STA 141 Homework 2

Jiewei Chen (999 494 235)

Honor Code: "The codes and results derived by using these codes constitute my own work. I have consulted the following resources regarding this assignment: Da Xue, Lingjun Ma, Zhihao Li.

1 Set up and Basic Analysis

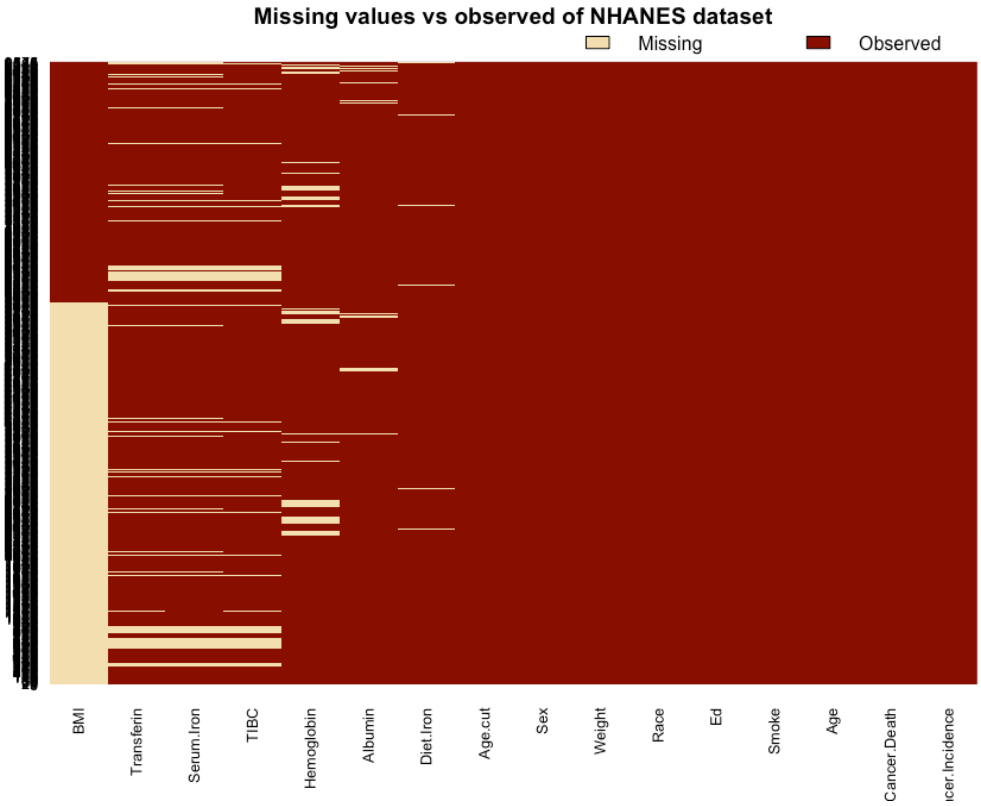
1.1 Load the dataset

```
# Load Dataset
load("NHANES.Rdata")
attach(NHANES)
```

This NHANES data set provides information of Cancer Incidence and Cancer Death of about 9575 people, as well as their Smoke status, Education, Race, Sex, Age, Weight, BMI and several relevant health index. The main purpose is to find the relation between each variable and Cancer, so that probably a model can be built to predict Cancer Incidence and Cancer Death. Thus, people can pay more attention to important factors that might cause cancer in their daily life.

1.2 Check for missing values

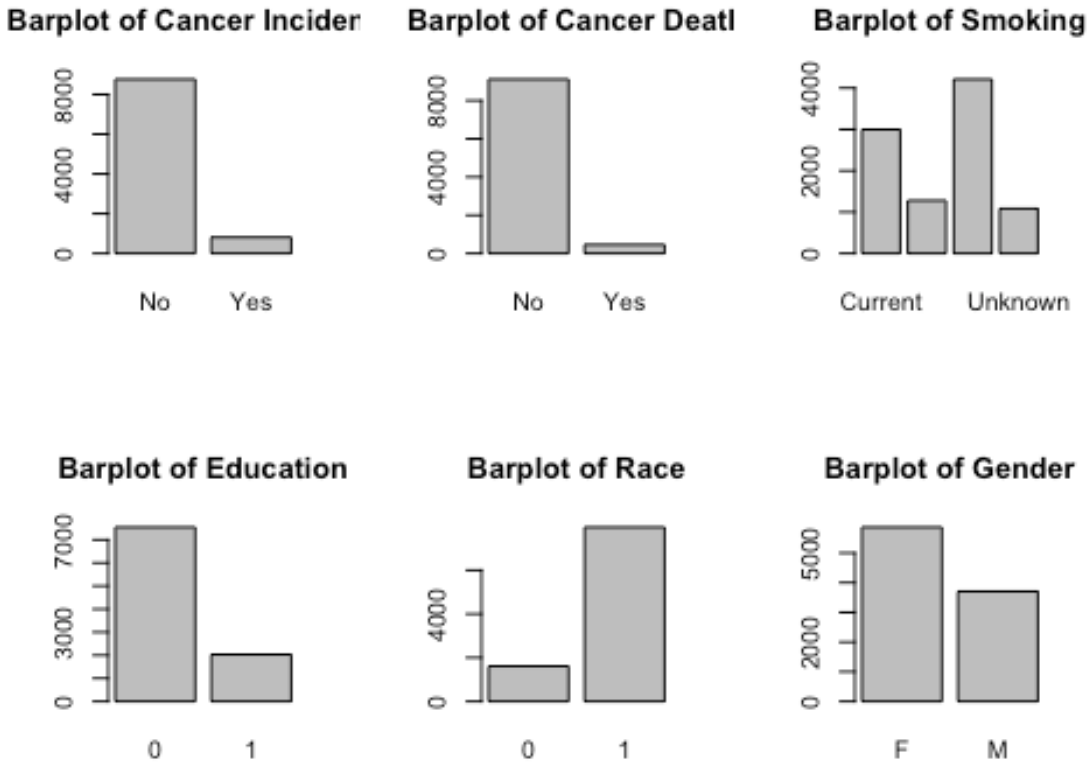
```
# install.packages("Amelia")
library(Amelia)
missmap(NHANES, main = "Missing values vs observed of NHANES dataset")
```



In the "Missin Values vs. Observed of NHANES dataset" graph, it can be observed that there are approximately 60% of missing values in BMI category. There are also some missing values in "Transfererin", "Serum.Iron", "TIBC", "Hemoglobin", "Albumin", and "Diet.Iron". So in the following analysis of the data, it is necessary to be cautious about what kind value is missing in BMI category.

1.3 Count for each factor variable

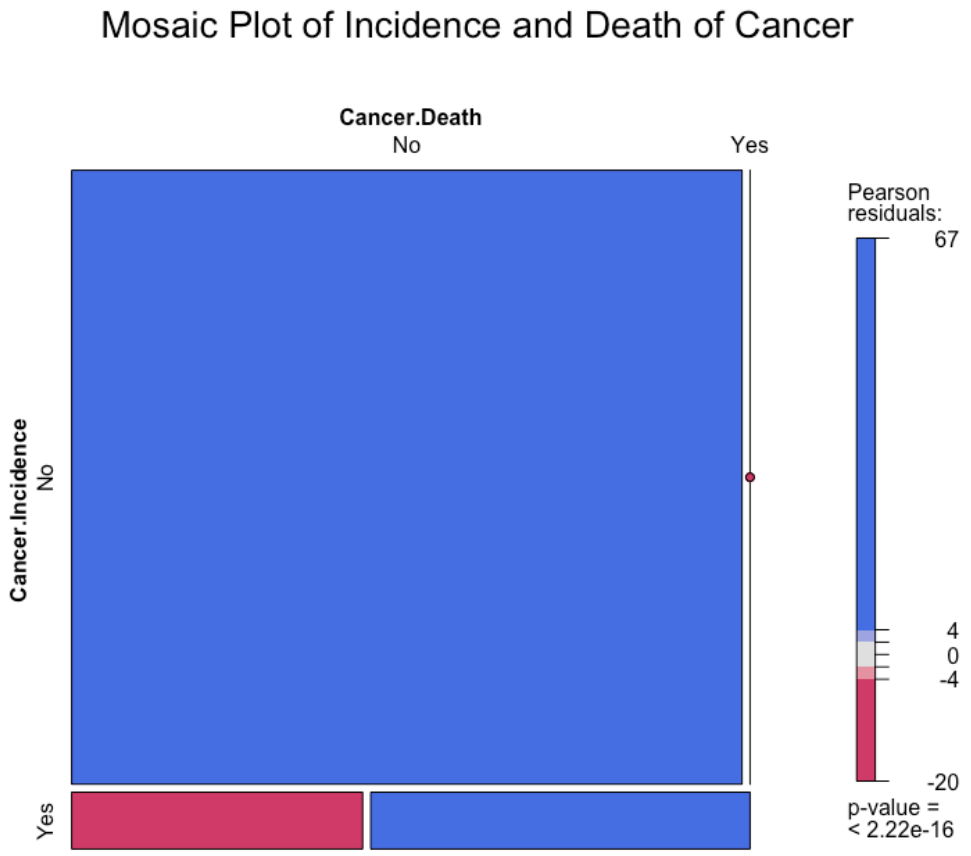
```
par(mfrow=c(2,3))
barplot(table(NHANES$Cancer.Incidence), main = "Barplot of Cancer Incidence")
barplot(table(NHANES$Cancer.Death), main = "Barplot of Cancer Death")
barplot(table(NHANES$Smoke), main = "Barplot of Smoking")
barplot(table(NHANES$Ed), main = "Barplot of Education")
barplot(table(NHANES$Race), main = "Barplot of Race")
barplot(table(NHANES$Sex), main = "Barplot of Gender")
```



Note here that, from the Ed result, the people who have "College Degree" and "Non-College Degree" are not evenly seperated. Neither are "Caucasian" and "Non-caucasian", which can be seen from Race result.

1.4 Relation b/t Cancer.Incidence and Cancer.Death

```
# install.packages("vcd")
library(vcd)
mosaic(~ Cancer.Incidence + Cancer.Death, data = NHANES, shade = TRUE, legend
 = TRUE, main = "Mosaic Plot of Incidence and Death of Cancer")
```



From this plot and calculation, it can be observed that people who are diagnosed with cancer are 8% of the population. Within these people who were diagnosed with cancer, about 43% was not died. However, since we don't know when these data were collected, ie. whether they were recorded before death or any other time, the cancer death category seems to be of little value in the following analysis.

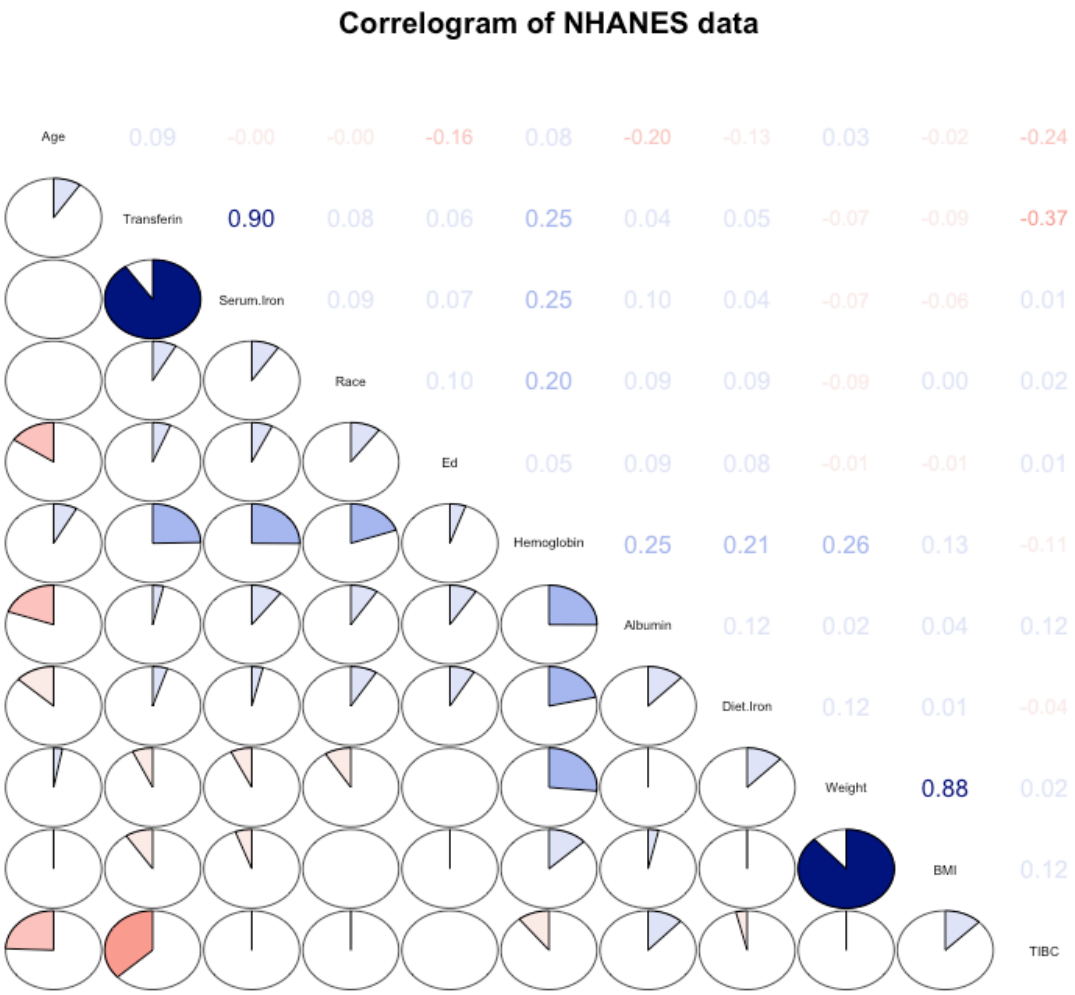
So in the following analysis, the goal is trying to use other categories to predict the incidence of cancer based on the information provided by this NHANES data set.

2 Analysis on Continous Variables

2.1 Correlation b/t Continuous Variables

```
# install.packages("corrgram")
# install.packages("diptest")
library(corrgram)
corrgram(NHANES, order = TRUE,
         lower.panel = panel.pie, upper.panel = panel.cor,
```

```
        text.panel = panel.txt,
        main = "Correlogram of NHANES data")
```

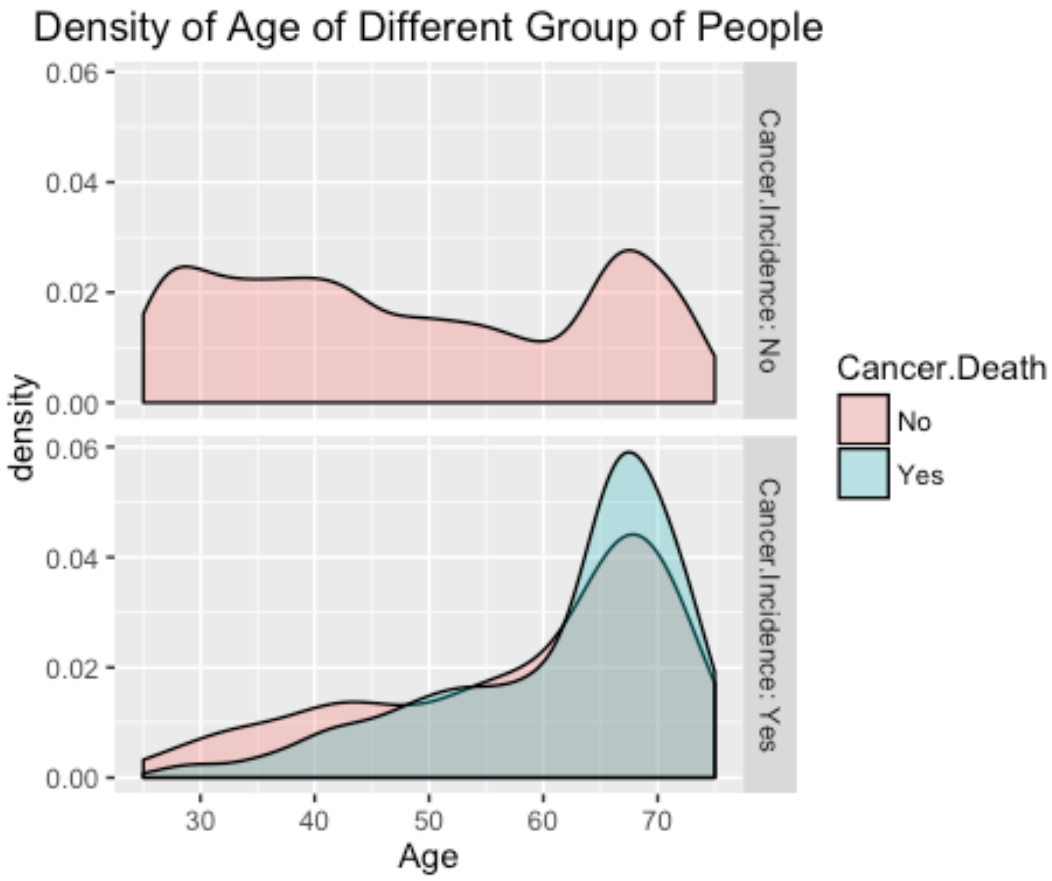


From the Correlogram of NHANES data, it can be observed that most of the variables don't have correlation. Strong correlation can be seen in BMI ~ Weight, Transferin ~ Serum.Iron and Transferin ~ TIBC pairs. Since Trasnferin is defined by "100 * Serum.Iron/TIBC" equation, the observation of high covariance among Serum.Iron, TIBC and Transferin is not surprising.

Weak correlation (Cov > 0.2) can be seen within the following pairs, Weight ~ Hemoglobin, Transferin ~ Hemoglobin, Diet.Iron ~ Hemoglobin, Albumin ~ Hemoglobin, Serum.Iorn ~ Hemoglobin, Alumbin ~ Age and TIBC ~ Age.

