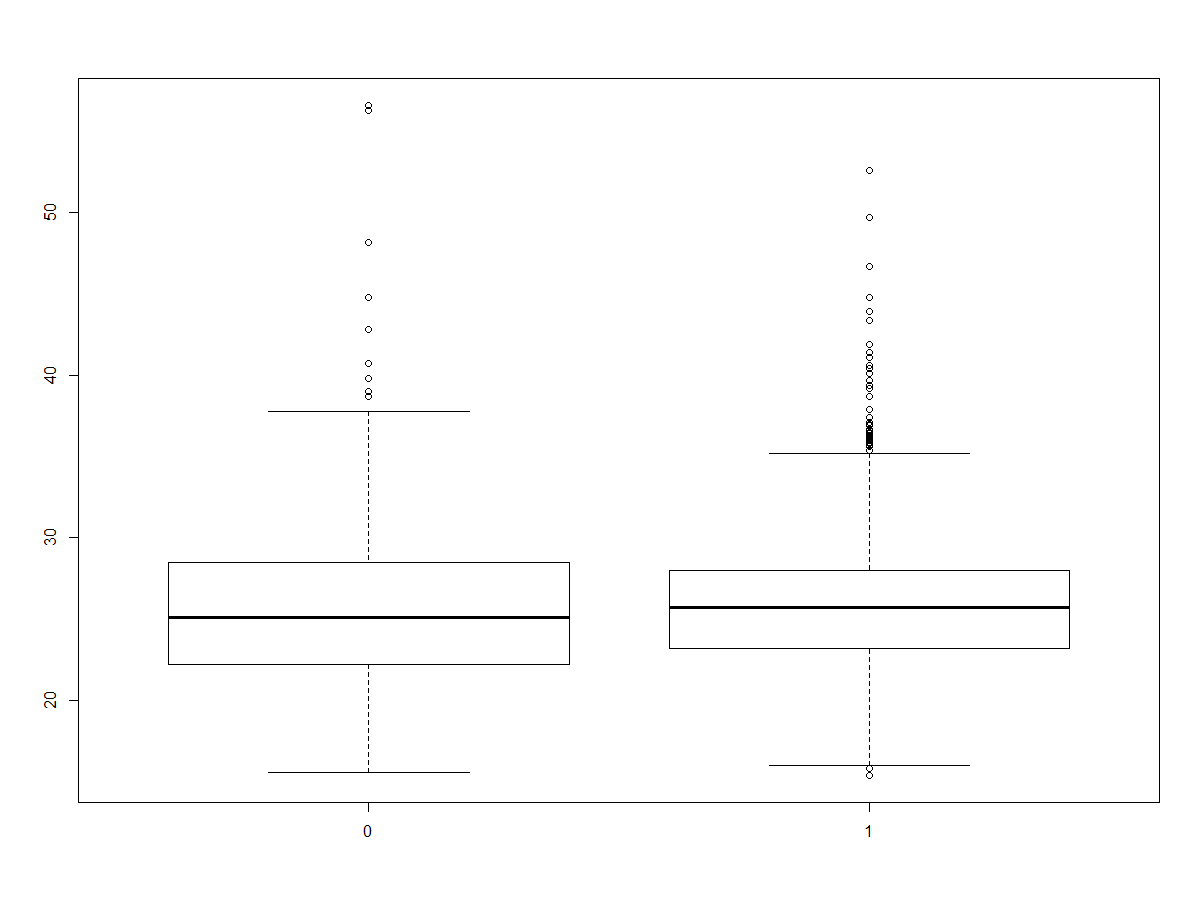
|  |
| --- |
| boxplot(BMI ~ Smoke, data = data) |



|  |
| --- |
| boxplot(BMI ~ Ed, data = data) |



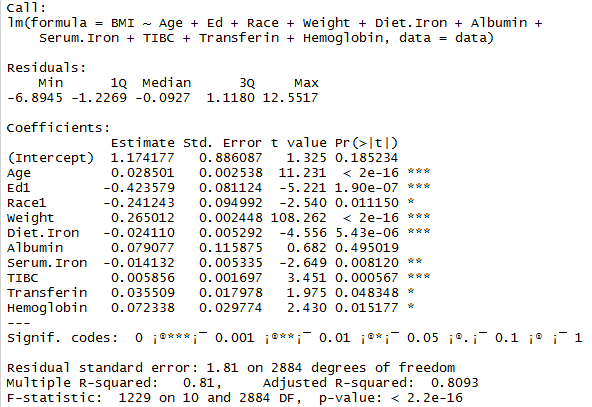
|  |
| --- |
| boxplot(BMI ~ Race, data = data) |



The Ed and Race boxplots suggest that BMI varies over the level of Ed and different Race, so we are going to including these in our model. However, the other three boxplots show that the BMI does not change significantly, so I will omit those three variables.

