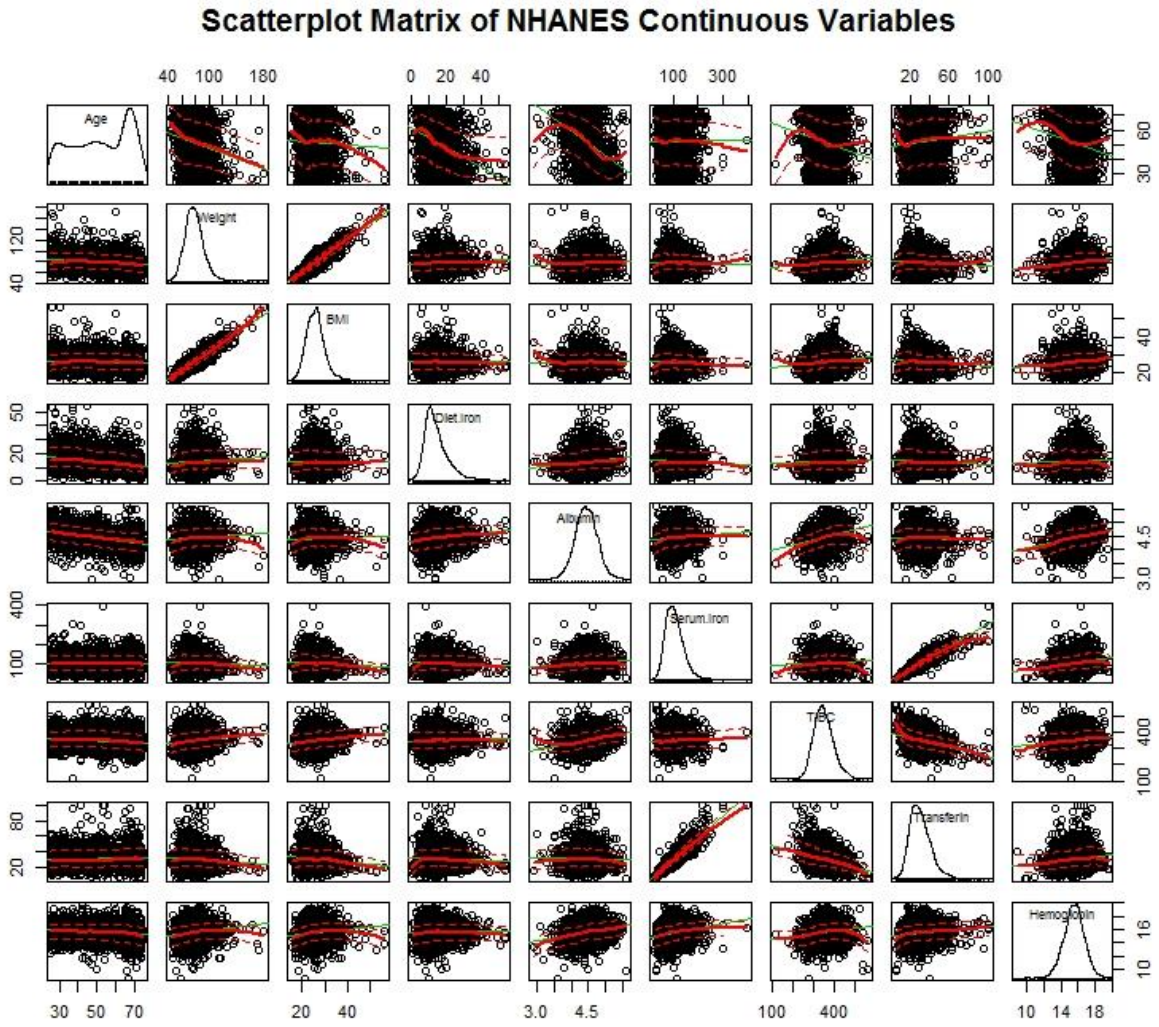
KATHLEEN ZHEN

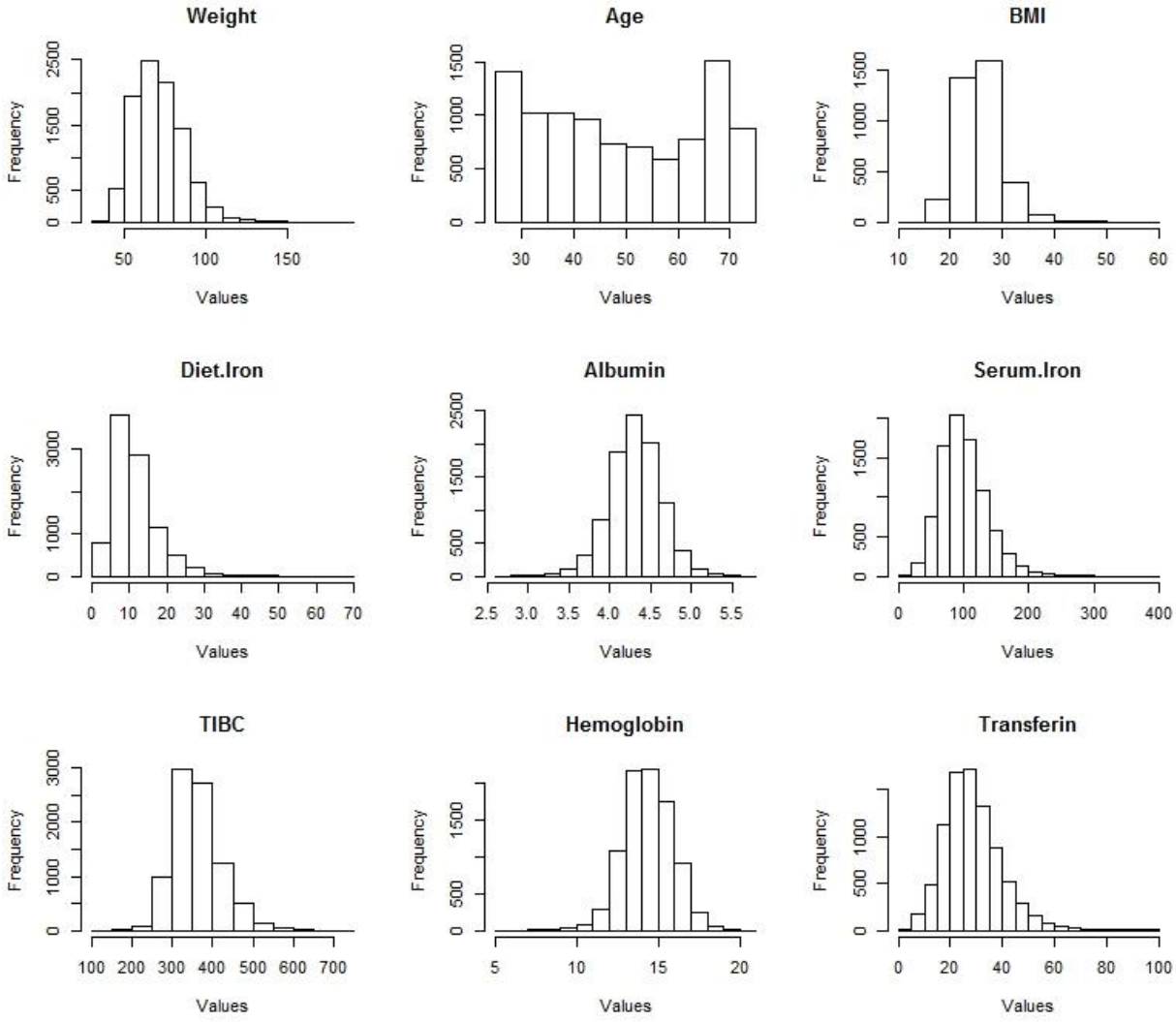
STA 141 HW 2

999210972

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

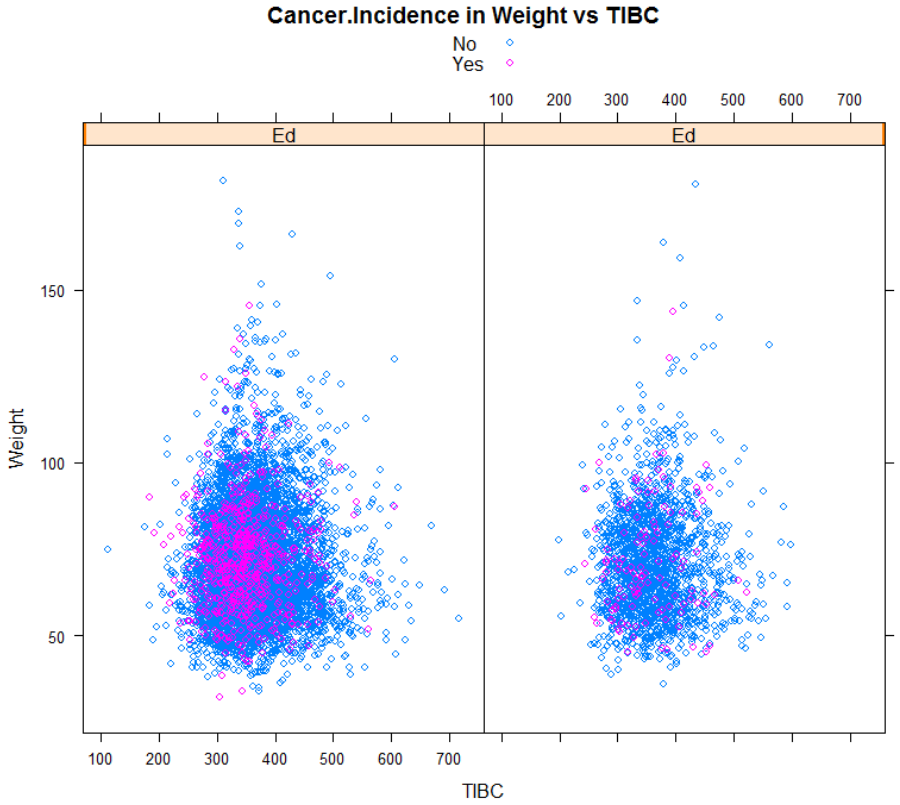
PART 1: LOOKING INTO THE CONTINUOUS VARIABLES

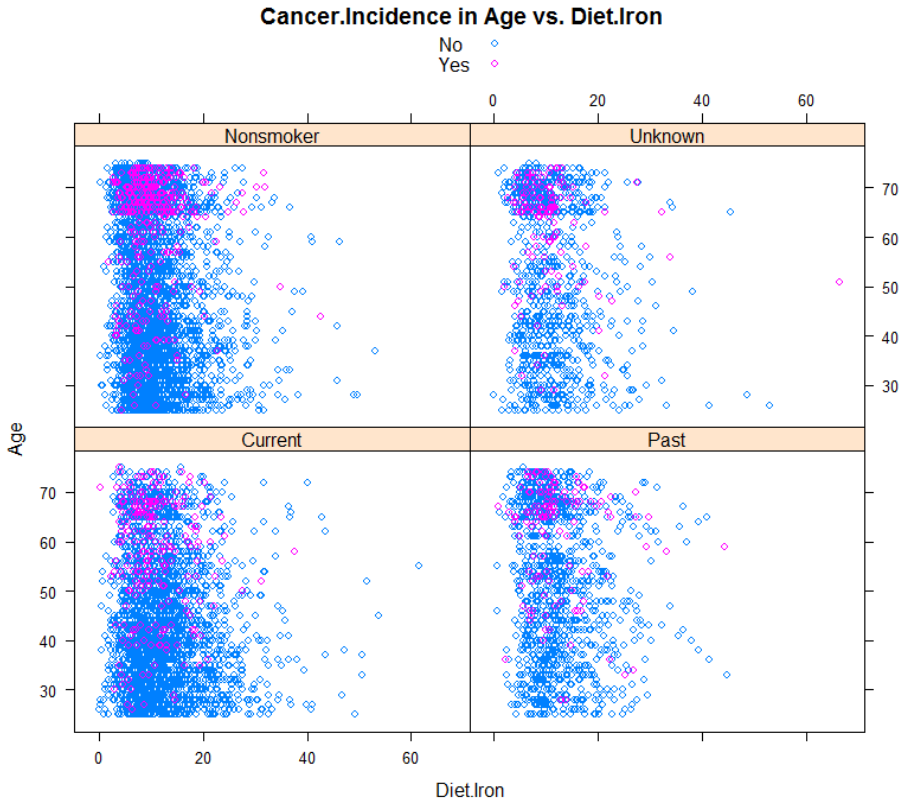