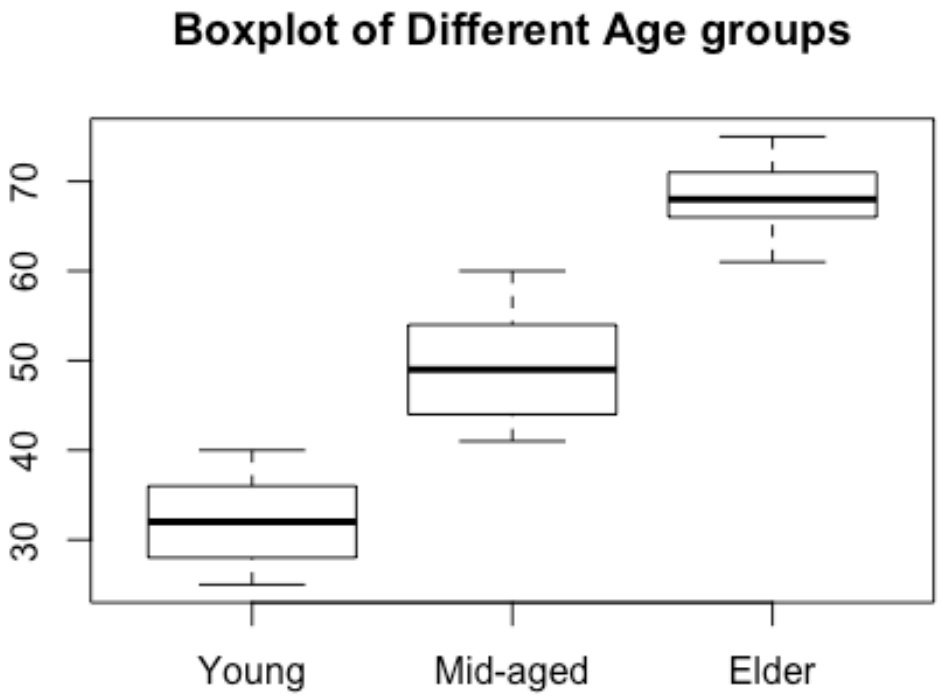
The relation between BMI ~ Weight and Transferin ~ Serum.Iron + TIBC will be examined in later part. ## 2.2 Is there any relation b/t Age and Cancer Incidence?

```
require(ggplot2)
plt = ggplot(NHANES, aes(Age, fill = Cancer.Death))
plt + geom_density(alpha = 0.3)+ facet_grid(Cancer.Incidence ~. , labeller =
label_both) + labs( title = "Density of Age of Different Group of People" )
```



It can be observed from this graph that the age distribution of people who are not diagnosed with cancer is quite different from that of people who are diagnosed with cancer. The distribution of age of cancer incidence reaches its peak at around 65 years old, indicating that elder people seems to have higher chance to get cancer. However, within the people who have cancer, there is not much difference in the distribution. So, age is a very important factor related to the prediction of whether a person will get cancer. Later age will be transfered into a categorical variable for further analysis.

```
NHANES$Age.cut = cut(NHANES$Age,breaks=c(25, 40, 60, 75), include.lowest=TRUE
, labels=c('Young','Mid-aged','Elder'))
boxplot(Age ~ NHANES$Age.cut, main = "Boxplot of Different Age groups")
```

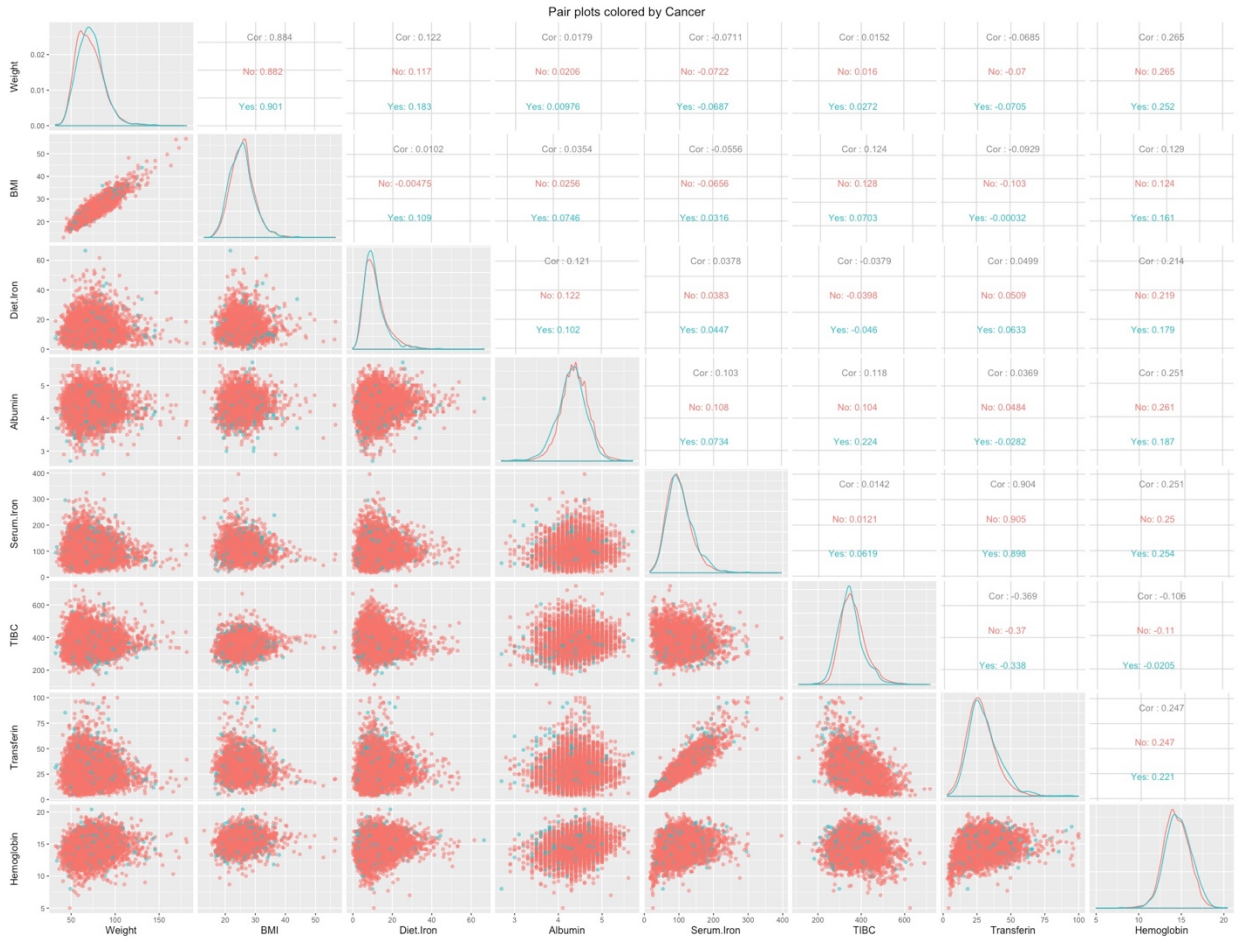


People are divided according to their age. Young, Mid-aged, and Elder correspond to [25,40], (40,60], and (60,75], respectively.

2.3 Is there any relation b/t Continuous Variables and Categorical Variables?

2.3.1 Continuous Variables vs. Cancer

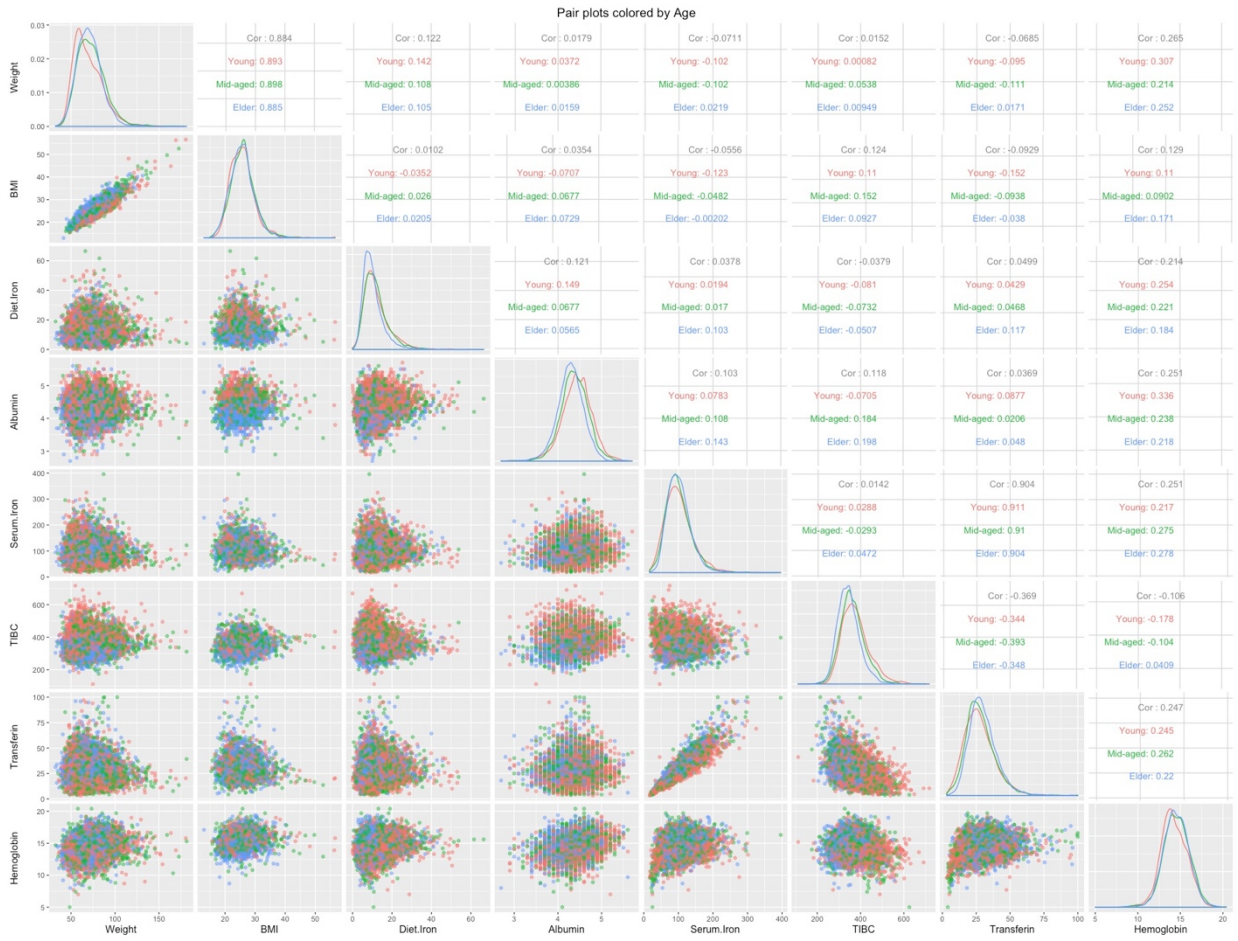
```
# install.packages("GGally")
require(GGally)
ggpairs(NHANES, mapping = aes(color = Cancer.Incidence, alpha = 0.5),
        columns = c(7:14),
        title = "Pair plots colored by Cancer",
        lower = list(continuous = wrap("points", alpha = 0.5)),
        diag = list(continuous = "density"),
        upper = list(continuous = "cor"))
```



From this graph, first, it can be seen that the density plot of each continuous variables in people who are diagnosed with cancer or not are almost laying on the top of each other, indicating that these health index have no direct influence on cancer. This can be further supported by the correlation value presented on the upper right. Most of the correlation between each health index of different groups are similar to the correlation of each health index of all people as a whole. There is no natural cluster here seen in this pair plots, indicating that these health indicators doesn't have significant relation with different groups of people with or without cancer. Increasing in the correlation between TIBC ~ Albumin, Diet.Iron ~ Weight can be observed after seperation by cancer.

2.3.2 Continuous Variables vs. Age Group

```
require(GGally)
ggpairs(NHANES, mapping = aes(color = Age.cut, alpha = 0.5),
        columns = c(7:14),
        title = "Pair plots colored by Age",
        lower = list(continuous = wrap("points", alpha = 0.5)),
        diag = list(continuous = "density"),
        upper = list(continuous = "cor"))
```



From this graph,

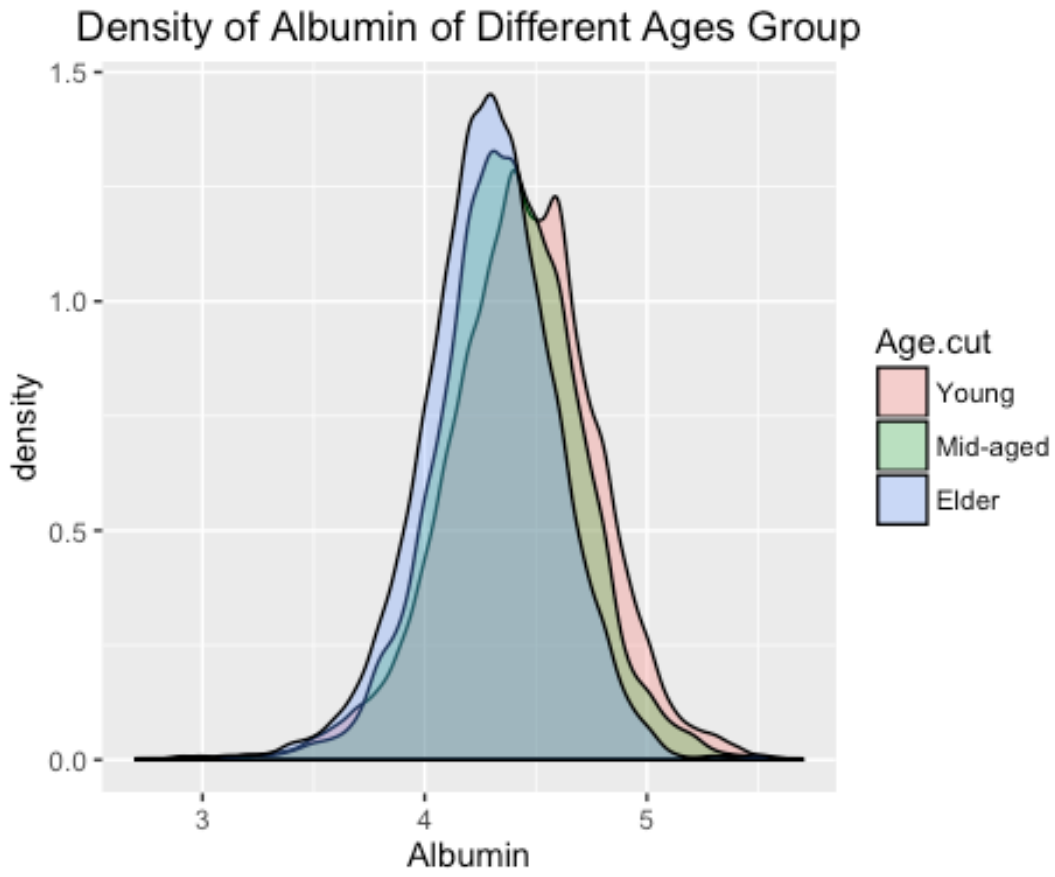
1) The weight of young people seems to be lower than the others. This can be seen from the density plot of weight. The peak of weight of young people shifts to the right comparing with the peak of mid-aged and elders.

2) People who is younger tend to have higher Albumin, however, the difference is not pronounced. This can be seen from the density plot of Albumin.

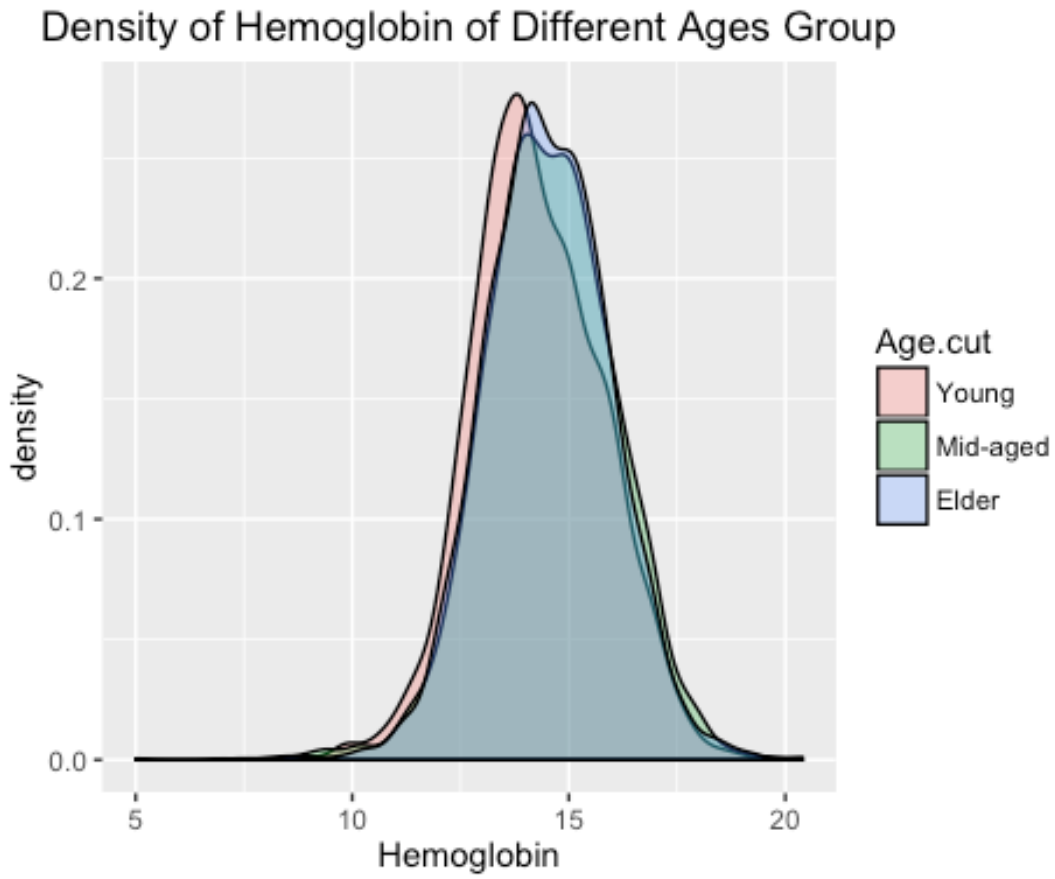
3) Slight difference can also be observed in Hemoglobin index. Younger people with age from 25 to 40 shows lower hemoglobin index, while people with 40+ age shows similar hemoglobin index.