We start fitting the model

|  |
| --- |
| fit <- lm(BMI~ Age+ Ed+ Race+ Weight + Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin + Hemoglobin, data = data)  summary(fit) |



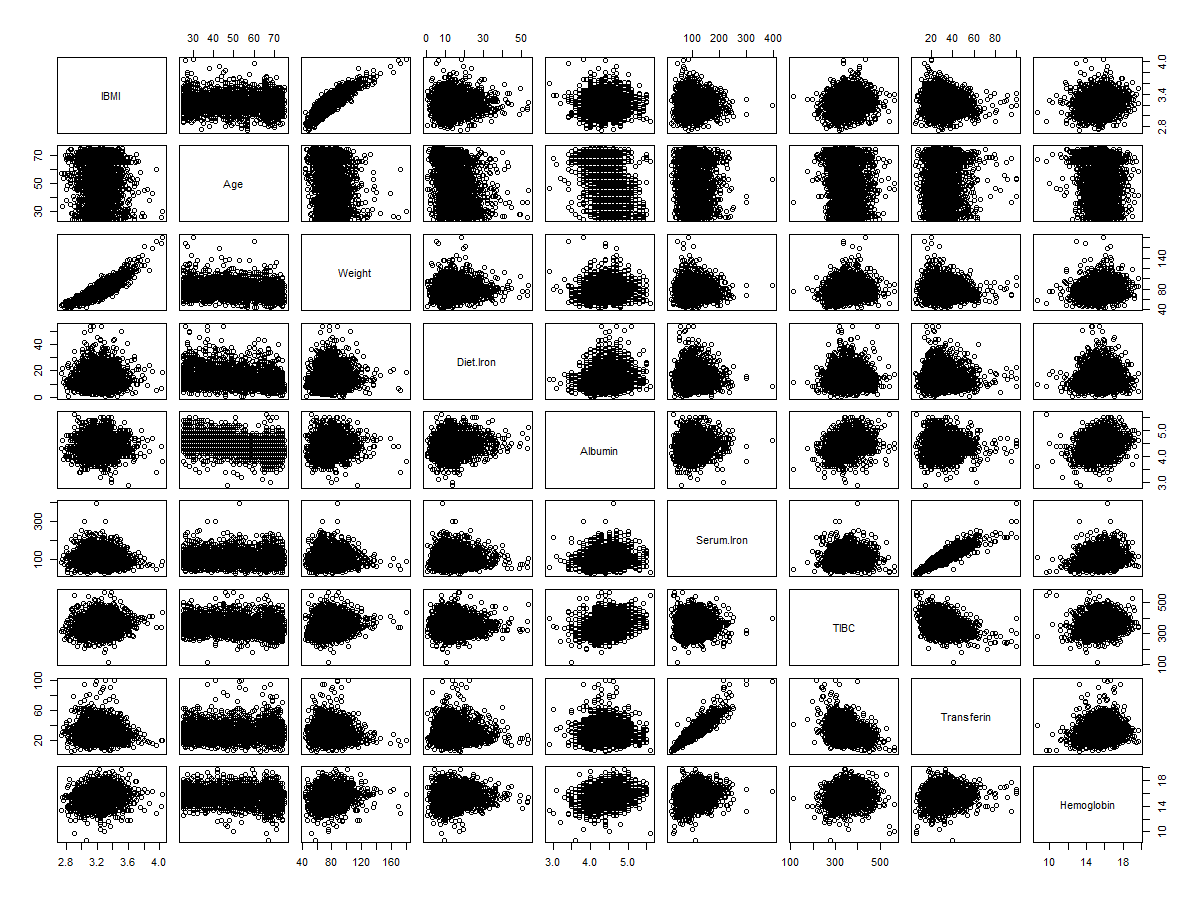
The p-value is the probability of achieving the t-value if the null hypothesis were true.

The multiple R-squared and adjusted R-squared both correspond to the amount of variability in the dependent variable that is explained by the model

The F-statistic is a goodness of fit test for the model.

We try to transform the data to see whether we can make the relationship more linear, so we use log transformation on the variable BMI.

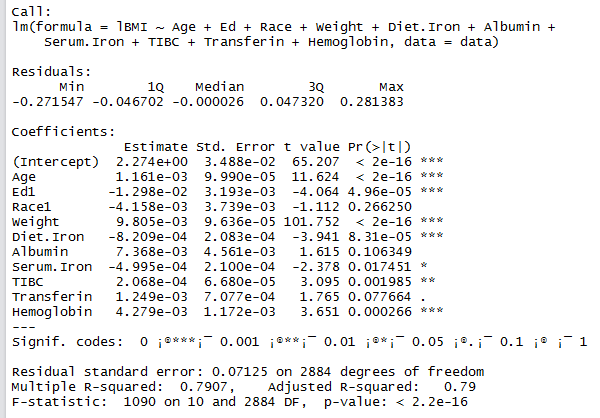
|  |
| --- |
| data$lBMI <- log(data$BMI)  data$BMI <- NULL  pairs(lBMI ~ Age+ Weight + Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin + Hemoglobin, data = data) |



The plot does not seem to have much change.