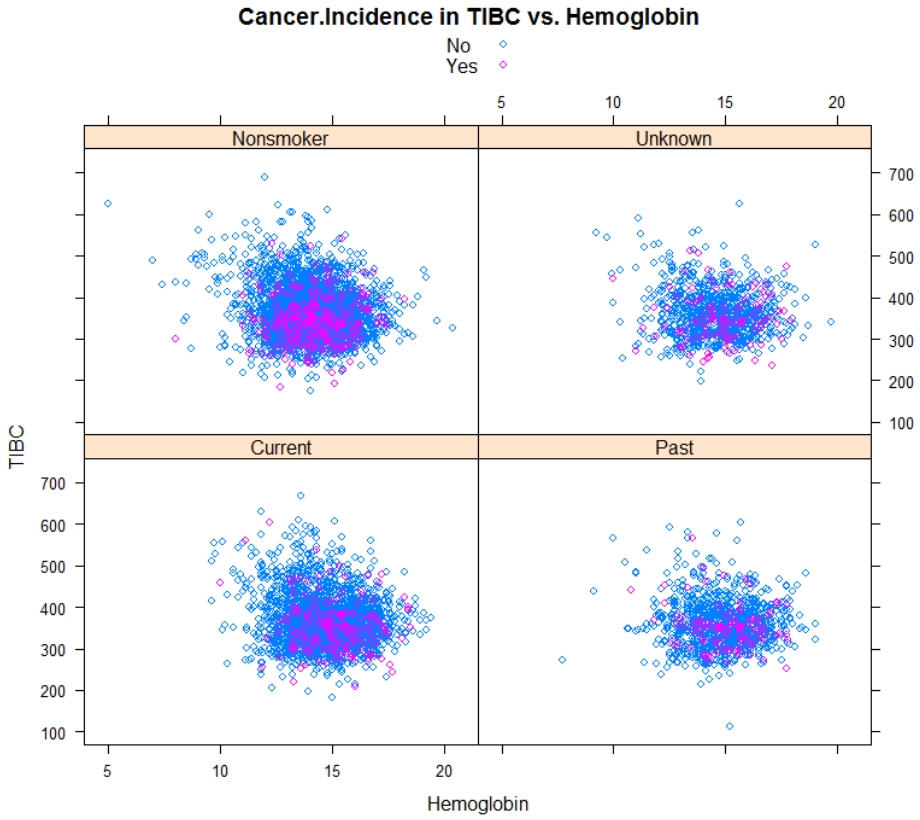


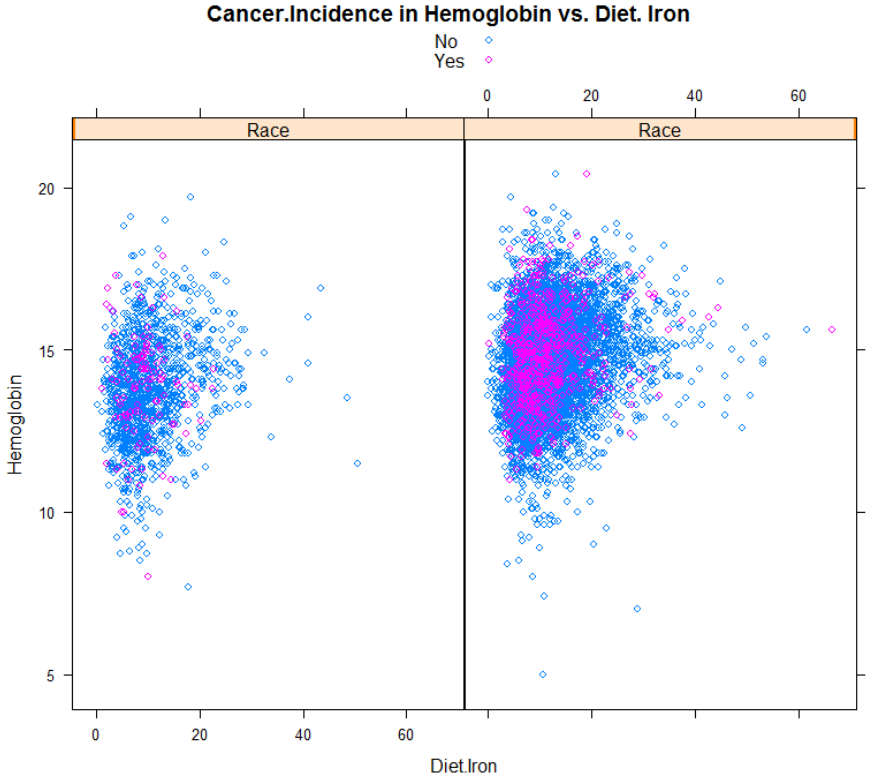
In the continuous variables in NHANES, I did a scatterplot matrix with the line of best fit through the data. The dotted lines represent the upper and lower bounds of the line of best fit. Looking at the data, most of the continue variables are in clutters when plotted against each other. The two pairs that are positively linear are Age vs. Weight and Serum.Iron vs Transferin. Looking at the histograms for each continuous variable, we can see that all the variables have a centralized value that is more dense. All the variables besides Age have this effect since the data is from all ages.