4) For the rest health index, there seems not much relation between them and age. Below are more clear plots of Albumin and Hemoglobin in different age group.

```
require(ggplot2)
plt_Al_Age = ggplot(NHANES, aes(Albumin, fill = Age.cut))
plt_Al_Age + geom_density(alpha = 0.3) + labs( title = "Density of Albumin of
 Different Ages Group" )
```

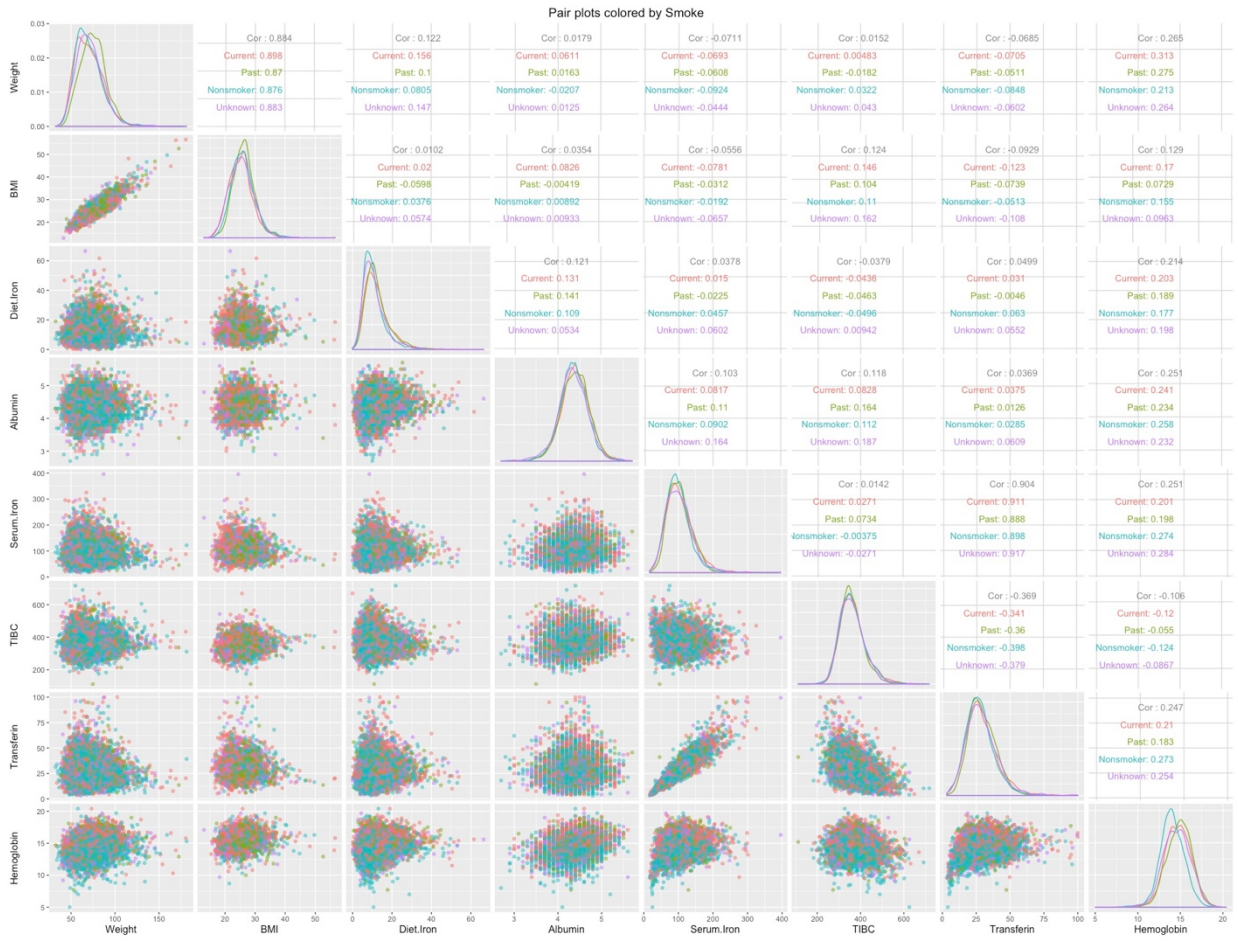



```
require(ggplot2)
plt_Al_Age = ggplot(NHANES, aes(Hemoglobin, fill = Age.cut))
plt_Al_Age + geom_density(alpha = 0.3) + labs( title = "Density of Hemoglobin
 of Different Ages Group" )
```



2.3.3 Continuous Variables vs. Smoke

```
require(GGally)
ggpairs(NHANES, mapping = aes(color = Smoke, alpha = 0.5),
        columns = c(7:14),
        title = "Pair plots colored by Smoke",
        lower = list(continuous = wrap("points", alpha = 0.5)),
        diag = list(continuous = "density"),
        upper = list(continuous = "cor"))
```



From this graph, it can be concluded that

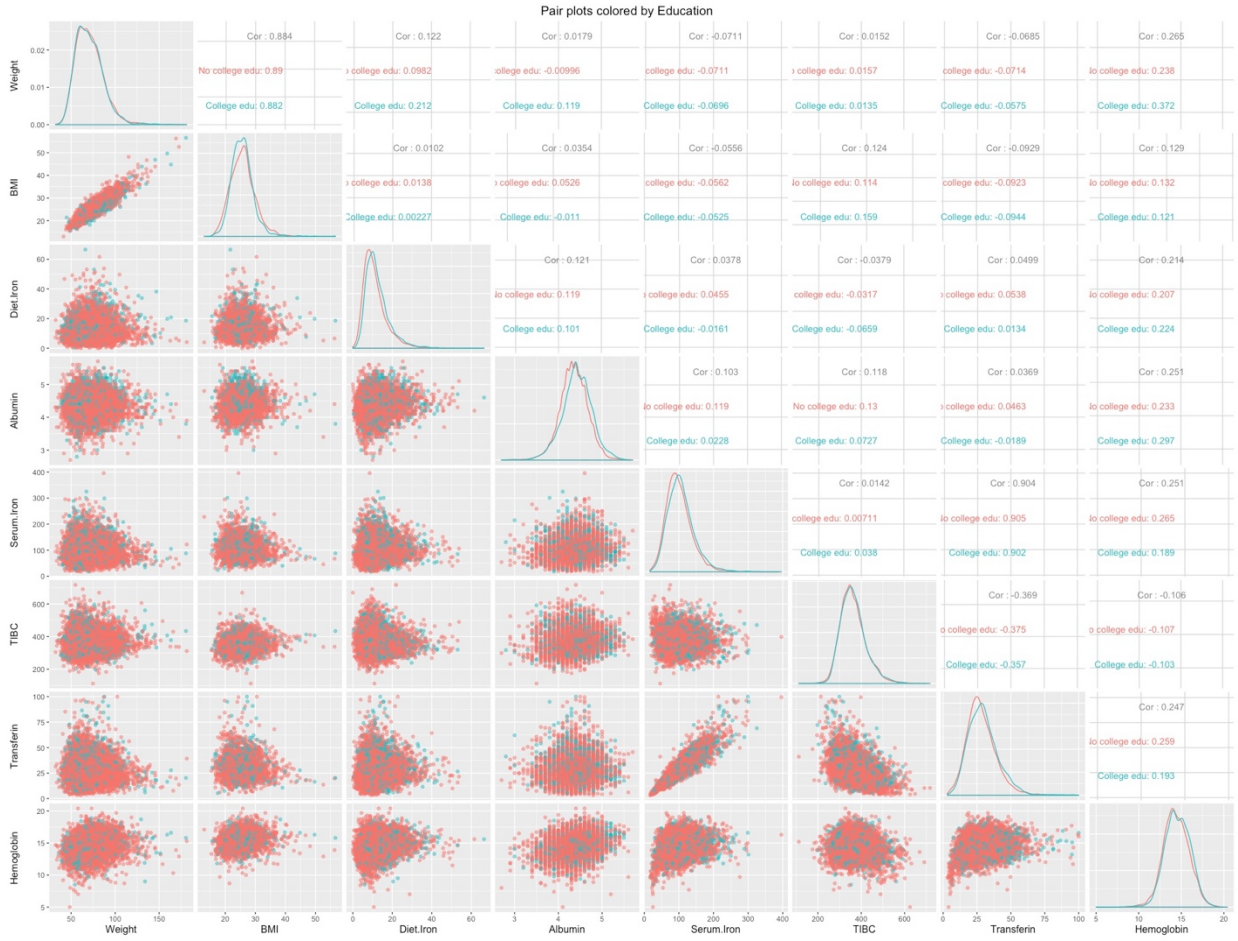
1) Smoking has a signficant influence on the Hemoglobin level of human. This can be concluded from the Hemoglobin density plot. People who smoke will have higher hemoglobin level. People who don't smoke have lower hemoglobin value. This finding have been confirmed by many scientific paper. One possible reason could be related to carbon monoxide. (REF: http://www.masimo.com/pdf/clinical/carboxyhemoglobin/light-carboxyhemoglobin-levels-in-smokers-nov-2007.pdf)

2) There is a slight difference in the Diet Iron index of different group of people. People who are currently smoking or are past smoker tend to have higher diet iron. Non-smoker have lower diet iron. But diet iron is a indicator of dietary intake of iron. So the relation between diet iron and smoking should be non-dependent.

2.3.4 Continuous Variables vs. Education

```
NHANES$Ed = factor(NHANES$Ed, levels = c(0, 1), labels = c("No college edu",
"College edu"))
require(GGally)
ggpairs(NHANES, mapping = aes(color = Ed, alpha = 0.5),
        columns = c(7:14),
```

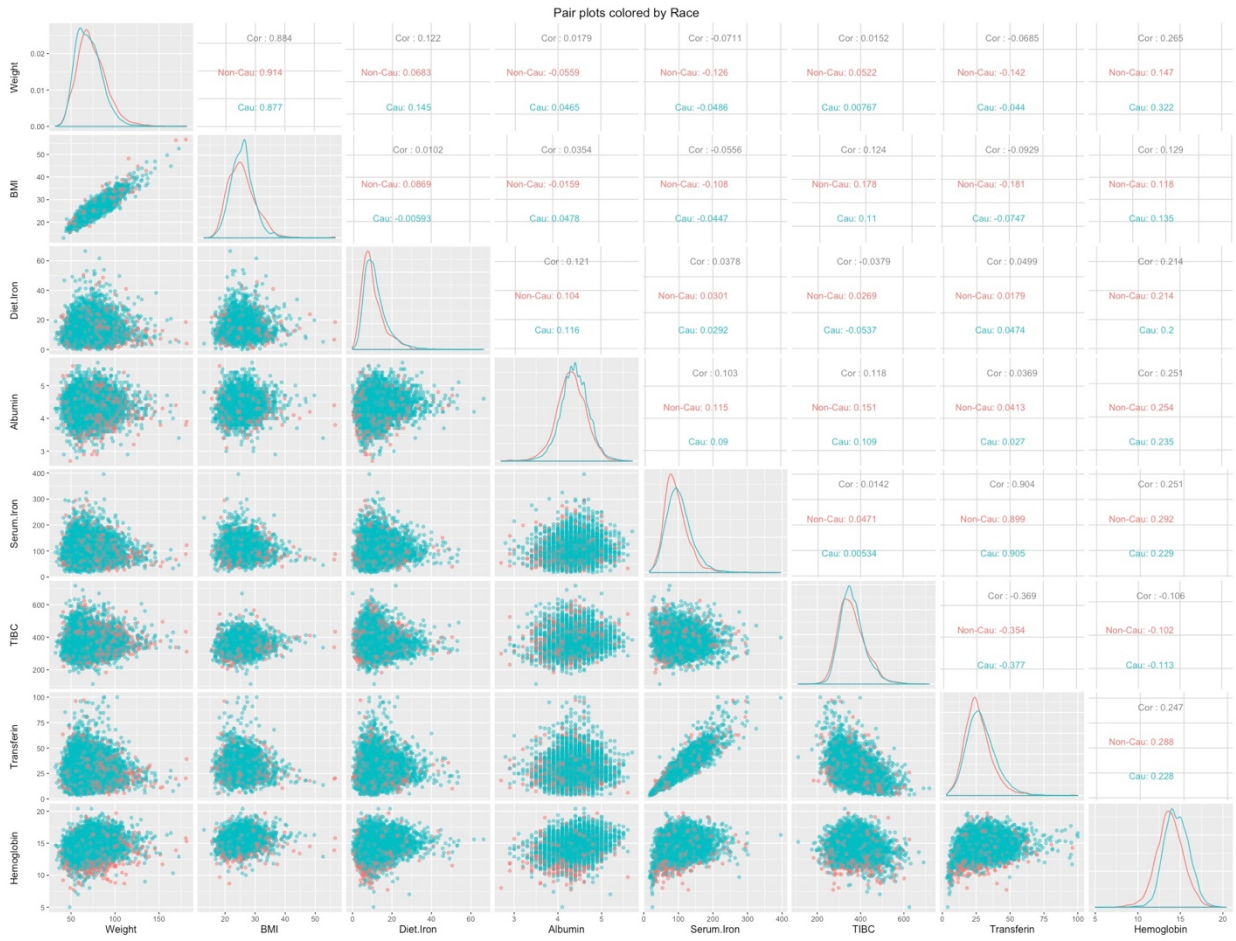
```
        title = "Pair plots colored by Education",
        lower = list(continuous = wrap("points", alpha = 0.5)),
        diag = list(continuous = "density"),
        upper = list(continuous = "cor"))
```



From this graph, it seems there doesn't exist significant relation between these health index and education. No natrual cluster can be observed from this graph.

2.3.5 Continuous Variables vs. Race

```
NHANES$Race = factor(NHANES$Race, levels = c(0, 1), labels = c("Non-Cau", "Ca
u"))
require(GGally)
ggpairs(NHANES, mapping = aes(color = Race, alpha = 0.5),
        columns = c(7:14),
        title = "Pair plots colored by Race",
        lower = list(continuous = wrap("points", alpha = 0.5)),
        diag = list(continuous = "density"),
        upper = list(continuous = "cor"))
```



From this graph, it can be concluded that

1) There is significant difference in the Hemoglobin level in caucasian, and non-caucasian. caucasian, in general have higher hemoglobin level.

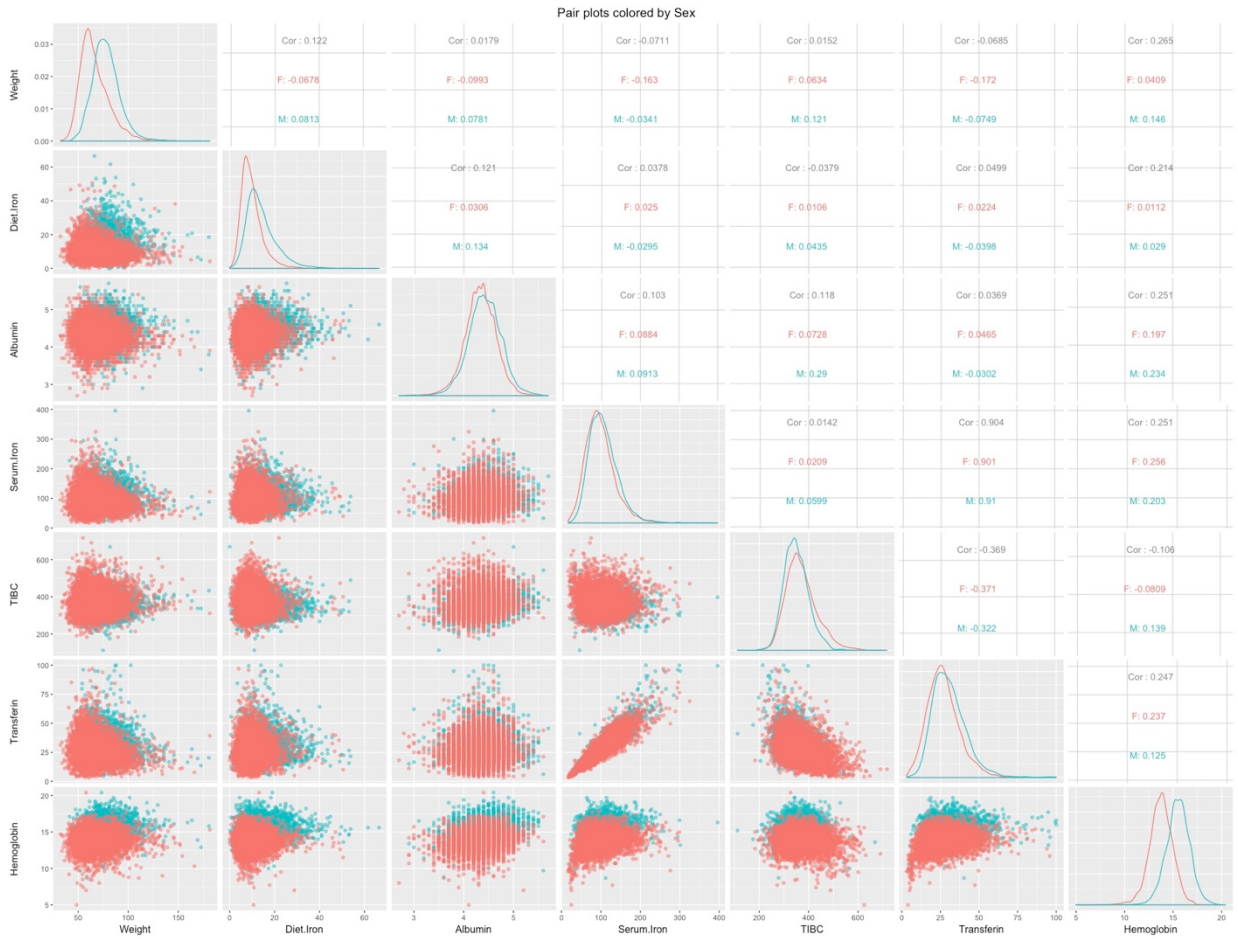
2) Slight difference in Serum Iron and Transferin level can be observed for different race. caucasian tend to have higher level in both Serum Iron and Transferin.

3) No other significant difference can be seen in other variables.

Note that the number of caucasian people in this dataset is much higher than that of non-caucasian people, so the conlusions above seems suspicious.