We try to fit the model again

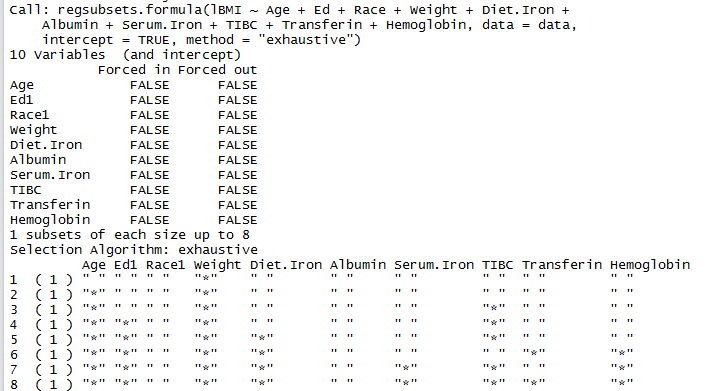
|  |
| --- |
| fit <- lm(lBMI~ Age+ Ed+ Race+ Weight + Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin + Hemoglobin, data = data)  summary(fit) |



The p-value and adjusted R-squared does not change much base on the result, the log transformation does not help much. We can omit the transformation.

Not all of the variables appear to be significant, and some are correlated with each other and thus give no additional information. Therefore, we have to select model variables.

|  |
| --- |
| library(leaps)  fit <- regsubsets(lBMI~ Age+ Ed+ Race+ Weight + Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin + Hemoglobin, data = data, intercept = TRUE, method = "exhaustive")  summary(fit)  summary(fit)$adjr2 |



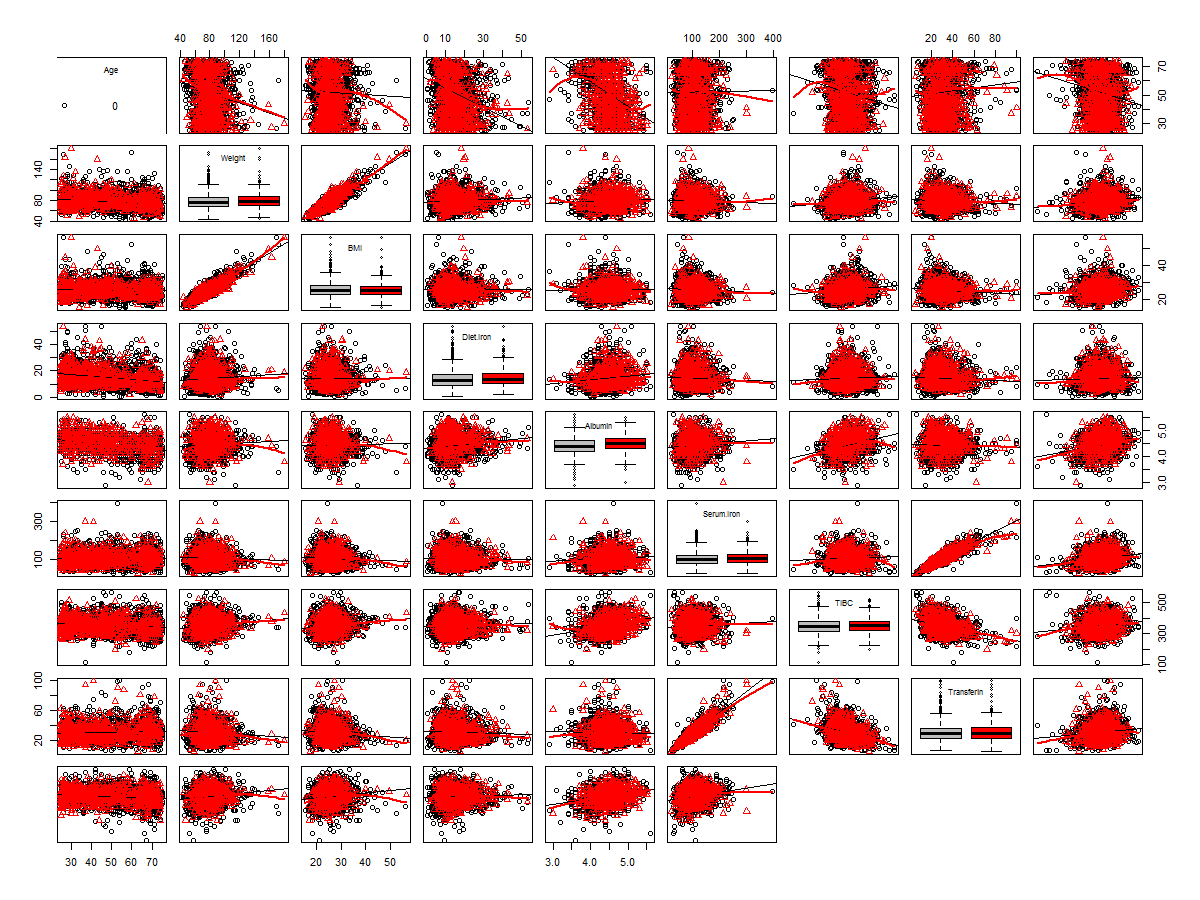


The function will return the model with best R-squared for each size. Overall, the best model is a model that drops Race1, Albumin, and it has an R-squared of 0.7898905.

Next we look at the natural clusters in the data. There are natural clusters in the data and we try to see whether these clusters related to different natural strata determined by the values of the categorical variables like Ed, Race, Sex, Smoke. We do a scatterplot matrix by those categorical variables.