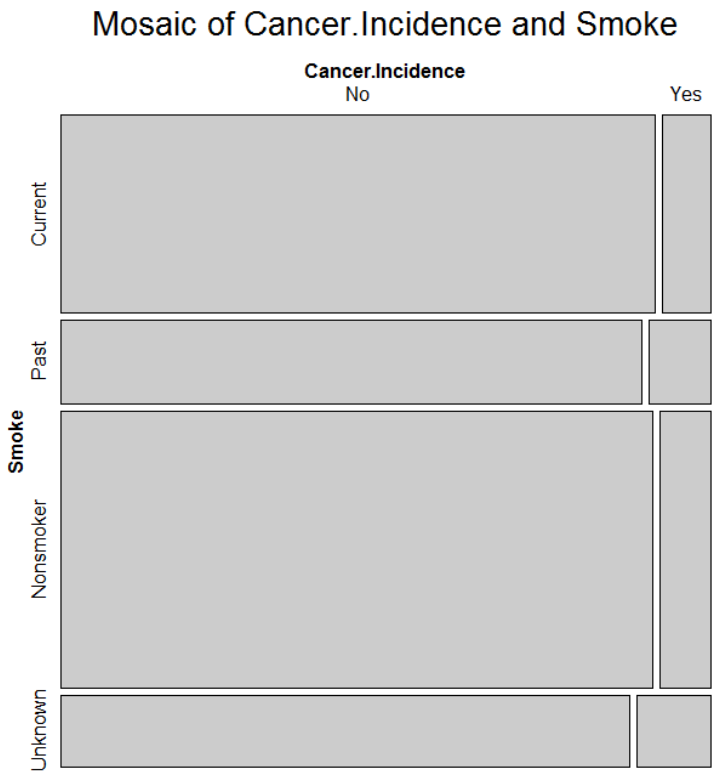
PART 2: LOOKING INTO NATURAL CLUSTERS IN THE DATA IN RELATION TO CATEGORICAL VARIABLES

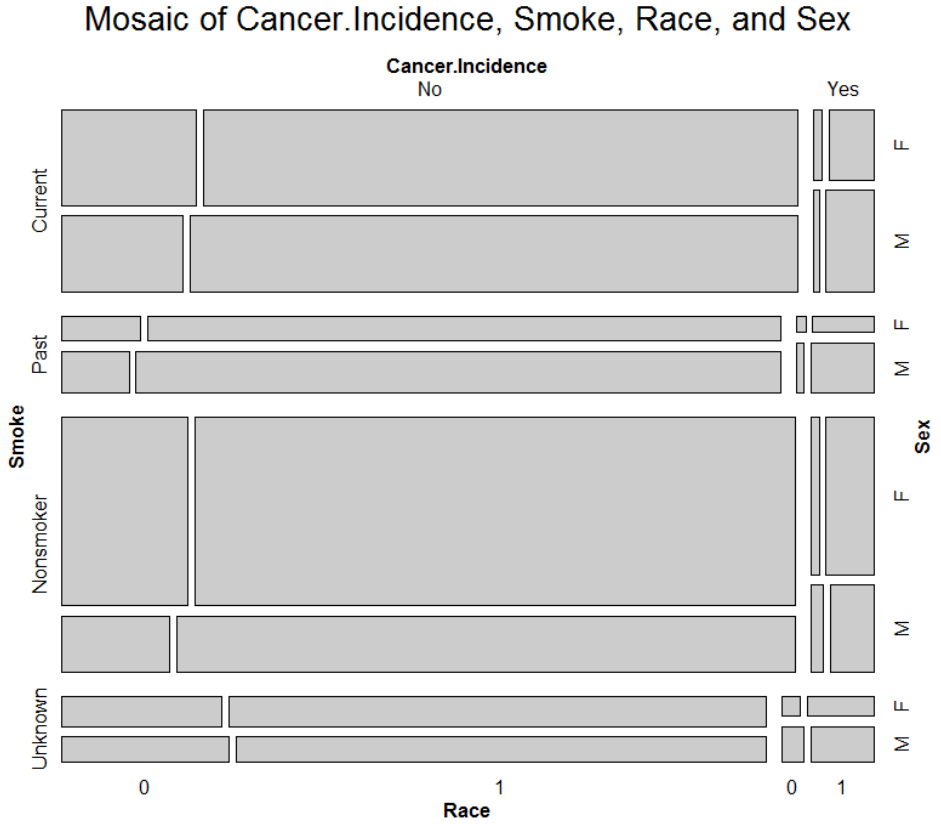

Cancer.Incidence in Weight vs TIBC
No
Yes
Ed
Ed
Weight
TIBC
100
200
300
400
500
600
700
50
100
150


Cancer.Incidence in Age vs. Diet.Iron
No
Yes
Nonsmoker
Unknown
Current
Past
Age
Diet.Iron
0
20
40
60
30
40
50
60
70