2.3.6 Continuous Variables vs. Sex

```
require(GGally)
ggpairs(NHANES, mapping = aes(color = Sex, alpha = 0.5),
        columns = c(7,9:14), # No BMI column
        title = "Pair plots colored by Sex",
        lower = list(continuous = wrap("points", alpha = 0.5)),
        diag = list(continuous = "density"),
        upper = list(continuous = "cor"))
```



From this pair plots colored by sex, it can be observed that

1) Men in general have higher weight than female. This can be seen from the density plot of weight, as well as from the natural cluster observed in plots of weight column

2) The dietary iron value is different for two genders, see the density plot of diet iron density.

3) Significant difference can be seen in the density plot of Hemoglobin for two gender. Clear natural cluster can be seen in all plots related to Hemoglobin. In general, men have higher Hemoglobin level than women.

2.4 Highlight of Findings

A brief summary on these pair plots colored by different categorical variables:

1. Plots colored by age, smoking, and gender show significant difference in several plots, especially hemoglobin level.

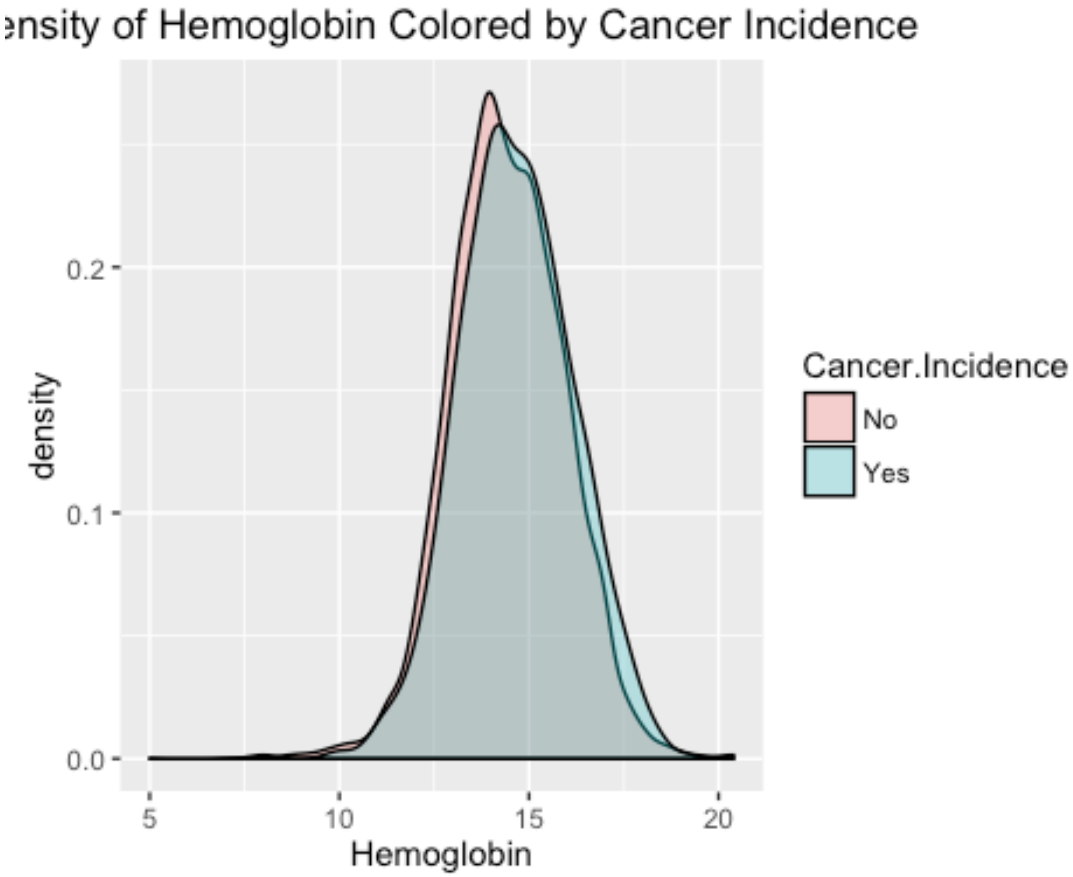
2. Significant influence of smoking status on Hemoglobin can be observed. Influence of race, sex and age on hemoglobin are shown as well. Although the influence is not that significant from age.

3. Non-pronouced influence of age on Albumin index can be observed.
4. Smoke and sex can have a slight impact on Diet iron value.
5. Serum Iron and Transferin have a weak relation with race.

2.5 More Analysis on Hemoglobin

2.5.1 Hemoglobin vs. Cancer

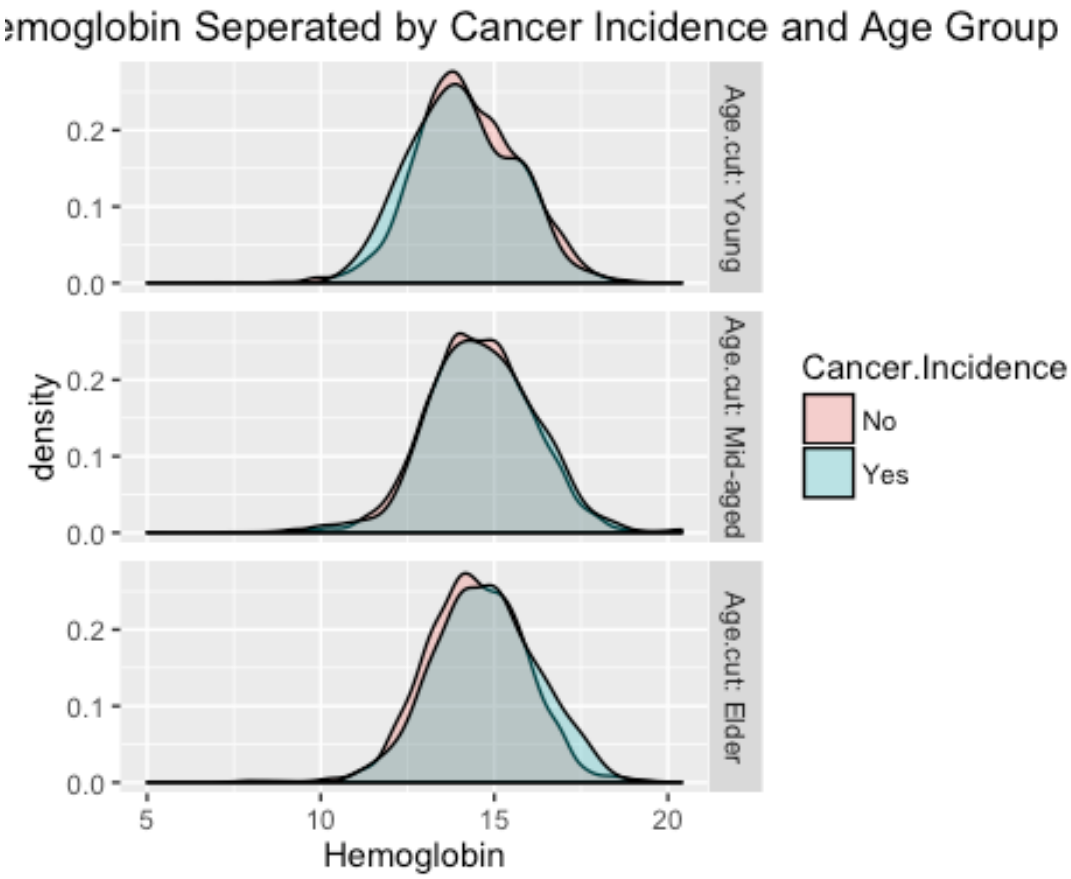
```
require(ggplot2)
plt = ggplot(NHANES, aes(Hemoglobin, fill = Cancer.Incidence))
plt + geom_density(alpha = 0.3) + labs( title = "Density of Hemoglobin Colore
d by Cancer Incidence" )
```



Recall that in the previous analysis, it is observed that for people who have cancer, the hemoglobin level is higher.

2.5.1.2 Does Hemoglobin ~ Cancer Relation Hold Across Age

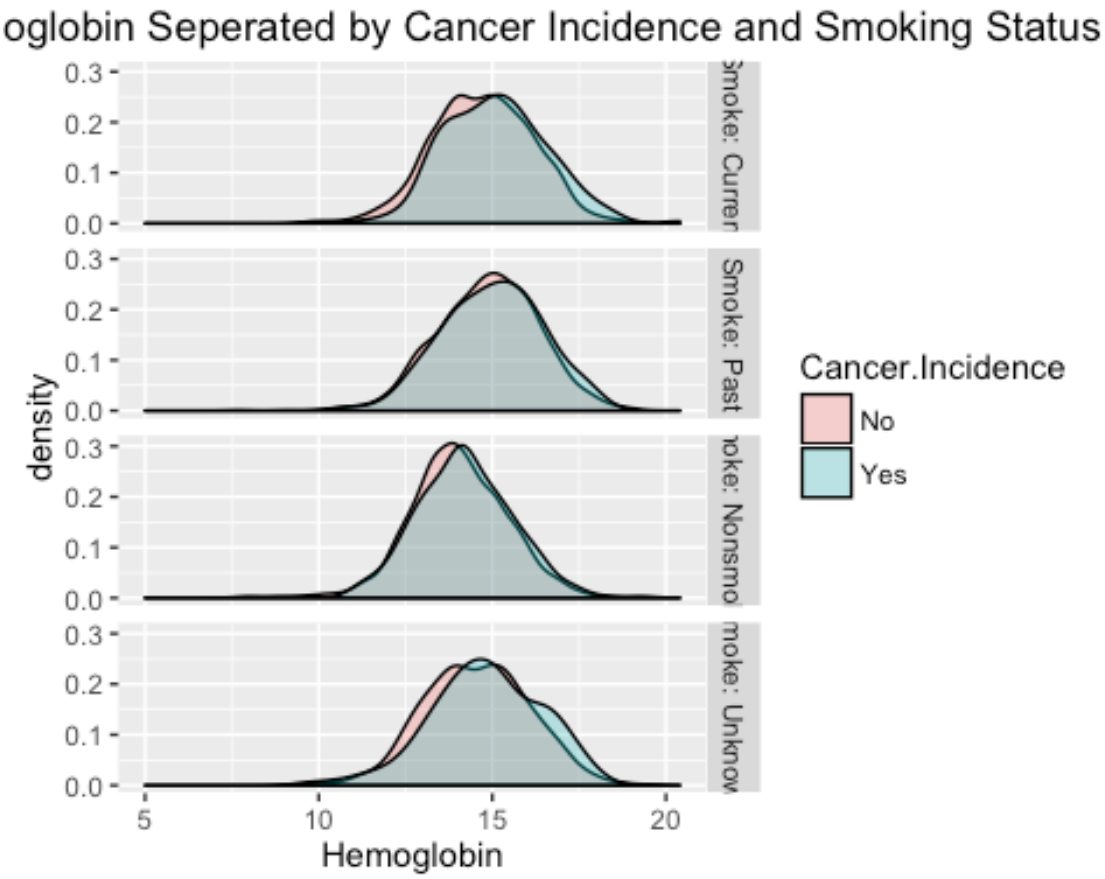
```
require(ggplot2)
plt = ggplot(NHANES, aes(Hemoglobin, fill = Cancer.Incidence))
plt + geom_density(alpha = 0.3)+ facet_grid(Age.cut ~ . , labeller = label_bo
th) + labs( title = "Density of Hemoglobin Seperated by Cancer Incidence and
Age Group" )
```



From this Hemoglobin vs. Age and Gender plot, no difference in the density plot can be observed in each part, indicating that the relation between hemoglobin and cancer incidence is not dependent on age.

2.5.1.3 Does Hemoglobin ~ Cancer Relation Hold Across Smoke

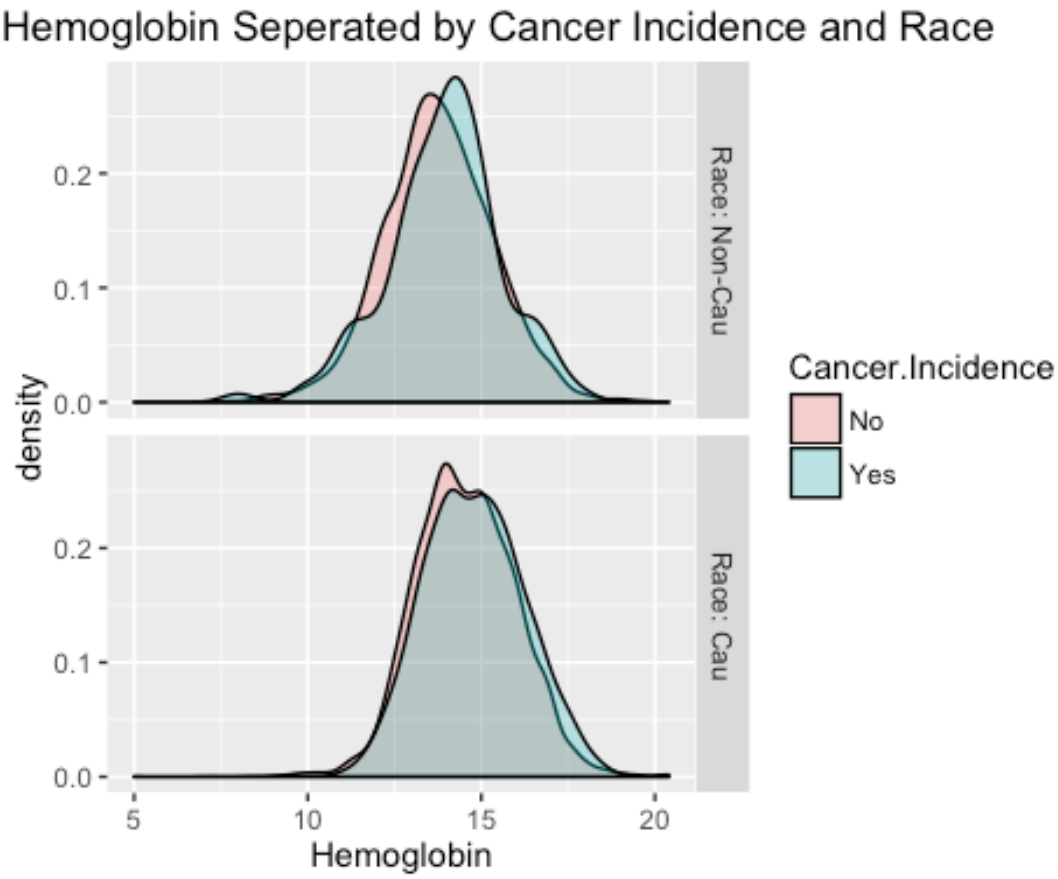
```
require(ggplot2)
plt = ggplot(NHANES, aes(Hemoglobin, fill = Cancer.Incidence))
plt + geom_density(alpha = 0.3)+ facet_grid(Smoke ~ ., labeller = label_both)
 + labs( title = "Density of Hemoglobin Seperated by Cancer Incidence and Smo
king Status" )
```



In "Past Smoker" facet, we cannot see the difference between the density of Hemoglobin for people who have cancer or not. However, for people who are currently smoking and who are non-smoker, the hemoglobin level seems to be a little higher for those who have cancer.

2.5.1.4 Does Hemoglobin ~ Cancer Relation Hold Across Race

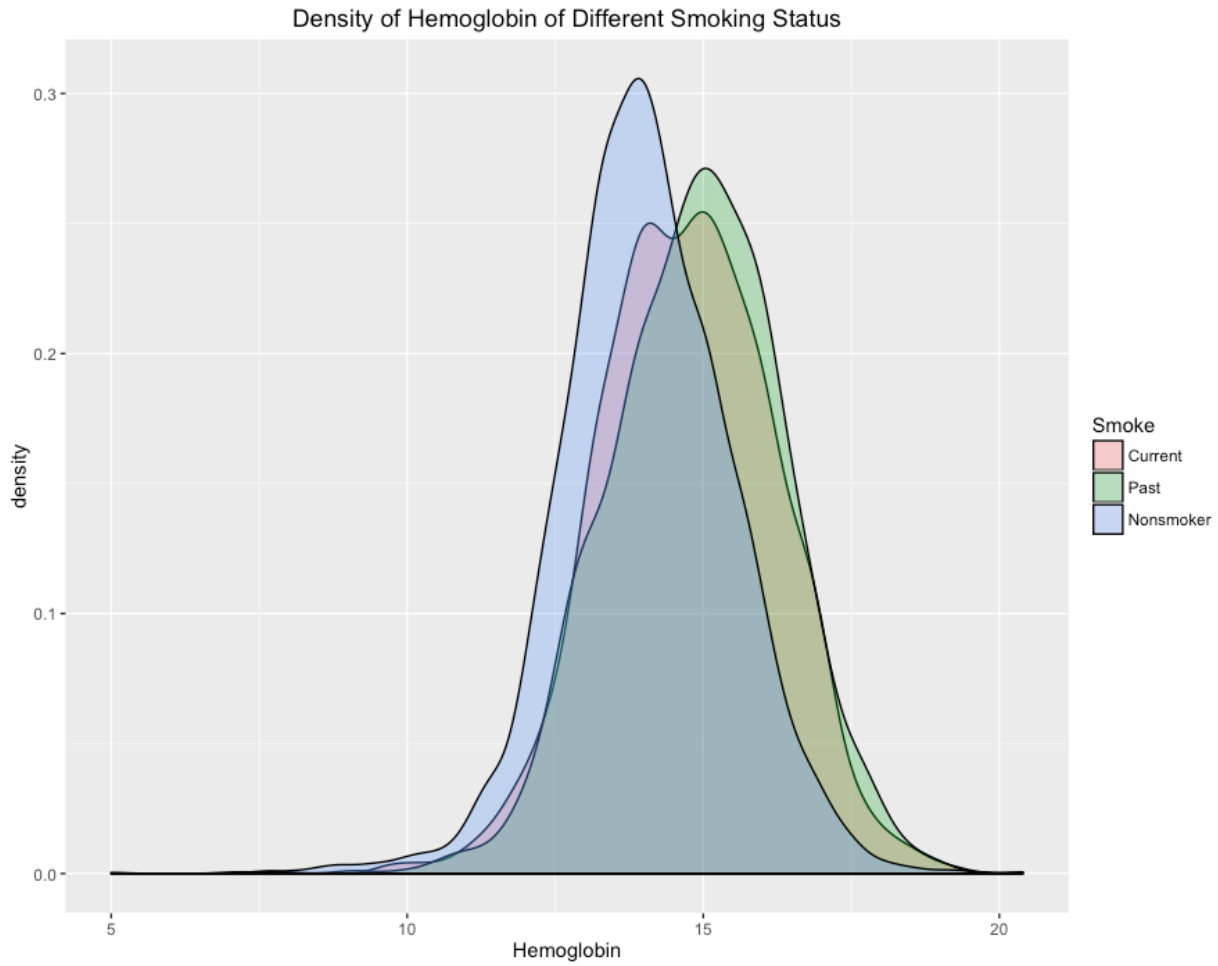
```
require(ggplot2)
plt = ggplot(NHANES, aes(Hemoglobin, fill = Cancer.Incidence))
plt+ geom_density(alpha = 0.3)+ facet_grid(Race ~ ., labeller = label_both) +
 labs( title = "Density of Hemoglobin Seperated by Cancer Incidence and Race"
 )
```



It can be seen clearly from this plot that, for both race, the relation between hemoglobin and cancer still hold.