First we look at Ed:

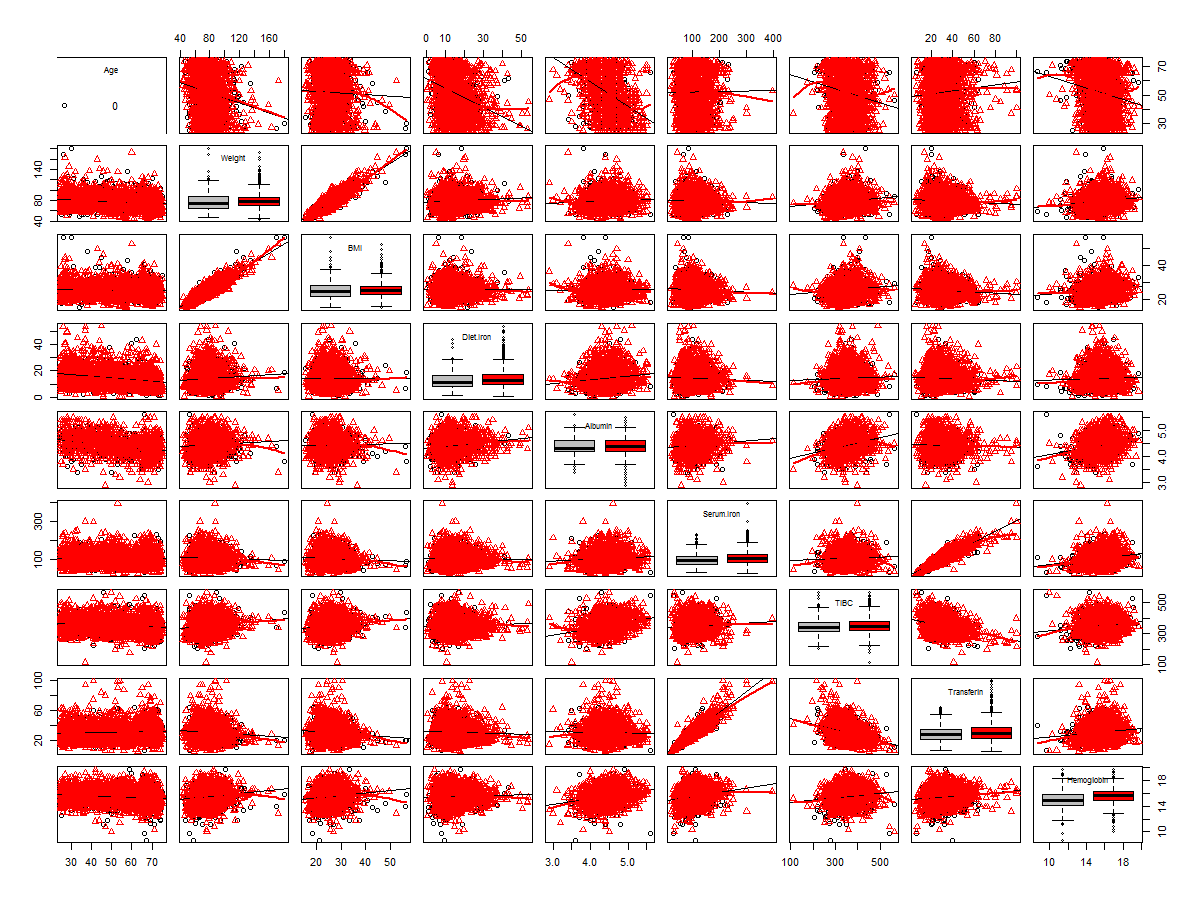
|  |
| --- |
| scatterplotMatrix(~ Age+ Weight + BMI+ Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin + Hemoglobin|Ed, data = NHANES, spread = FALSE, span = 0.8, diagonal = "boxplot") |



The clusters are mostly overlapped, so there is not significant difference between education levels.

Next we look at Race:

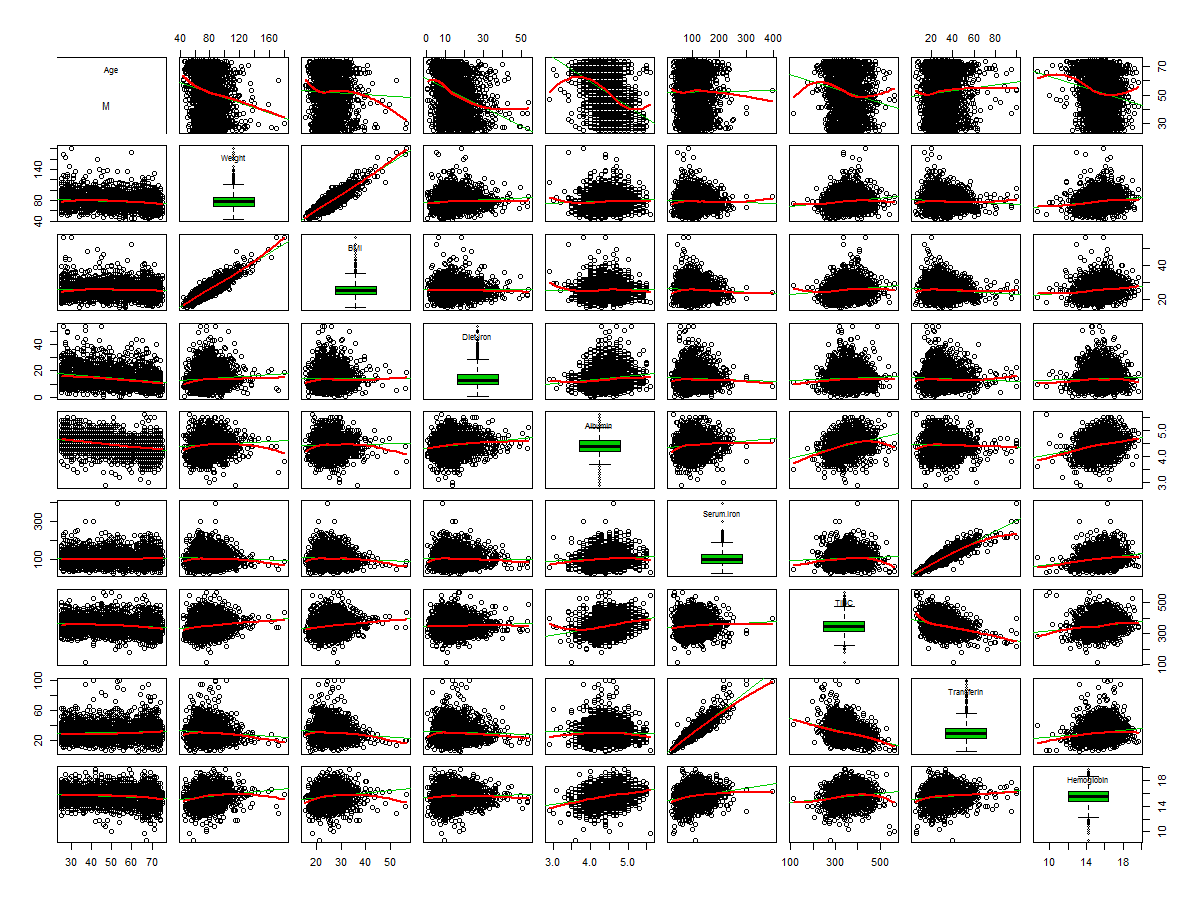
|  |
| --- |
| scatterplotMatrix(~ Age+ Weight + BMI+ Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin + Hemoglobin|Race, data = NHANES, spread = FALSE, span = 0.8, diagonal = "boxplot") |



The clusters are mostly overlapped, so there is not significant difference between different race.

Sex:

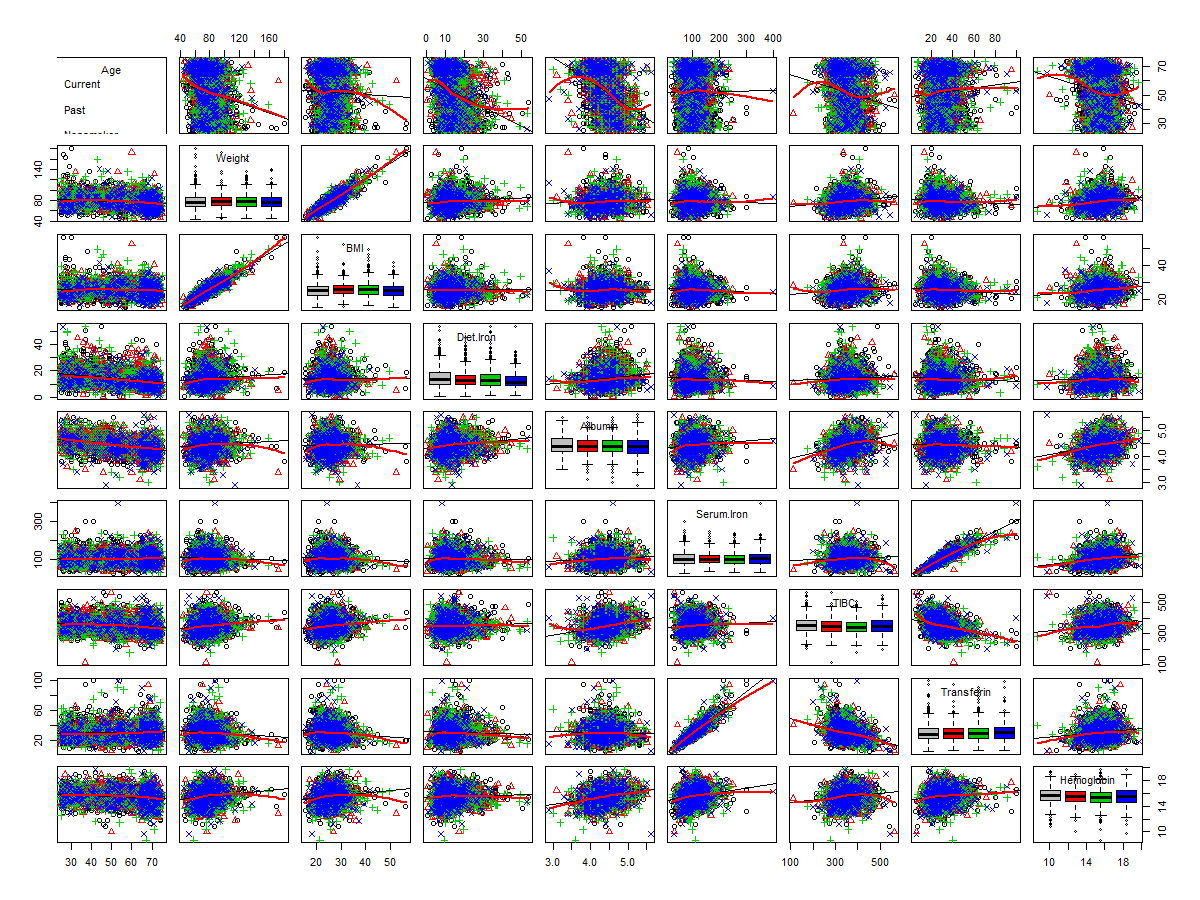
|  |
| --- |
| scatterplotMatrix(~ Age+ Weight + BMI+ Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin + Hemoglobin|Sex, data = NHANES, spread = FALSE, span = 0.8, diagonal = "boxplot") |