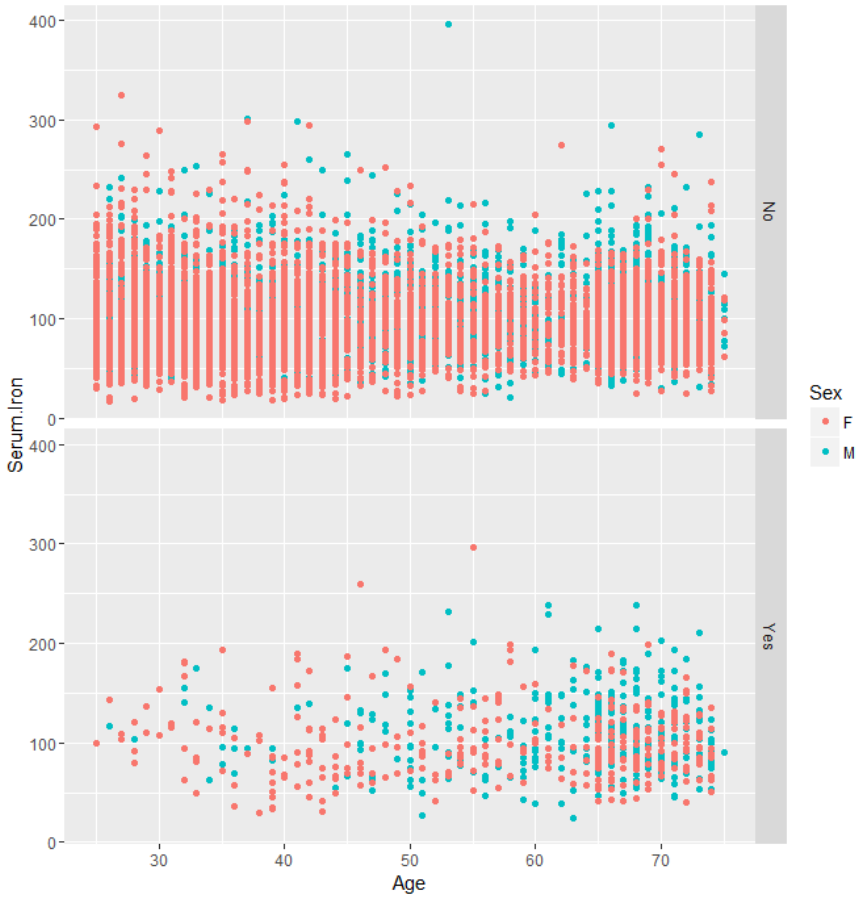

Cancer.Incidence in TIBC vs. Hemoglobin
No
Yes
Nonsmoker
Unknown
Current
Past
TIBC
Hemoglobin


Cancer.Incidence in Hemoglobin vs. Diet. Iron
No
Yes
Race
Race
Hemoglobin
Diet.Iron


Mosaic of Cancer.Incidence and Smoke
Cancer.Incidence
No
Yes
Smoke
Current
Past
Nonsmoker
Unknown

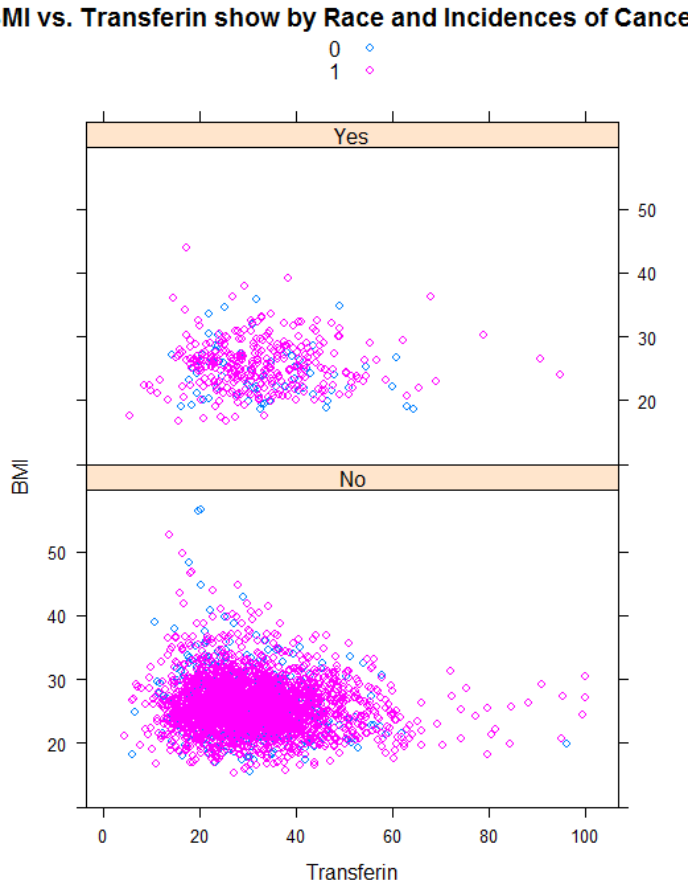

Mosaic of Cancer.Incidence, Smoke, Race, and Sex
Cancer.Incidence
No
Yes
Smoke
Current
Past
Nonsmoker
Unknown
Sex
F
M
Race
0
1