2.5.2 Hemoglobin vs. Smoke

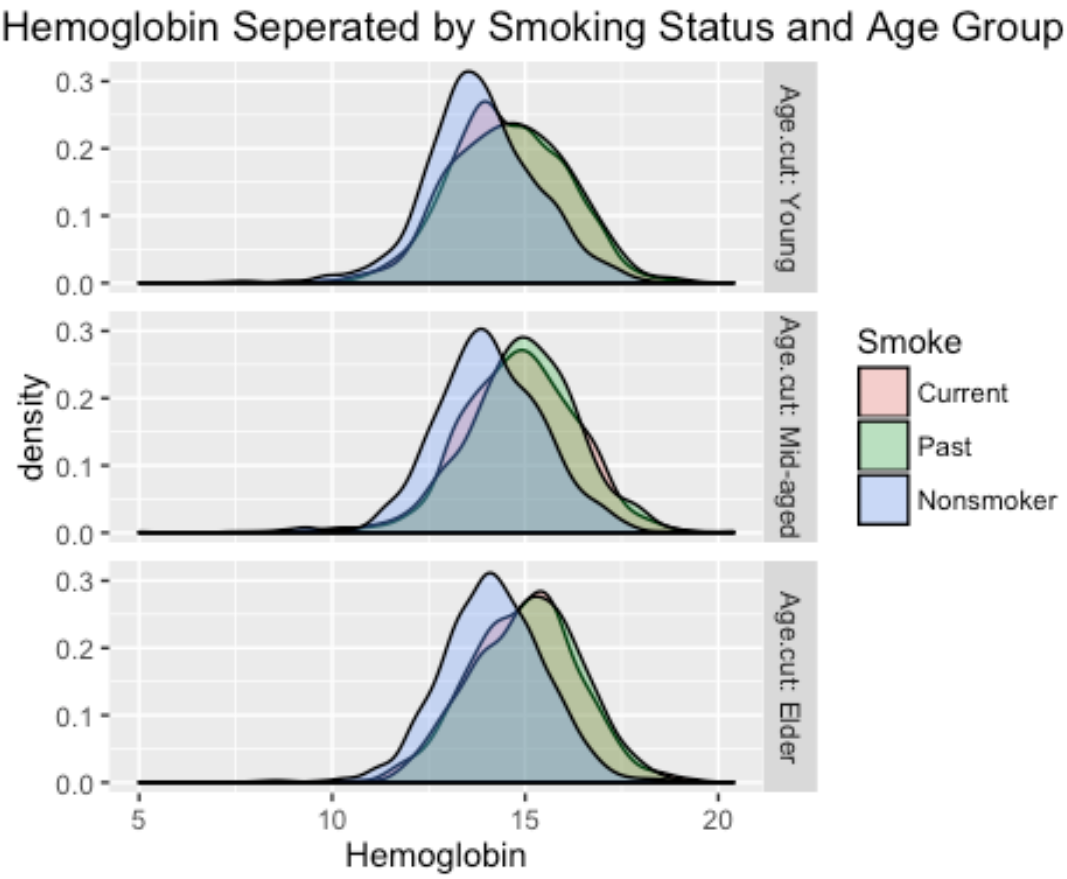
```
require(ggplot2)
plt = ggplot(NHANES[NHANES$Smoke != "Unknown",], aes(Hemoglobin, fill = Smoke
))
plt + geom_density(alpha = 0.3) + labs( title = "Density of Hemoglobin of Dif
ferent Age Group and Gender" )
```



In the previous analysis, it is observed that smoke has a significant influence on hemoglobin level. Smoking leads to higher hemoglobin level.

2.5.2.1 Does Hemoglobin ~ Smoke Relation Hold Across Age

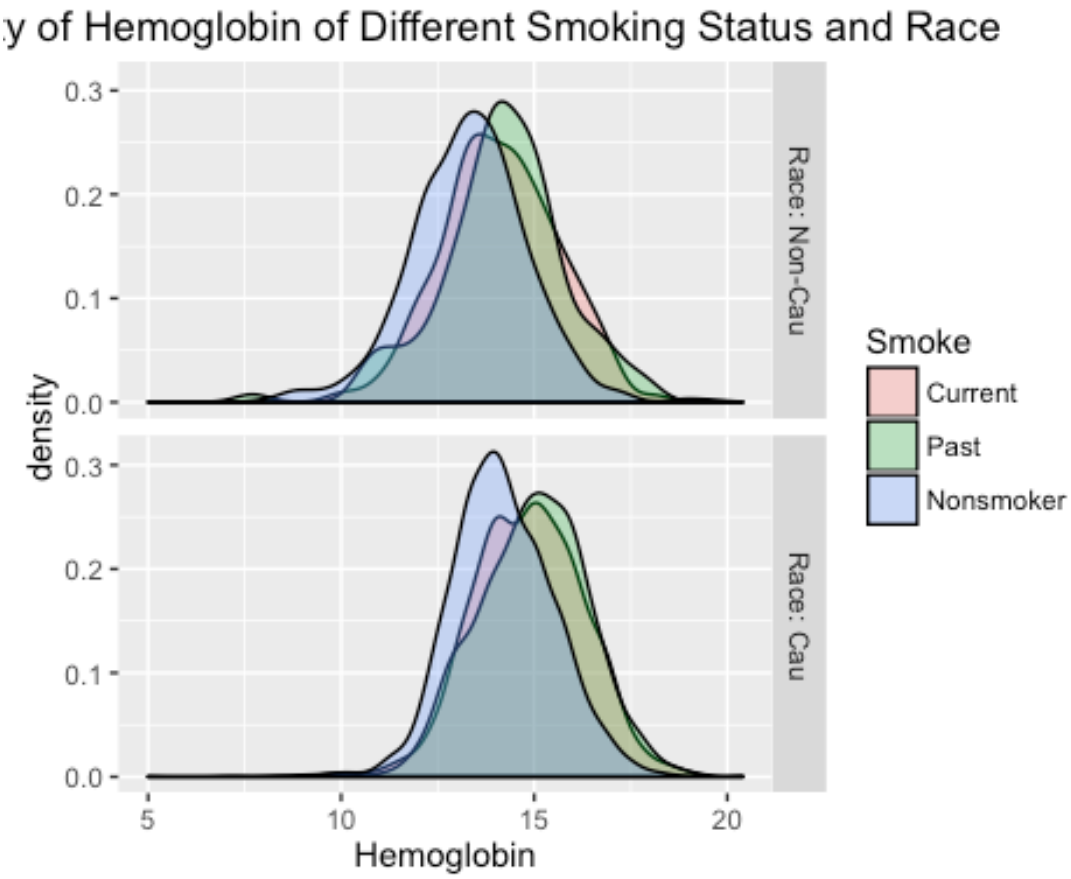
```
require(ggplot2)
plt = ggplot(NHANES[NHANES$Smoke != "Unknown",], aes(Hemoglobin, fill = Smoke
))
plt + geom_density(alpha = 0.3)+ facet_grid(Age.cut ~ . , labeller = label_bo
th) + labs( title = "Density of Hemoglobin Seperated by Smoking Status and Ag
e Group" )
```



Same relation between Hemoglobin and smoke can be seen across age groups.

2.5.2.2 Does Hemoglobin ~ Smoke Relation Hold Across Race

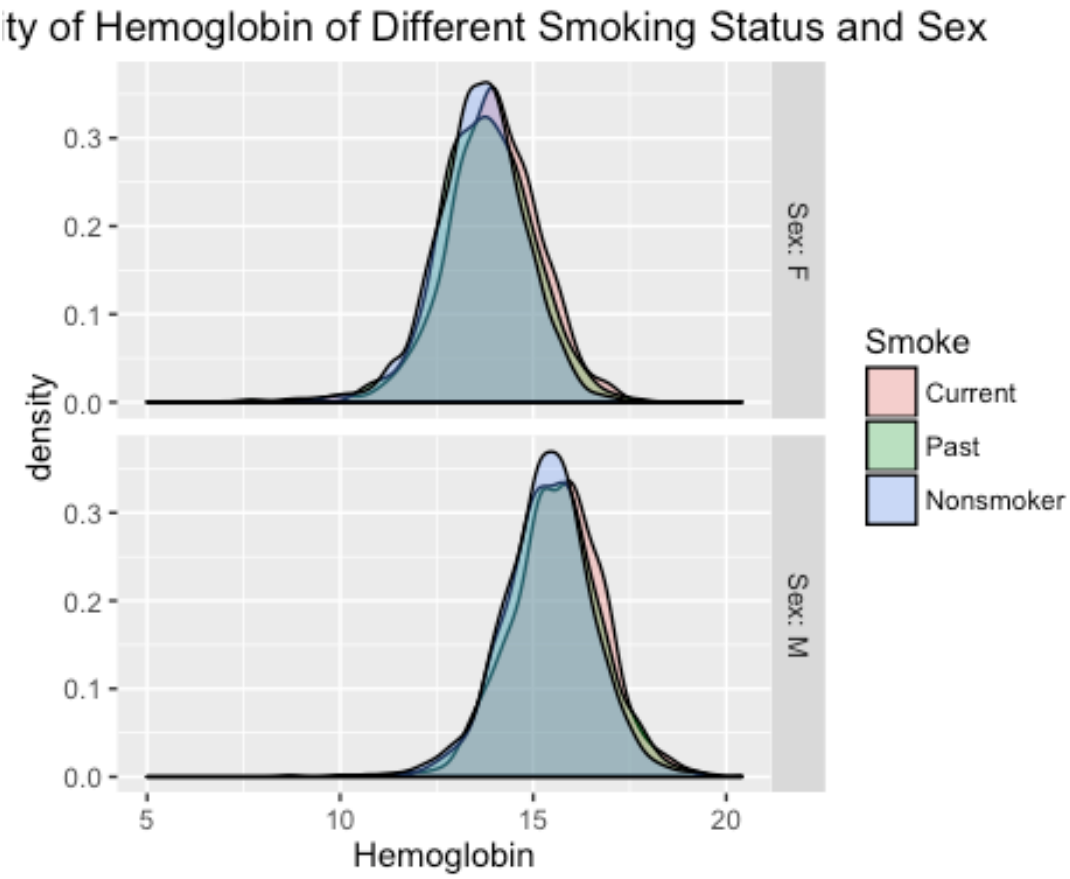
```
require(ggplot2)
plt = ggplot(NHANES[NHANES$Smoke != "Unknown",], aes(Hemoglobin, fill = Smoke
))
plt + geom_density(alpha = 0.3)+ facet_grid(Race ~ . , labeller = label_both)
 + labs( title = "Density of Hemoglobin of Different Smoking Status and Race"
 )
```



Same relation between Hemoglobin and smoke hold for different races.

2.5.2.3 Does Hemoglobin ~ Smoke Relation Hold Across Sex

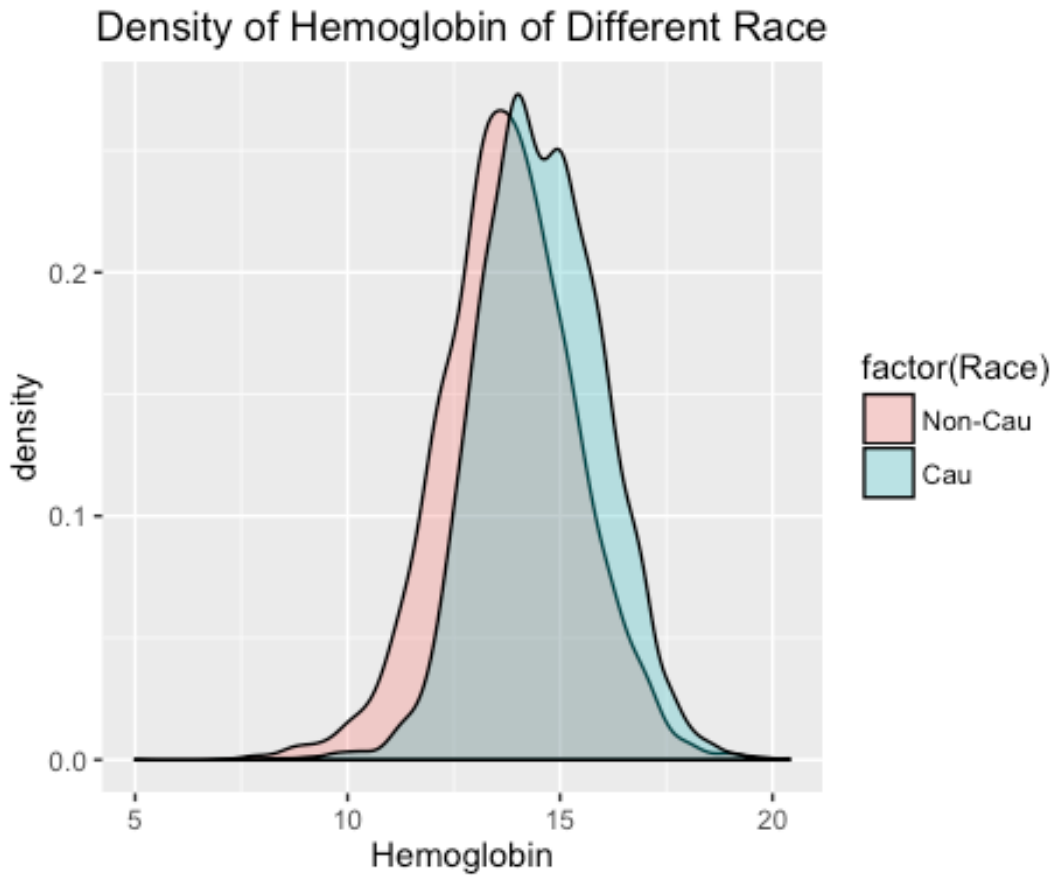
```
require(ggplot2)
plt = ggplot(NHANES[NHANES$Smoke != "Unknown",], aes(Hemoglobin, fill = Smoke
))
plt + geom_density(alpha = 0.3)+ facet_grid(Sex ~ . , labeller = label_both)
+ labs( title = "Density of Hemoglobin of Different Smoking Status and Sex" )
```



It can be found that the relation between hemoglobin and smoking no longer hold when seperated by gender. The density file are almost on the top of each other.

2.5.3 Hemoglobin vs. Race

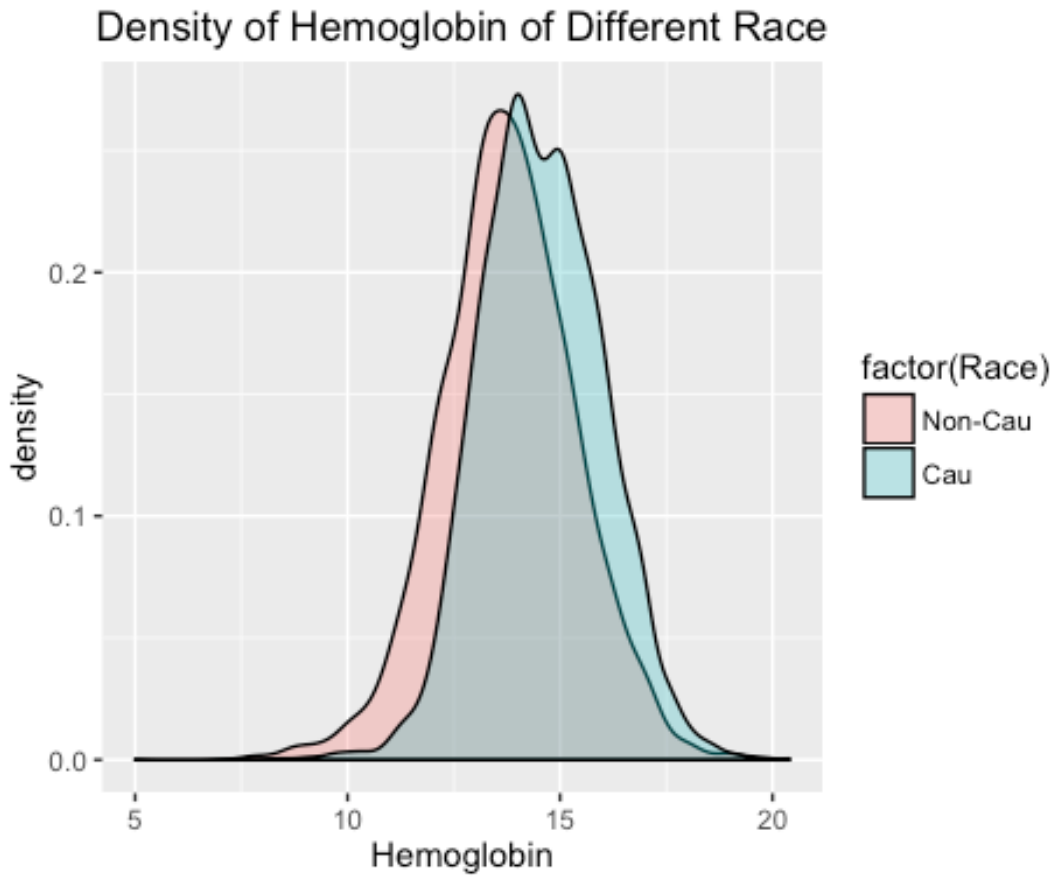
```
require(ggplot2)
plt = ggplot(NHANES, aes(Hemoglobin, fill = factor(Race)))
plt + geom_density(alpha = 0.3) + labs( title = "Density of Hemoglobin of Dif
ferent Race" )
```



In the previous analysis, it is observed that race has a significant influence on hemoglobin level. caucasian in general have higher hemoglobin level.