We can see the Sex variable is very different from the two variables above. The clusters do not overlap at all. As a result, the Sex variable is significant difference between male and female since some of the values of female are missing

Smoke:

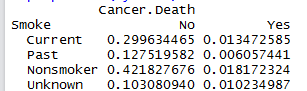
|  |
| --- |
| scatterplotMatrix(~ Age+ Weight + BMI+ Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin + Hemoglobin|Smoke, data = NHANES, spread = FALSE, span = 0.8, diagonal = "boxplot") |

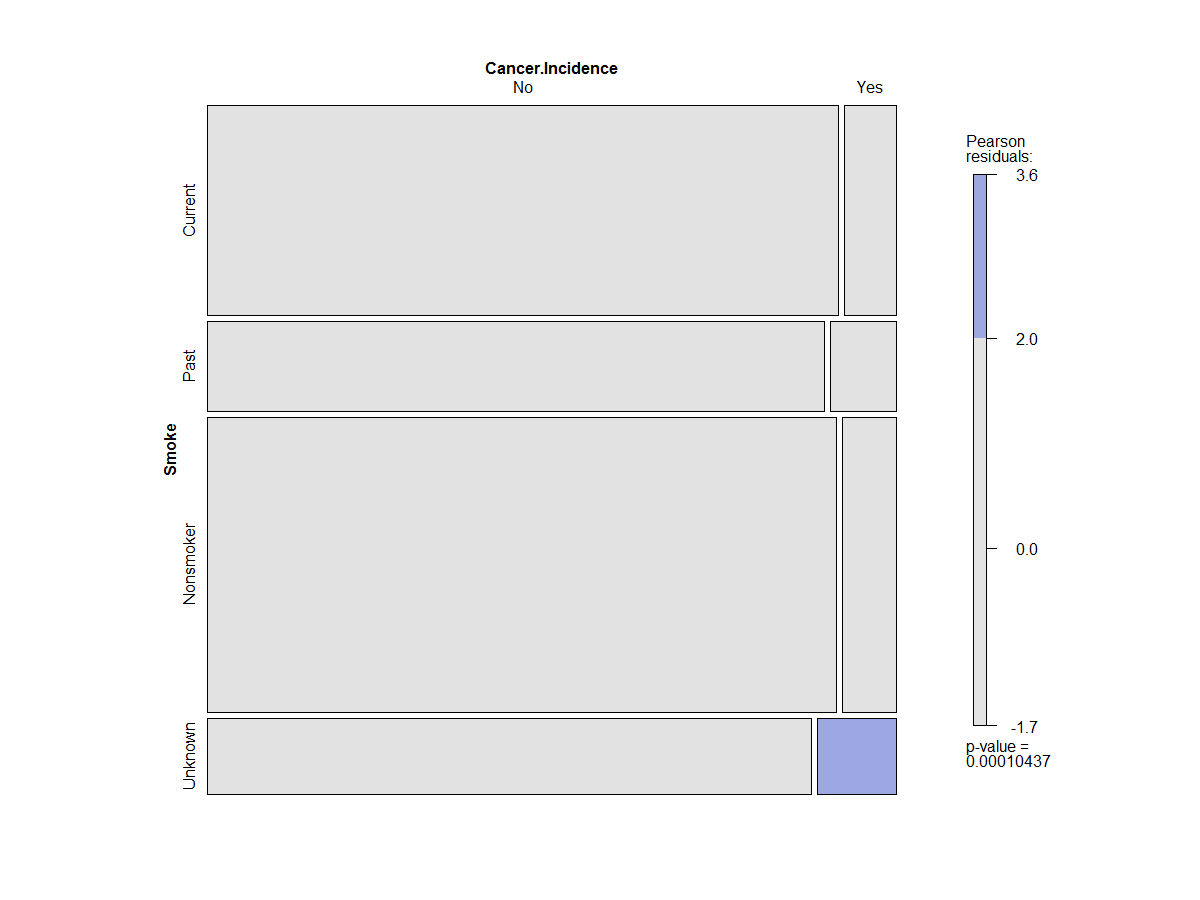


The plot shows the clusters are mostly overlapped again, so there is not significant difference between smoking or not.