Albumin
5
4
3
No
Yes
30
40
50
60
70
Age
Sex
F
M

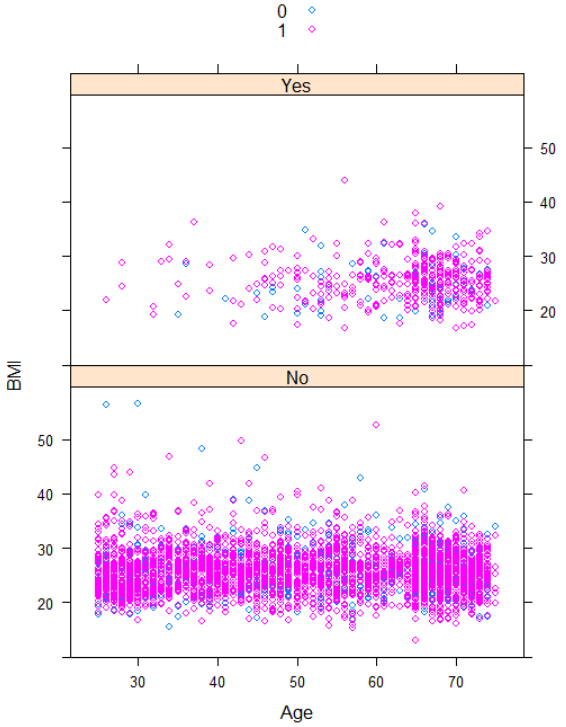

Serum.Iron
400
300
200
100
0
No
Yes
30
40
50
60
70
Age
Sex
F
M

In this section, I plotted various continuous variables against each other while it focuses on categorical variables. I will use Cancer.Incidence for most of the plots because I want to find if the continuous variables affects incidences of cancer. Looking at the various plots, one common trend that occurs is cancer is found more commonly in older ages. Most people intake roughly the same amount of Diet.Iron and cancer is most commonly found in nonsmokers at older ages. Another common thing that I notice is that all the categorical variables are skewed to one side. The side that is more dense generally has more incidences of cancer as shown in Weight vs TBIC and Hemoglobin vs Diet.Iron, just to name a few. In the breakdown of smokers, nonsmokers, past smokers, and unknown, the most data comes from nonsmokers. In proportion to the data, nonsmokers also have the most incidences of cancer. Female Caucasians appear to have more incidences of cancer. As shown by Albumin vs. Age and Serum.Iron vs Age, older female ages have more incidences of cancer.