2.5.3 Hemoglobin vs. Race

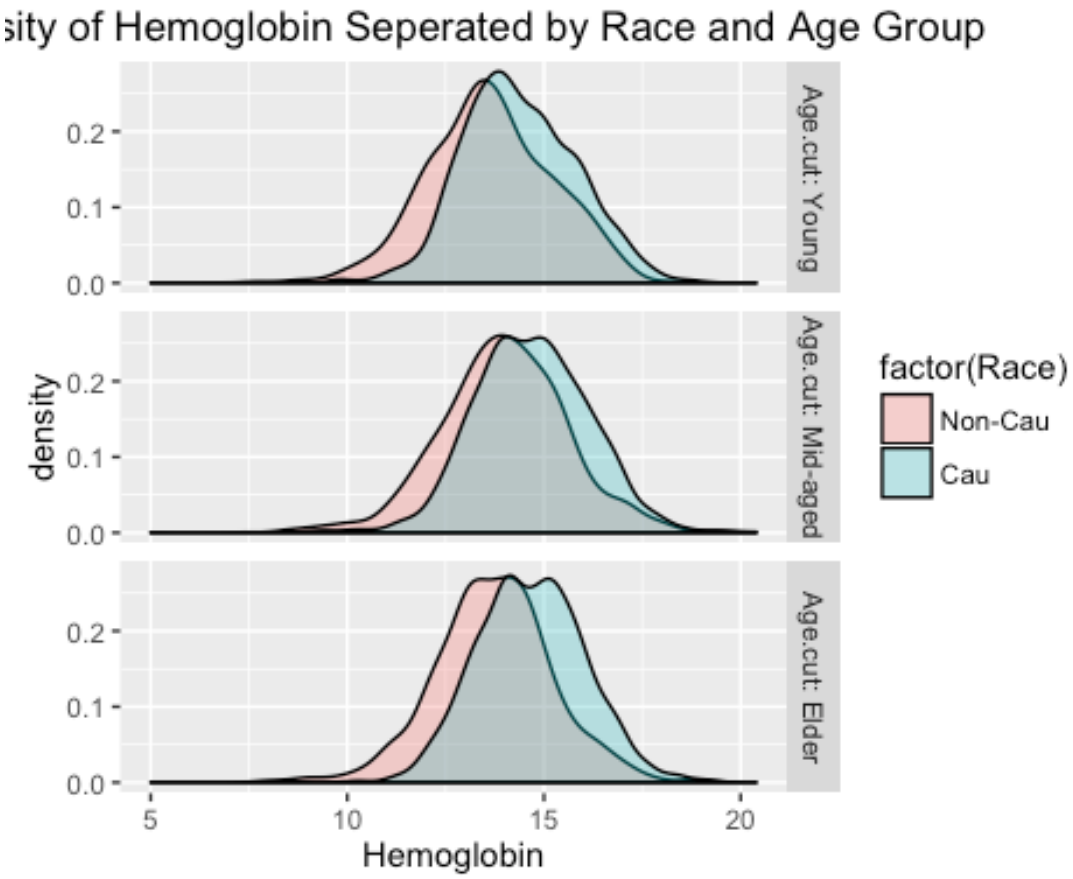
```
require(ggplot2)
plt = ggplot(NHANES, aes(Hemoglobin, fill = factor(Race)))
plt + geom_density(alpha = 0.3) + labs( title = "Density of Hemoglobin of Dif
ferent Race" )
```



In the previous analysis, it is observed that race has a significant influence on hemoglobin level. caucasian have higher hemoglobin level.

2.5.3.1 Does Hemoglobin ~ Race Relation Hold Across Age

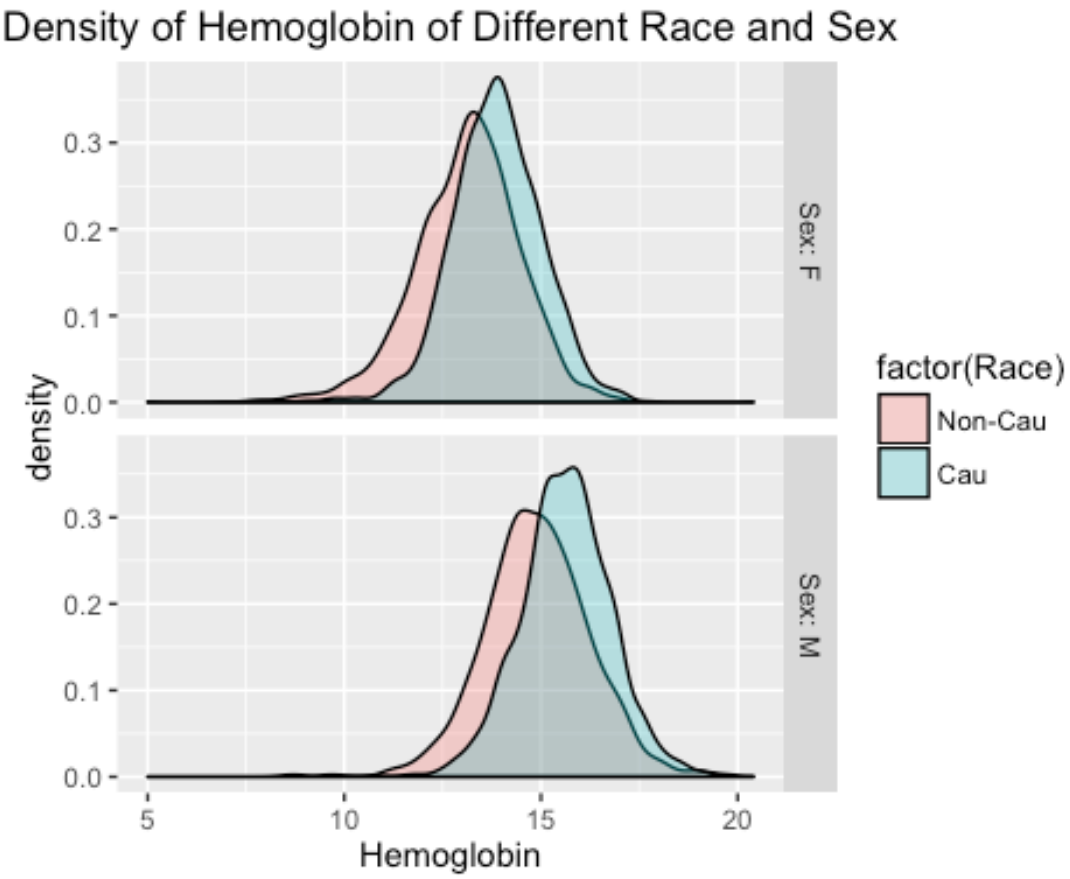
```
require(ggplot2)
plt = ggplot(NHANES, aes(Hemoglobin, fill = factor(Race)))
plt + geom_density(alpha = 0.3)+ facet_grid(Age.cut ~ . , labeller = label_bo
th) + labs( title = "Density of Hemoglobin Seperated by Race and Age Group" )
```



Same relation between Hemoglobin and race can be seen across different age groups.

2.5.3.2 Does Hemoglobin ~ Race Relation Hold Across Sex

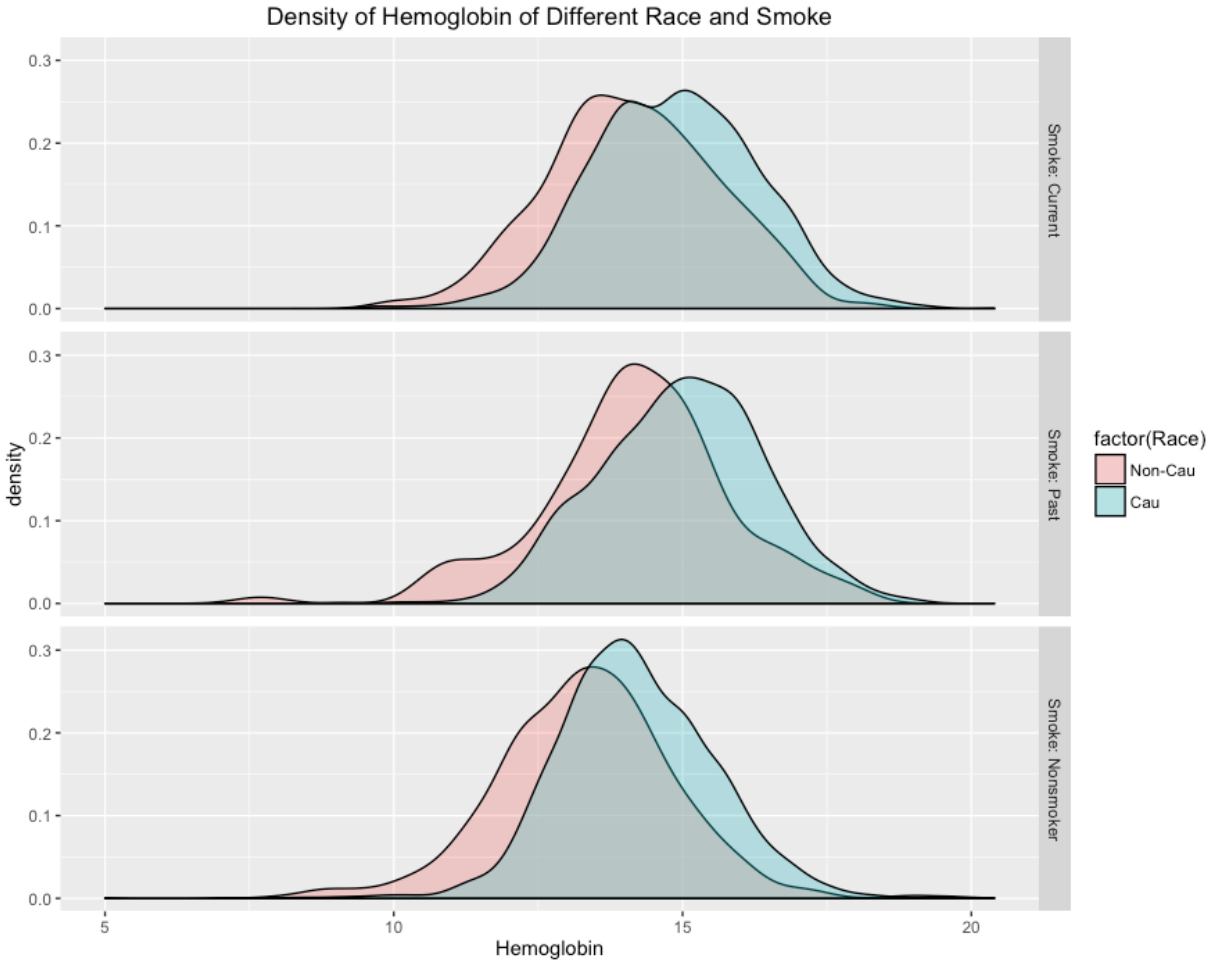
```
require(ggplot2)
plt = ggplot(NHANES, aes(Hemoglobin, fill = factor(Race)))
plt + geom_density(alpha = 0.3)+ facet_grid(Sex ~ . , labeller = label_both)
+ labs( title = "Density of Hemoglobin of Different Race and Sex" )
```



Same relation between Hemoglobin and race hold for different gender.

2.5.3.3 Does Hemoglobin ~ Race Relation Hold Across Smoke

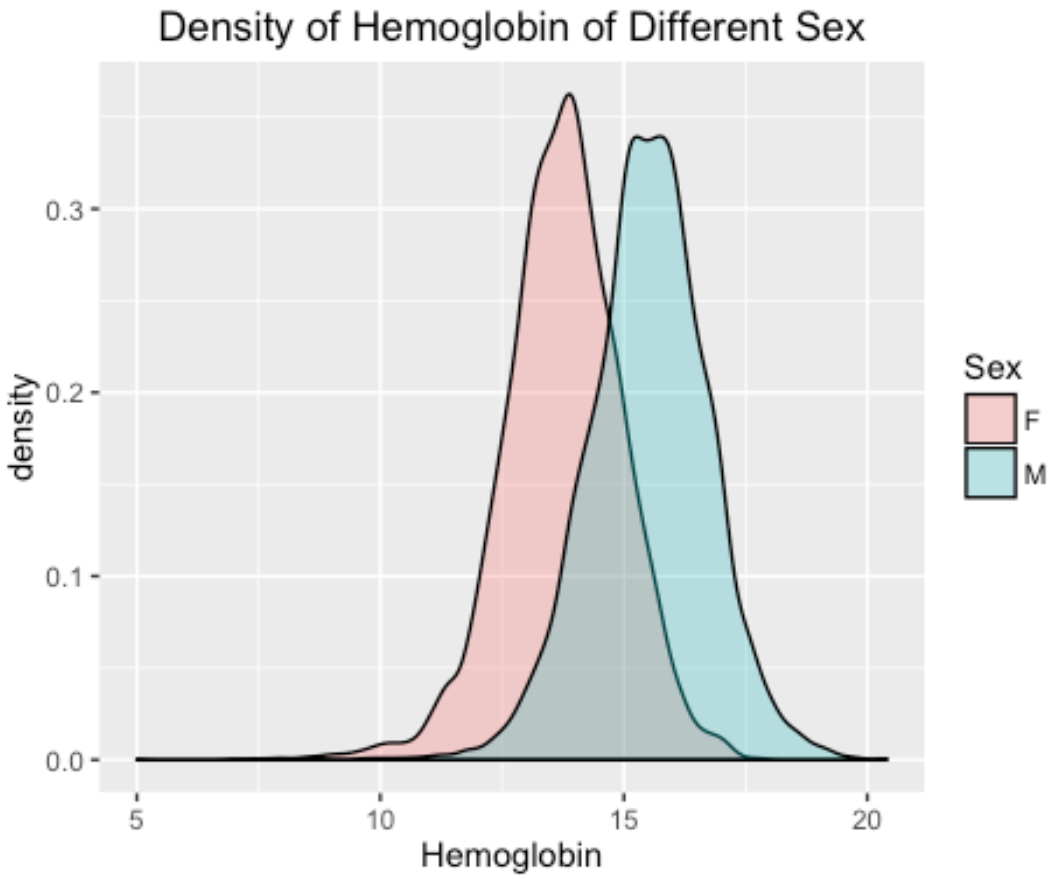
```
require(ggplot2)
plt = ggplot(NHANES[NHANES$Smoke != "Unknown",], aes(Hemoglobin, fill = factor(Race)))
plt + geom_density(alpha = 0.3)+ facet_grid(Smoke ~ . , labeller = label_both) + labs( title = "Density of Hemoglobin of Different Race and Smoke" )
```



Same relation between Hemoglobin and race hold for different smoking status.

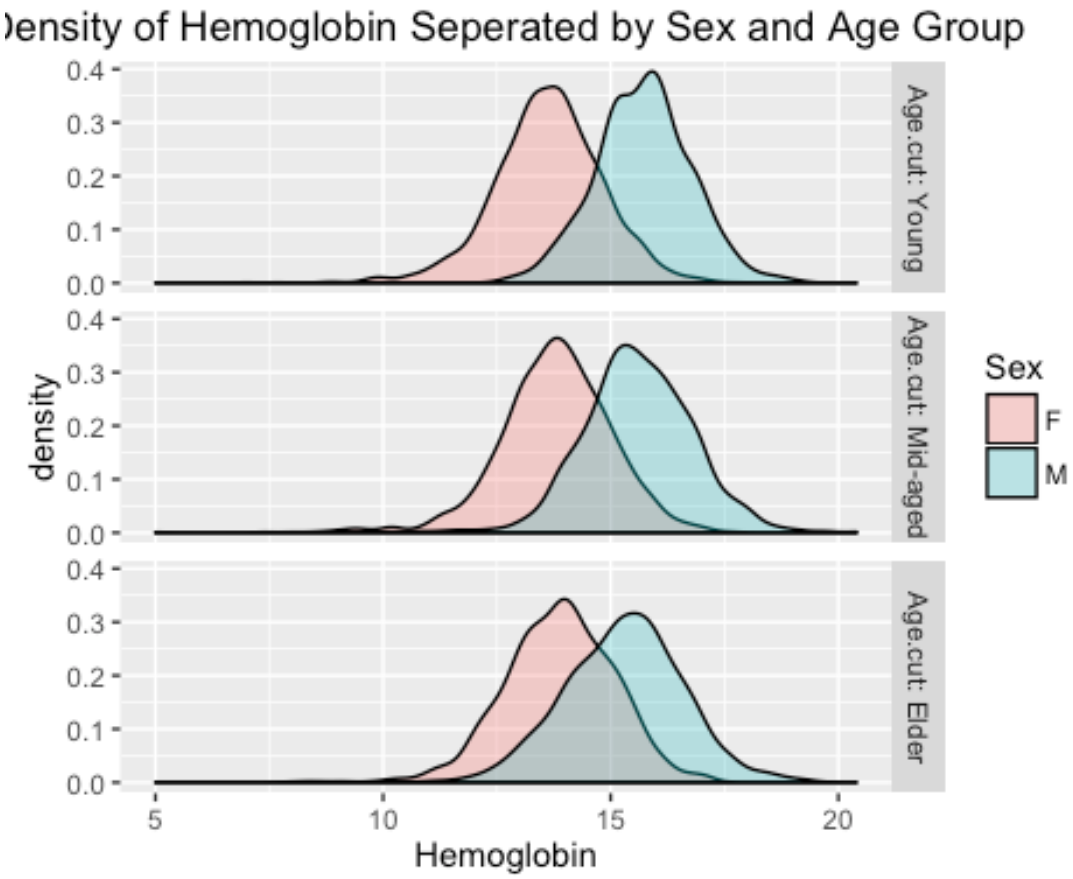
2.5.4 Hemoglobin vs. Sex

```
require(ggplot2)
plt = ggplot(NHANES, aes(Hemoglobin, fill = Sex))
plt + geom_density(alpha = 0.3) + labs( title = "Density of Hemoglobin of Dif
ferent Sex" )
```



In the previous analysis, it is observed that smoke has a significant influence on hemoglobin level. Smoking leads to higher hemoglobin level. #### 2.5.4.1 Does Hemoglobin ~ Sex Relation Hold Across Age