Finally, we look at the pattern is present in terms of the effect of smoking on incidences of cancer. We will do a contingency table between the categorical variables Smoke and Cancer.Incidence and then create the mosaic plot of the table.

|  |
| --- |
| attach(NHANES)  mytable <- table(Smoke, Cancer.Death)  prop.table(mytable)  mosaic(mytable, shade=TRUE, legend=TRUE) |



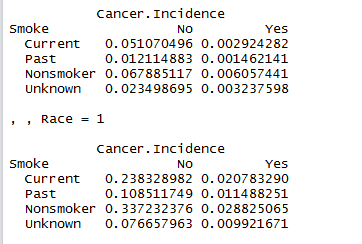


First we should omit the data of unknown since it does not relate to our research. From the table, we can see that people have smoked before or still smoking have about 2% higher population to incidence cancer. Also, from the plot we can see the similar information that the population of getting cancer is almost the same whether smoking or not. The reason maybe because the sample size is not big enough to state that smoking causes more cancer, or the face is that smoking does not have huge impact of getting cancer.

Next we look at how the patterns change across age groups, gender, and race.

First we look at Race:

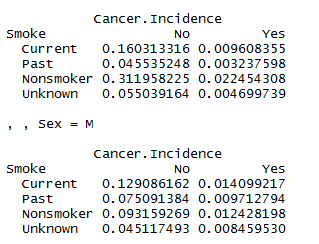
|  |
| --- |
| mytable <- table(Smoke, Cancer.Incidence, Race)  ptable = prop.table(mytable)  ptable  mosaic(mytable, shade=TRUE, legend=TRUE) |

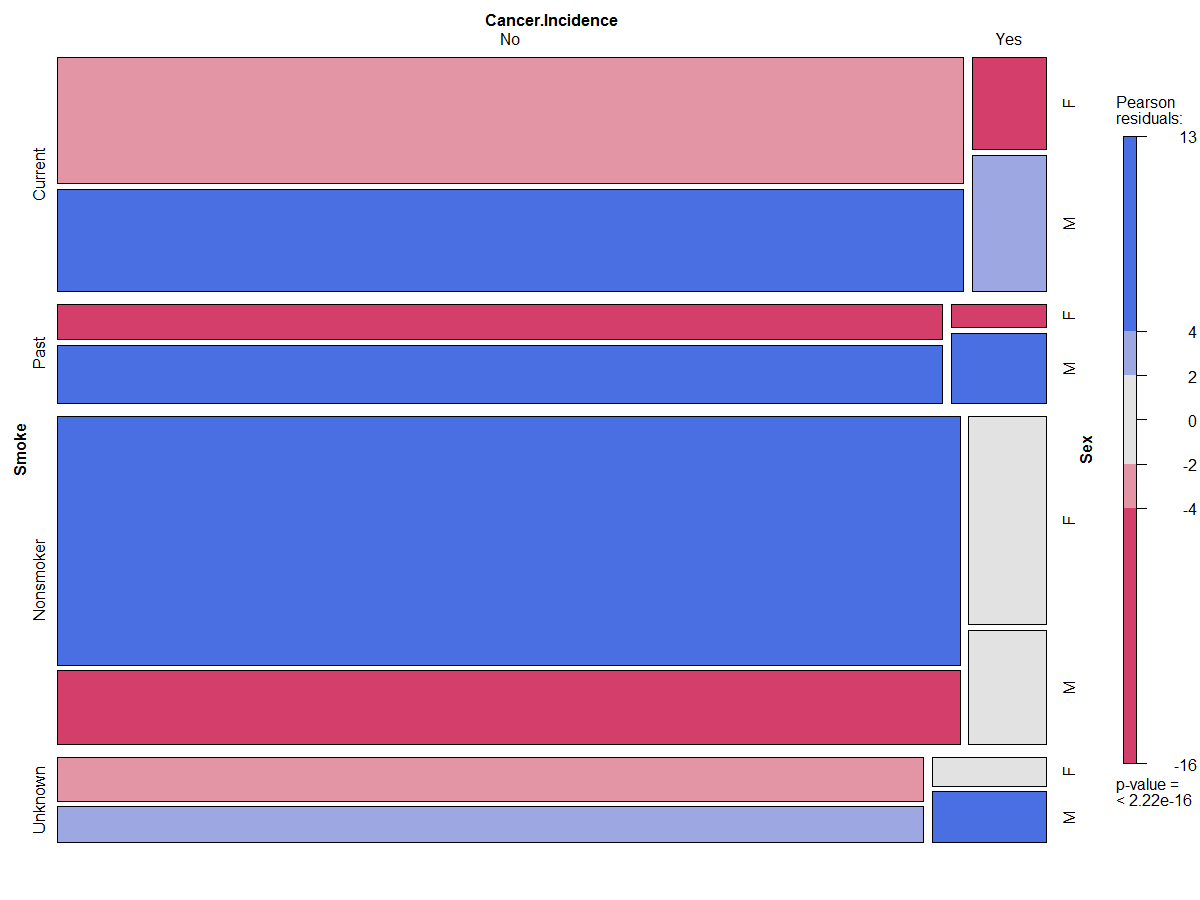




The plot shows the similar pattern of smoking on incidences of cancer. Therefore, race does not have effect of incidences of cancer.