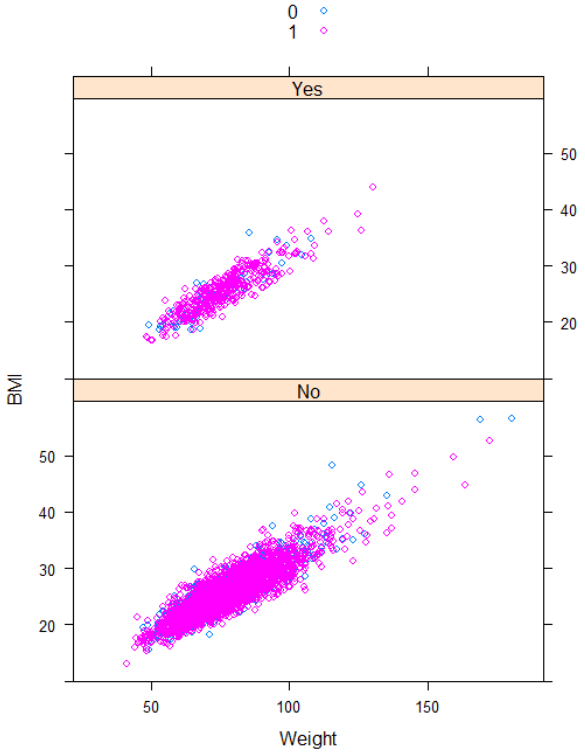
PART 3: VARIABLES THAT AFFECT BMI

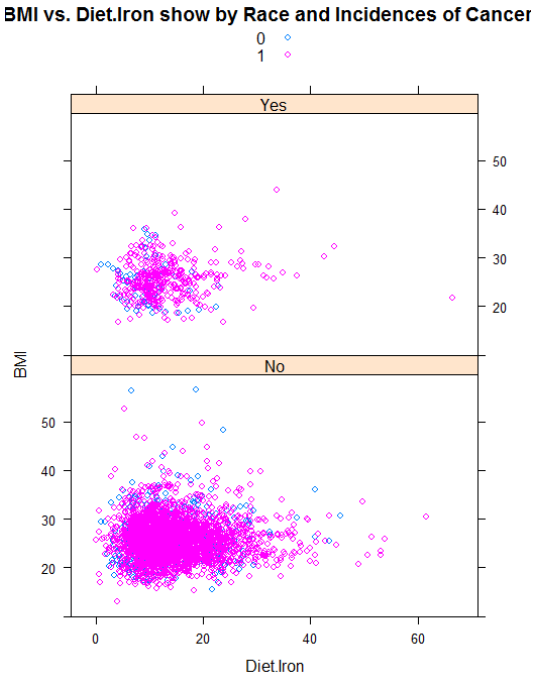


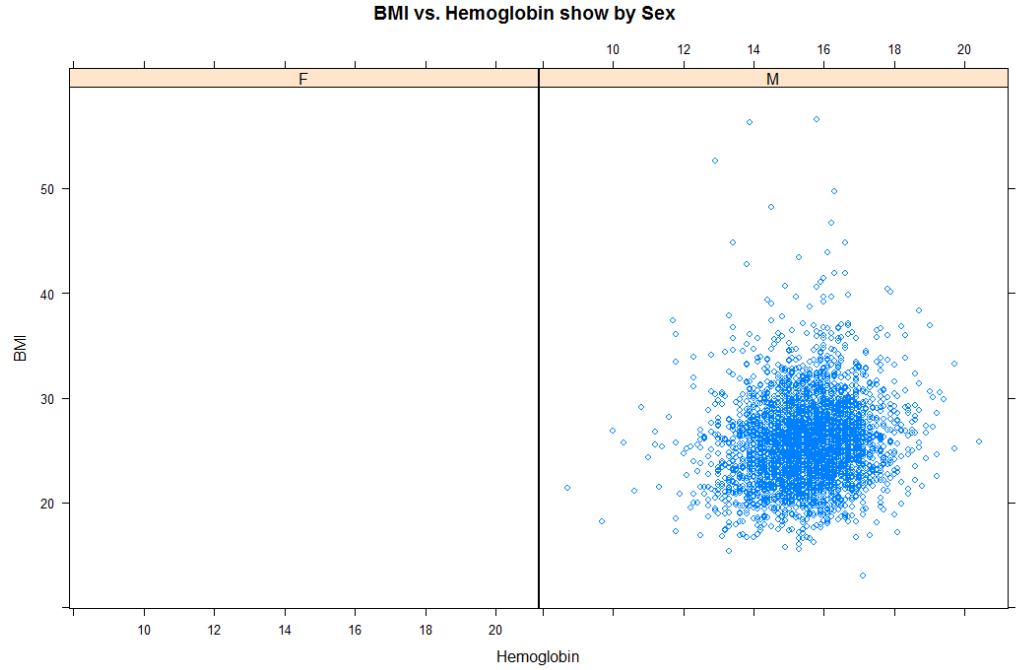
BMI vs. Age show by Race and Incidences of Cancer



BMI vs. Weight show by Race and Incidences of Cancer







When I am looking at how other continuous variables affect BMI, I also showed their classification by race and incidences of cancer. I reported some of the continuous variables above because the other graphs report the same thing. Most of the variables are in clutters except BMI vs. Weight which means weight should affect BMI the most. When I look into the classifications by Sex, I noticed there are no females data for BMI. Since there are no female data, I want to use all the continuous variables to predict BMI terms. I did a linear regression on the continuous variable with the y variable being BMI.

```
Residuals:
    Min      1Q  Median      3Q     Max
-7.1518 -1.2211 -0.0911  1.1278 12.8751

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.4257077  0.8895702   1.603  0.10911
Age          0.0304795  0.0025187  12.101  < 2e-16 ***
Weight       0.2644897  0.0024608 107.482  < 2e-16 ***
Diet.Iron   -0.0254019  0.0053033  -4.790 1.75e-06 ***
Albumin     -0.0002927  0.1158053  -0.003  0.99798
Serum.Iron  -0.0134904  0.0053656  -2.514  0.01198 *
TIBC         0.0057457  0.0017064   3.367  0.00077 ***
Transferin   0.0323225  0.0180754   1.788  0.07385 .
Hemoglobin   0.0605786  0.0294835   2.055  0.04000 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.821 on 2886 degrees of freedom
  (6680 observations deleted due to missingness)
Multiple R-squared:  0.8075,    Adjusted R-squared:  0.807
F-statistic:  1514 on 8 and 2886 DF,  p-value: < 2.2e-16
```

I get the following equation:

BMI = 1.4257077 + 0.0304(Age) + 0.2644(Weight) – 0.025(Diet.Iron) – 0.0002(Albumin) – 0.134(Serum.Iron) + 0.0057(TIBC) + 0.0323(Transferin) + 0.0605(Hemoglobin).

Then I tested to see if any of these variables affect the predictions. If they don’t affect it, it is safe to remove. I did the testing using a stepwise regression.

```
         Age Weight Diet.Iron Albumin Serum.Iron TIBC Transferin Hemoglobin
1  ( 1 ) " " "*"    " "       " "     " "        " "  " "        " "
2  ( 1 ) "*" "*"    " "       " "     " "        " "  " "        " "
3  ( 1 ) "*" "*"    "*"       " "     " "        " "  " "        " "
4  ( 1 ) "*" "*"    "*"       " "     " "        " "  "*"        " "
5  ( 1 ) "*" "*"    "*"       " "     "*"        "*"  " "        " "
6  ( 1 ) "*" "*"    "*"       " "     "*"        "*"  " "        "*"
7  ( 1 ) "*" "*"    "*"       " "     "*"        "*"  "*"        "*"
8  ( 1 ) "*" "*"    "*"       "*"     "*"        "*"  "*"        "*"
```

	R^2
1	0.7901516
2	0.8033375
3	0.8047414
4	0.8060639
5	0.8067235

6	0.8069211
7	0.8070682*
8	0.8070013

Looking at the R^2 table, number 7 has the highest value. Going back to the table above, row 7 means we remove Albumin. Our resulting regression is BMI = 1.4257077 + 0.0304(Age) + 0.2644(Weight) – 0.025(Diet.Iron) – 0.134(Serum.Iron) + 0.0057(TIBC) + 0.0323(Transferin) + 0.0605(Hemoglobin).