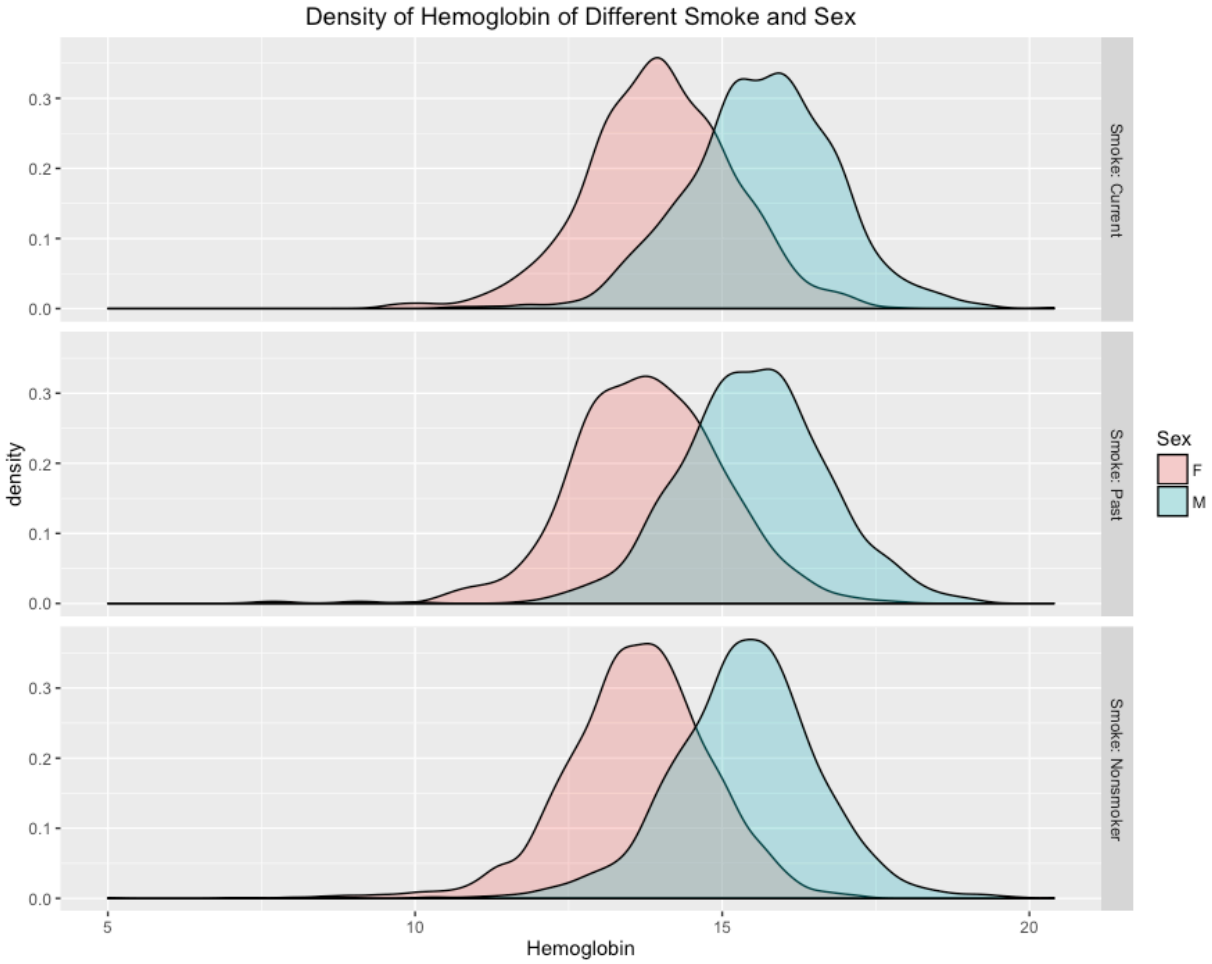
```
require(ggplot2)
plt = ggplot(NHANES, aes(Hemoglobin, fill = Sex))
plt + geom_density(alpha = 0.3)+ facet_grid(Age.cut ~ . , labeller = label_bo
th) + labs( title = "Density of Hemoglobin Seperated by Sex and Age Group" )
```



Same relation between Hemoglobin and sex can be seen across different age groups.

2.5.4.2 Does Hemoglobin ~ Sex Relation Hold Across Smoke

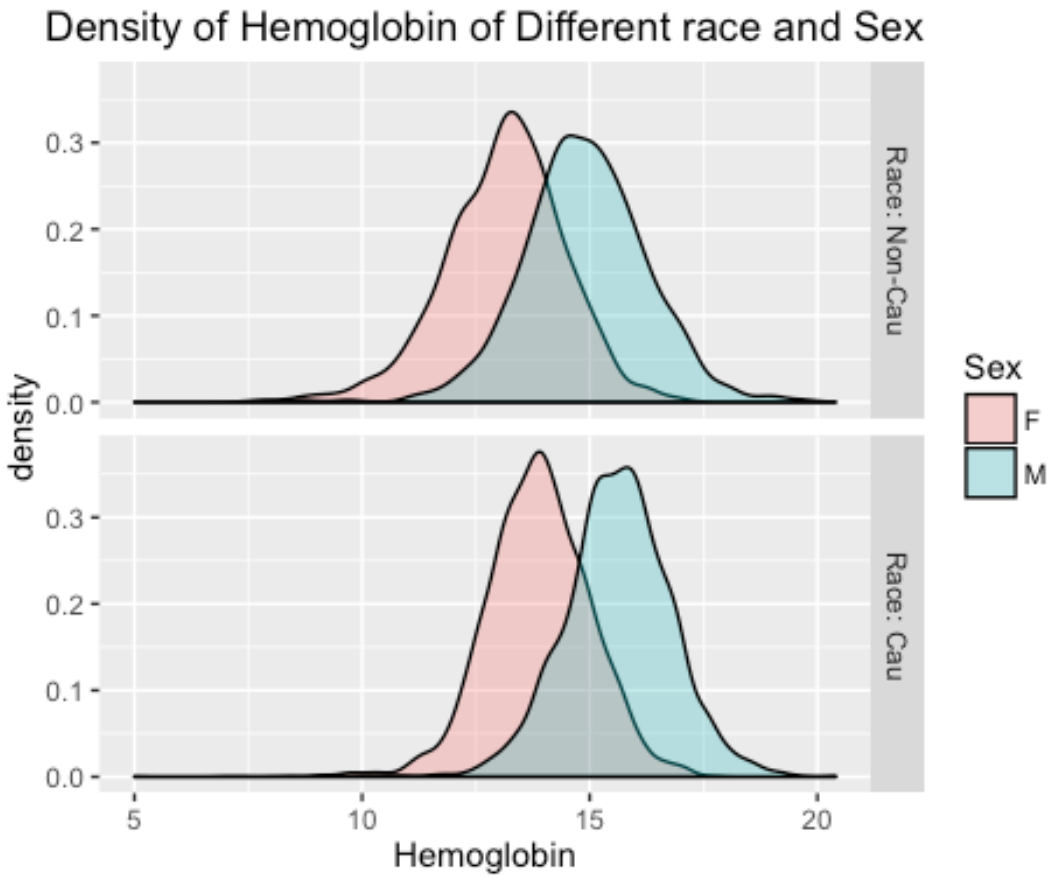
```
require(ggplot2)
plt = ggplot(NHANES[NHANES$Smoke != "Unknown",], aes(Hemoglobin, fill = Sex)
)
plt + geom_density(alpha = 0.3)+ facet_grid(Smoke ~ . , labeller = label_both
) + labs( title = "Density of Hemoglobin of Different Smoke and Sex" )
```



Same relation between Hemoglobin and sex hold for different smoking groups.

2.5.4.3 Does Hemoglobin ~ Sex Relation Hold Across Race

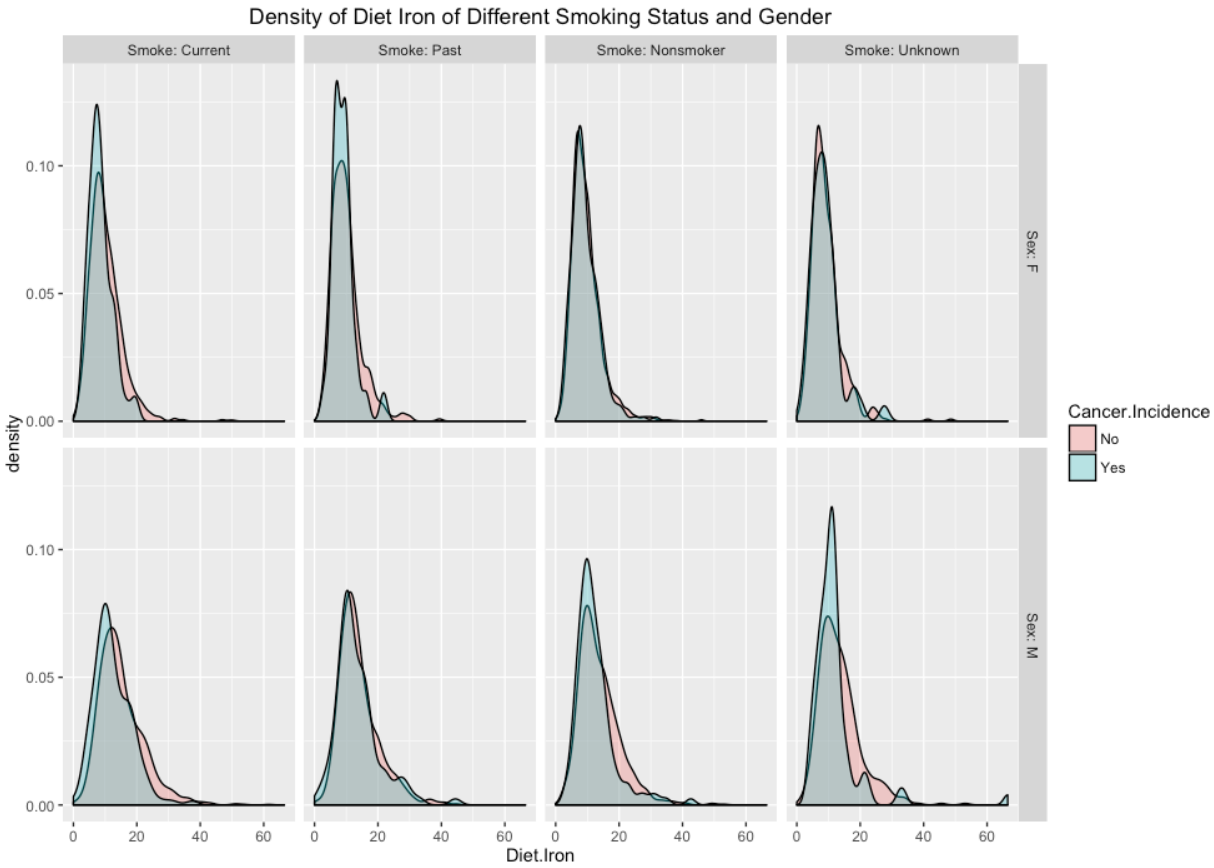
```
require(ggplot2)
plt = ggplot(NHANES, aes(Hemoglobin, fill = Sex))
plt + geom_density(alpha = 0.3)+ facet_grid(Race ~ . , labeller = label_both)
 + labs( title = "Density of Hemoglobin of Different race and Sex" )
```



Same relation between Hemoglobin and sex hold for different race.

2.6 More Analysis on Diet Iron

```
require(ggplot2)
plt = ggplot(NHANES, aes(Diet.Iron, fill = Cancer.Incidence))
plt + geom_density(alpha = 0.3)+ facet_grid(Sex ~ Smoke , labeller = label_bo
th) + labs( title = "Density of Diet Iron of Different Smoking Status and Gen
der" )
```

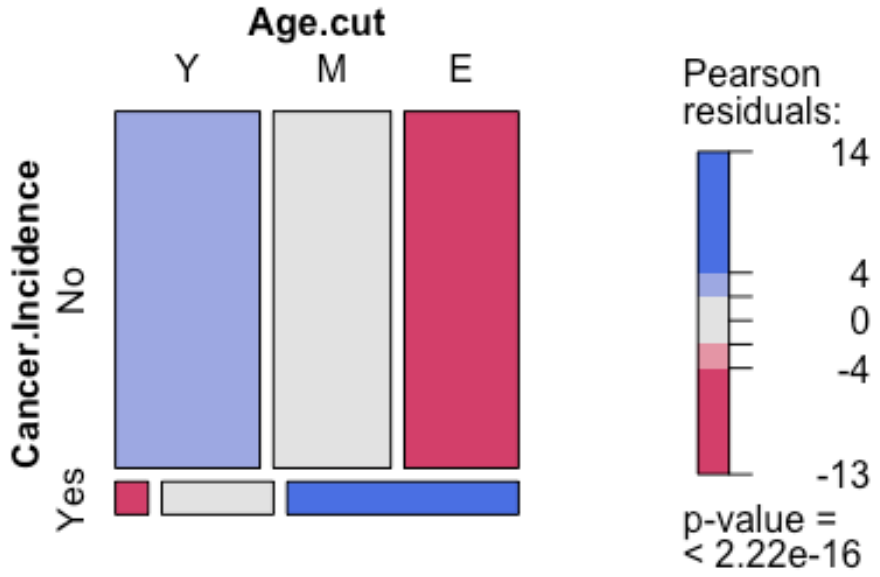


No siginificant difference observed.

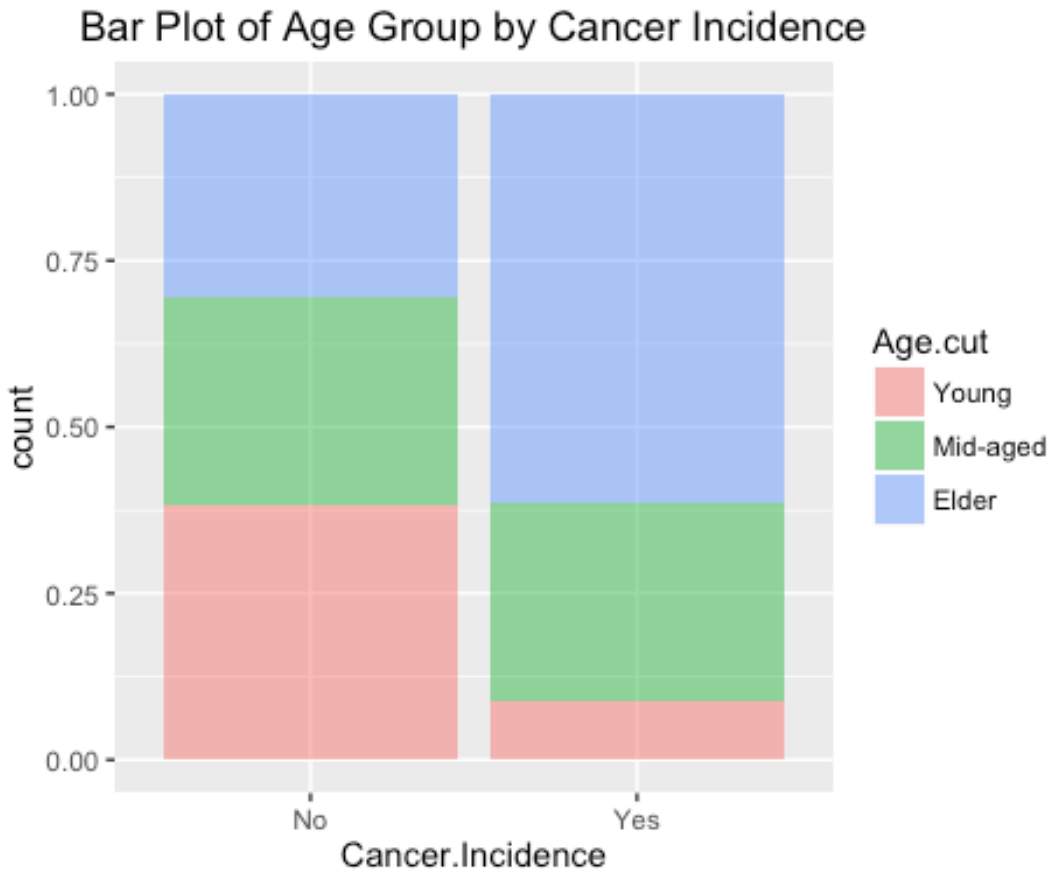
3 Analysis on Categorical Variables

3.1 Relation between Cancer + Age Group

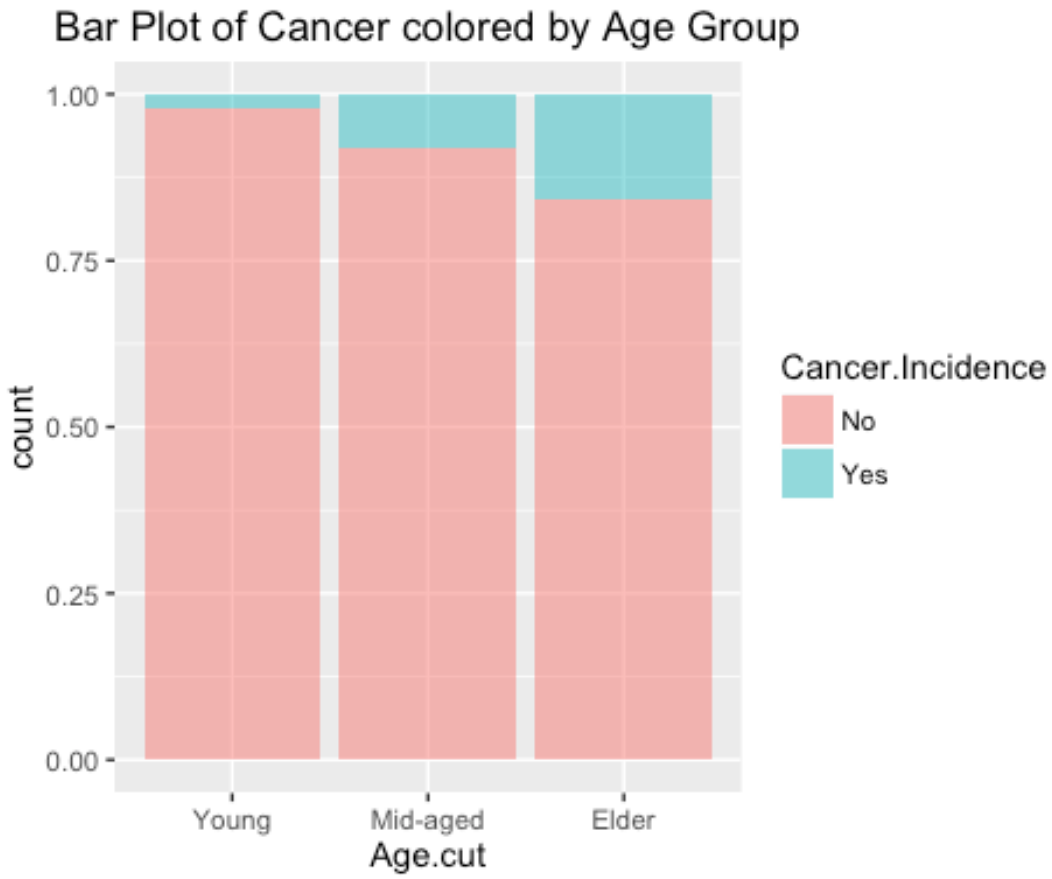
```
mosaic(~ Cancer.Incidence + Age.cut , data = NHANES, shade = TRUE, legend = T
RUE,   labeling_args=list(abbreviate=c(Age.cut=1)), main = "Mosaic Plot of Ca
ncer vs. Age ")
```




```
plt = ggplot( NHANES, aes( x = Cancer.Incidence, fill = Age.cut ))
plt + geom_bar(position="fill", alpha = 0.5) +  labs(title = "Bar Plot of Age
 Group by Cancer Incidence")
```