Next we look at Sex variable:

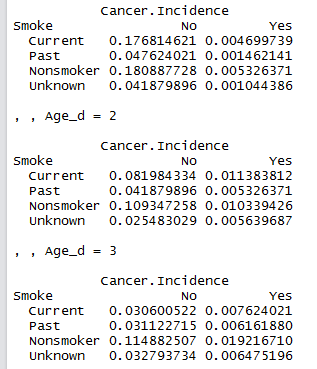


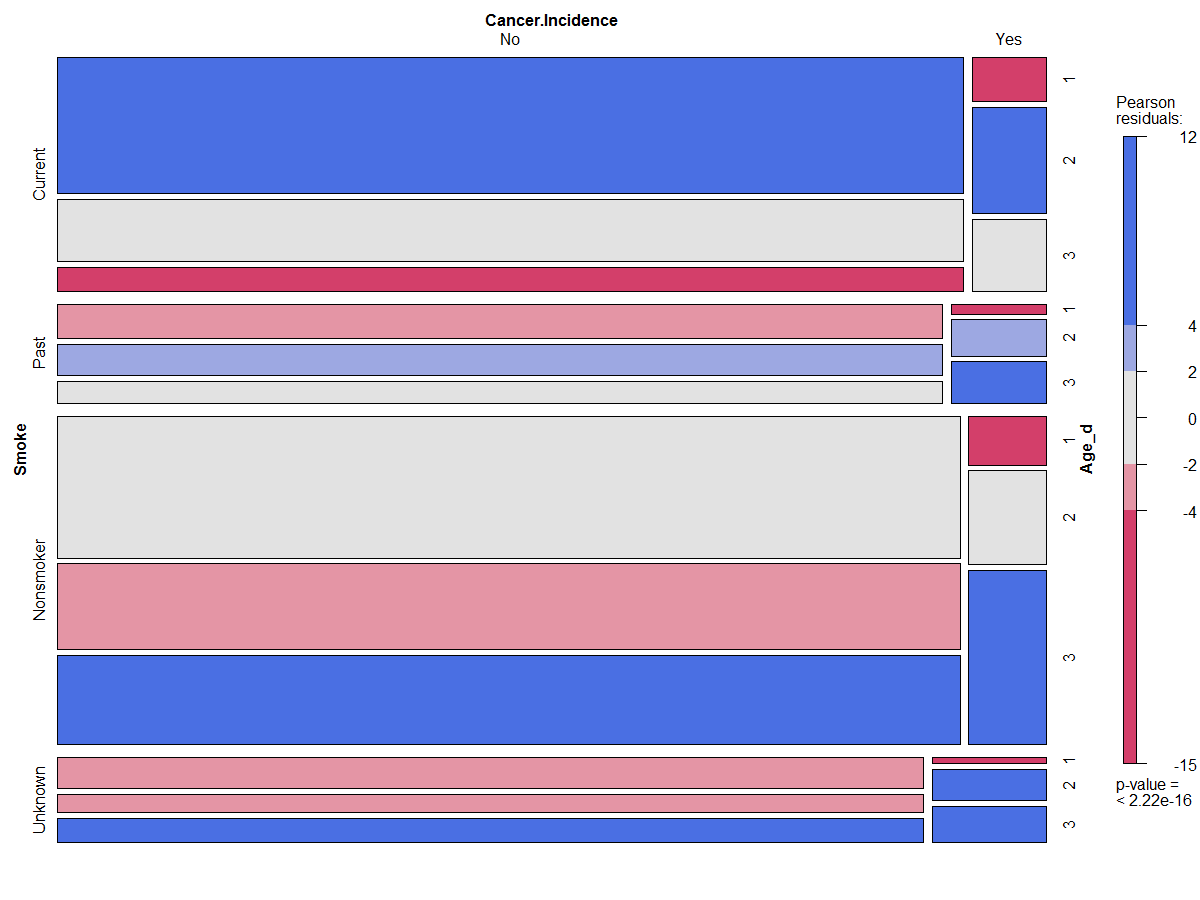


The plot shows the female and male have the same chance getting cancer. The pattern of smoking on incidences of cancer does not change. Therefore, gender does not have effect of incidences of cancer.

Finally, we look at the age group, I divide the age into 3 groups 20 to 45, 45 to 65 and 65-85

|  |
| --- |
| NHANES$Age\_d = cut(Age,c(20,45,65,85),c("1","2","3"))  data <- NHANES  attach(data)  mytable <- table(Smoke, Cancer.Incidence, Age\_d)  ptable = prop.table(mytable)  ptable  mosaic(mytable, shade=TRUE, legend=TRUE) |





Once again the plot shows the different age groups have similar chances getting cancer. The pattern of smoking on incidences of cancer does not change. Therefore, age groups do not have effect of incidences of cancer.