APPENDIX

```
install.packages("hexbin")
install.packages("car")
library("hexbin")
library("tidyverse")
library("car")
#load("~/R/STA 141A/HW2/NHANES.Rdata")

data("NHANES")

##PART 1
pairs(~Age + Weight + BMI + Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin + Hemoglobin, data =
NHANES, main = "Scatterplot Matrix of NHANES Continuous Variables")

scatterplotMatrix(~Age + Weight + BMI + Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin +
Hemoglobin, data = NHANES, main = "Scatterplot Matrix of NHANES Continuous Variables")

qplot(Weight, Diet.Iron, data = NHANES, colour = factor(Ed), main = "hi") + geom_smooth()

par(mfrow = c(3,3))
hist(NHANES$Weight, xlab = "Values", ylab = "Frequency", main = "Weight")
hist(NHANES$Age, xlab = "Values", ylab = "Frequency", main = "Age")
hist(NHANES$BMI, xlab = "Values", ylab = "Frequency", main = "BMI")
hist(NHANES$Diet.Iron, xlab = "Values", ylab = "Frequency", main = "Diet.Iron")
hist(NHANES$Albumin, xlab = "Values", ylab = "Frequency", main = "Albumin")
hist(NHANES$Serum.Iron, xlab = "Values", ylab = "Frequency", main = "Serum.Iron")
hist(NHANES$TIBC, xlab = "Values", ylab = "Frequency", main = "TIBC")
hist(NHANES$Hemoglobin, xlab = "Values", ylab = "Frequency", main = "Hemoglobin")
hist(NHANES$Transferin, xlab = "Values", ylab = "Frequency", main = "Transferin")

#PART 2
library("lattice")
xyplot(Age ~ Diet.Iron | Smoke, groups = Cancer.Incidence, type = 'p', auto.key = TRUE, data = NHANES,
main = "Cancer.Incidence in Age vs. Diet.Iron") #keep
xyplot(Weight ~ TIBC | Ed, groups = Cancer.Incidence, type = 'p', auto.key = TRUE, data = NHANES, main
= "Cancer.Incidence in Weight vs TIBC") #keep
xyplot(TIBC ~ Hemoglobin | Smoke, groups = Cancer.Incidence, type = 'p', auto.key = TRUE, data =
NHANES, main = "Cancer.Incidence in TIBC vs. Hemoglobin") #keep
xyplot(Hemoglobin ~ Diet.Iron | Race, groups = Cancer.Incidence, type = 'p', auto.key = TRUE, data =
NHANES, main = "Cancer.Incidence in Hemoglobin vs. Diet. Iron") #keep
xyplot(Serum.Iron ~ TIBC | Ed, groups = Cancer.Incidence, type = 'p', auto.key = TRUE, data = NHANES)
#keep
xyplot(Weight ~ Hemoglobin | Smoke, groups = Sex, type = 'p', auto.key = TRUE, data = NHANES) #keep

#PART 3
install.packages('vcd')
library('vcd')
```

```
xyplot(BMI ~ Transferin | Cancer.Incidence, groups = Smoke, type = 'p', auto.key = T, data = NHANES)
mosaic(~Smoke + Cancer.Incidence, data = NHANES, main = "Mosaic of Cancer.Incidence and Smoke")

#PART 4
mosaic(~Smoke + Cancer.Incidence + Sex + Race, data = NHANES, main = "Mosaic of Cancer.Incidence,
Smoke, Race, and Sex")

#PART 5
plotr = ggplot(NHANES, aes(Age, Weight, color = Cancer.Incidence))
plotr+geom_point()
plotr+geom_point() + facet_grid(Cancer.Incidence ~.)

plotr = ggplot(NHANES, aes(Age, Serum.Iron, color = Sex))
plotr+geom_point()
plotr+geom_point() + facet_grid(Cancer.Incidence ~.)

plotr = ggplot(NHANES, aes(Age, Diet.Iron, color = Cancer.Incidence))
plotr+geom_point()
plotr+geom_point() + facet_grid(Cancer.Incidence ~.)

plotr = ggplot(NHANES, aes(Age, Albumin, color = Sex))
plotr+geom_point()
plotr+geom_point() + facet_grid(Cancer.Incidence ~.)

#PART 6
xyplot(BMI ~ Transferin | Cancer.Incidence, groups = Race, type = 'p', auto.key = T, data = NHANES,
main= "BMI vs. Transferin show by Race and Incidences of Cancer")
xyplot(BMI ~ Age | Cancer.Incidence, groups = Race, type = 'p', auto.key = T, data = NHANES, main=
"BMI vs. Age show by Race and Incidences of Cancer")
xyplot(BMI ~ Weight | Cancer.Incidence, groups = Race, type = 'p', auto.key = T, data = NHANES, main=
"BMI vs. Weight show by Race and Incidences of Cancer")
xyplot(BMI ~ Diet.Iron | Cancer.Incidence, groups = Race, type = 'p', auto.key = T, data = NHANES, main=
"BMI vs. Diet.Iron show by Race and Incidences of Cancer")
xyplot(BMI ~ Albumin | Cancer.Incidence, groups = Race, type = 'p', auto.key = T, data = NHANES, main=
"BMI vs. Albumin show by Race and Incidences of Cancer")
xyplot(BMI ~ Serum.Iron | Cancer.Incidence, groups = Race, type = 'p', auto.key = T, data = NHANES,
main= "BMI vs. Serum.Iron show by Race and Incidences of Cancer")
xyplot(BMI ~ TIBC | Cancer.Incidence, groups = Race, type = 'p', auto.key = T, data = NHANES, main=
"BMI vs. TIBC show by Race and Incidences of Cancer")
xyplot(BMI ~ Hemoglobin | Cancer.Incidence, groups = Race, type = 'p', auto.key = T, data = NHANES,
main= "BMI vs. Hemoglobin show by Race and Incidences of Cancer")

xyplot(BMI ~ Hemoglobin | Sex, type = 'p', auto.key = T, data = NHANES, main= "BMI vs. Hemoglobin
show by Sex")

#linear regression
fit = lm(BMI ~ Age + Weight + Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin + Hemoglobin, data =
NHANES)
```

```
summary(fit)
coefficients(fit)
fitted(fit)

install.packages('leaps')
library('leaps')
step(fit)
summary(regsubsets(BMI ~ Age + Weight + Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin +
Hemoglobin, data = NHANES, nbest=1, nvmax = 12))
summary(regsubsets(BMI ~ Age + Weight + Diet.Iron + Albumin + Serum.Iron + TIBC + Transferin +
Hemoglobin, data = NHANES, nbest=1, nvmax = 12))$adjr2
```