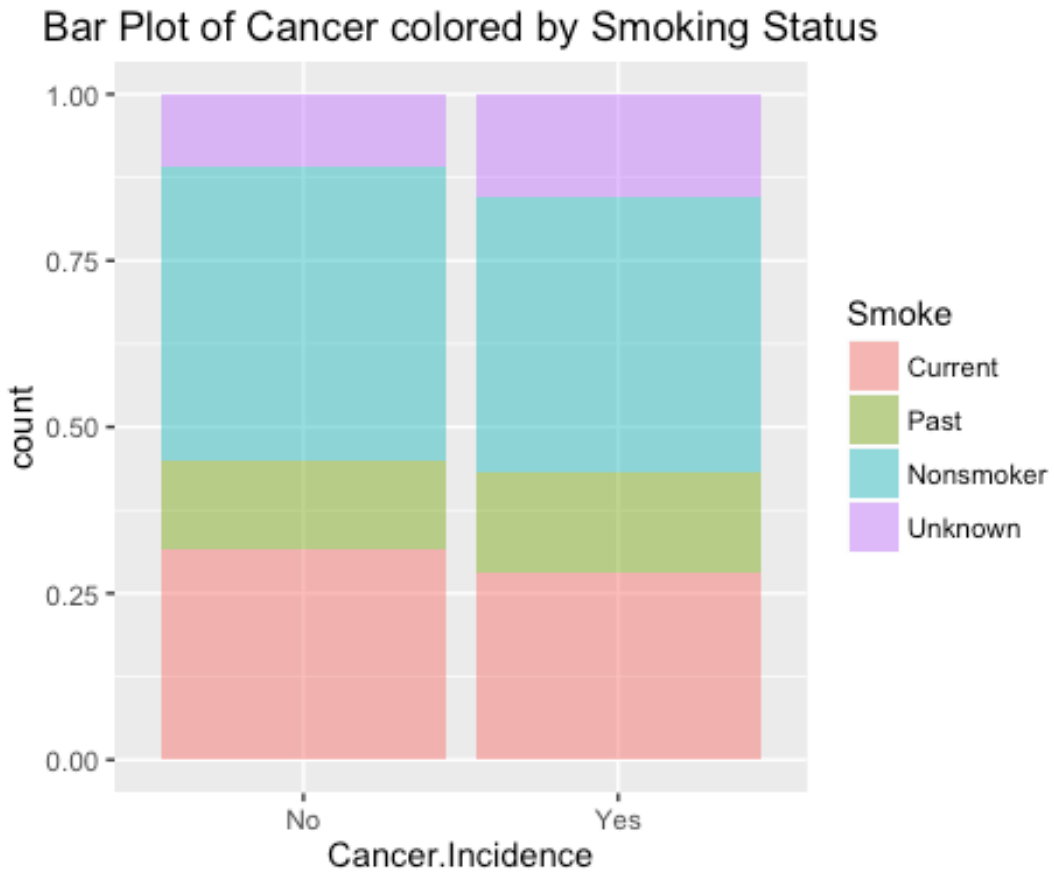

```
plt = ggplot( NHANES, aes( x = Age.cut, fill = Cancer.Incidence ))
plt + geom_bar(position="fill", alpha = 0.5) +  labs(title = "Bar Plot of Can
cer colored by Age Group")
```



It can be observed clearly from these plots that age has a significant influence on cancer. Over 60% of people who get cancer falls into the elder category. People who are over 40 years old takes up almost 90% of the people who get cancer.

3.2 Relation between Cancer + Smoke

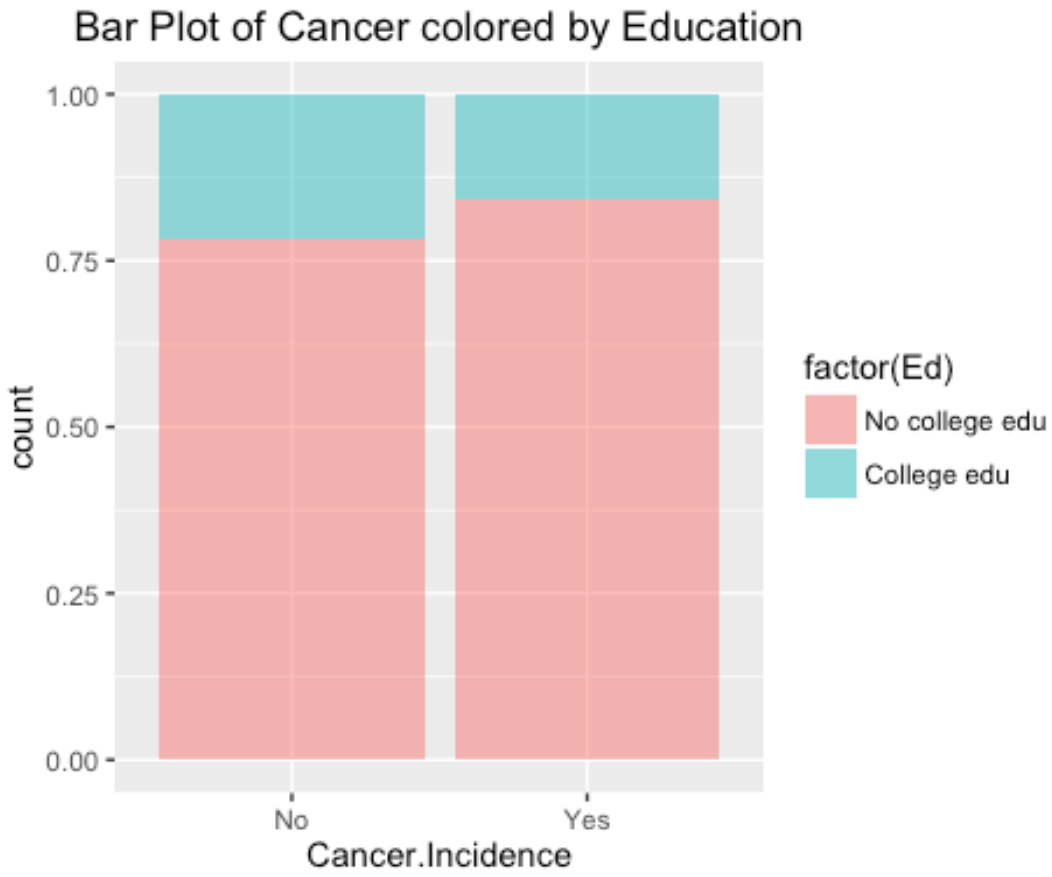
```
plt = ggplot( NHANES, aes( x = Cancer.Incidence, fill = Smoke ))
plt + geom_bar(position="fill", alpha = 0.5) +  labs(title = "Bar Plot of Can
cer colored by Smoking Status")
```



This plot shows that the percentage of each smoking status is similar for different cancer group. So the cancer incidence is independent of smoking status.

3.3 Relation between Cancer + Ed

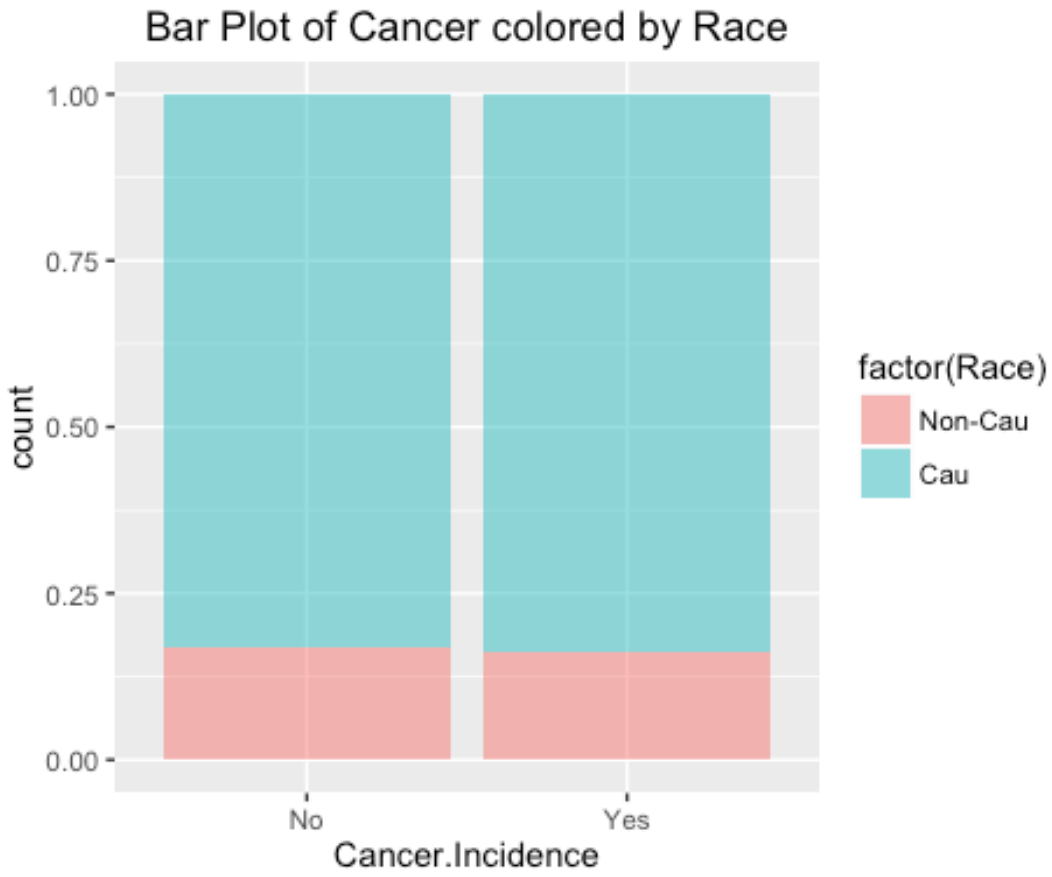
```
plt = ggplot( NHANES, aes( x = Cancer.Incidence, fill = factor(Ed)))
plt + geom_bar(position="fill", alpha = 0.5) + labs(title = "Bar Plot of Canc
er colored by Education")
```



It can be observed from this bar plot that, education has a trivial influence on cancer. In the category where people don't have cancer, the percentage of people with college degree is about 20%. While for people who have cancer, the educated percentage is about 5% lower.

3.4 Relation between Cancer + Race

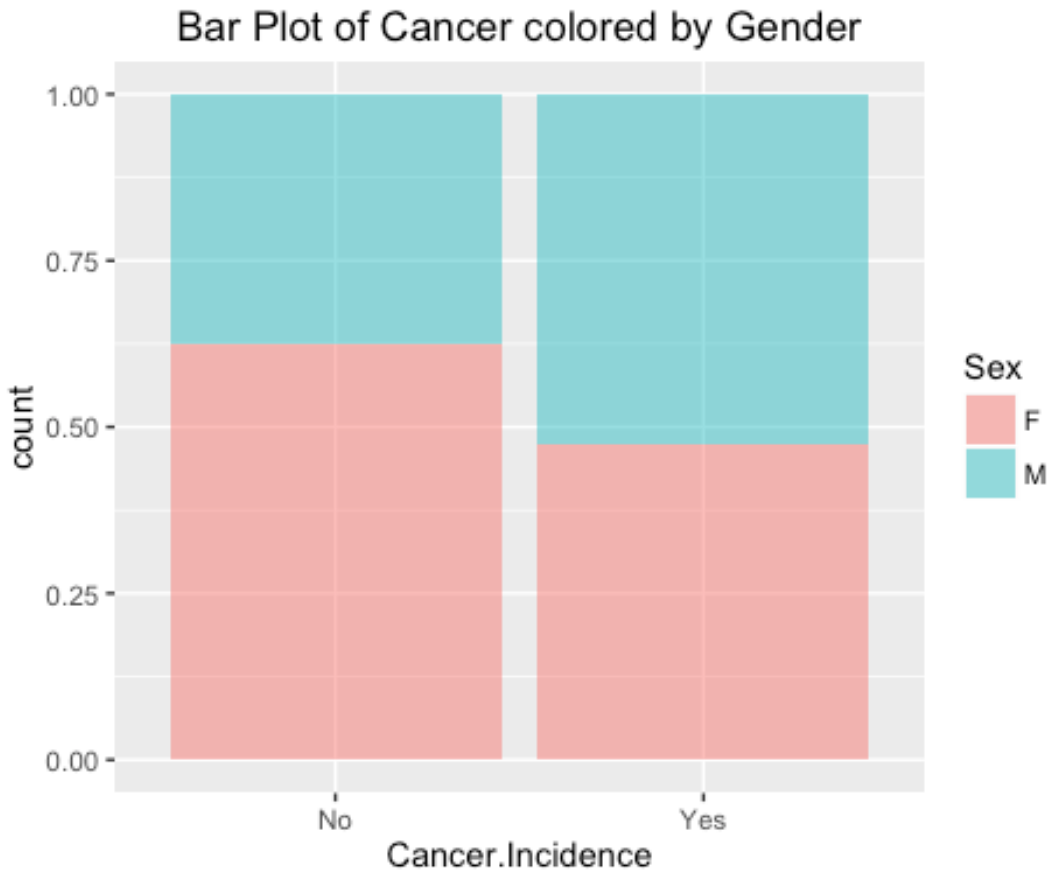
```
plt = ggplot( NHANES, aes( x = Cancer.Incidence, fill = factor(Race)))
plt + geom_bar(position="fill", alpha = 0.5) +  labs(title = "Bar Plot of Can
cer colored by Race")
```



No significant relation can be seen from race and cancer.

3.5 Relation between Cancer + Sex

```
plt = ggplot( NHANES, aes( x = Cancer.Incidence, fill = Sex))
plt + geom_bar(position="fill", alpha = 0.5) + labs(title = "Bar Plot of Canc
er colored by Gender")
```

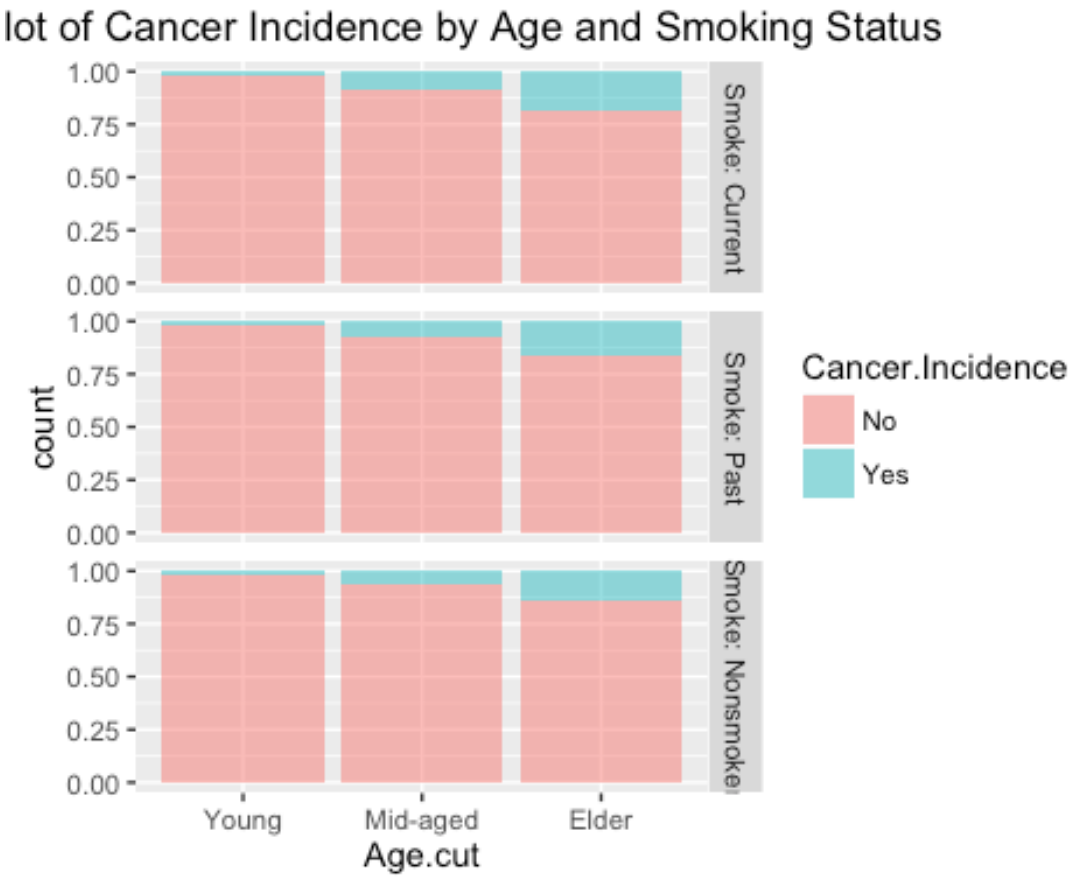


This plot shows that gender is an important factor in cancer. Higher percentage of male in the population of "Cancer Incidence" can be seen. Thus it seems that female are less likely to get cancer.

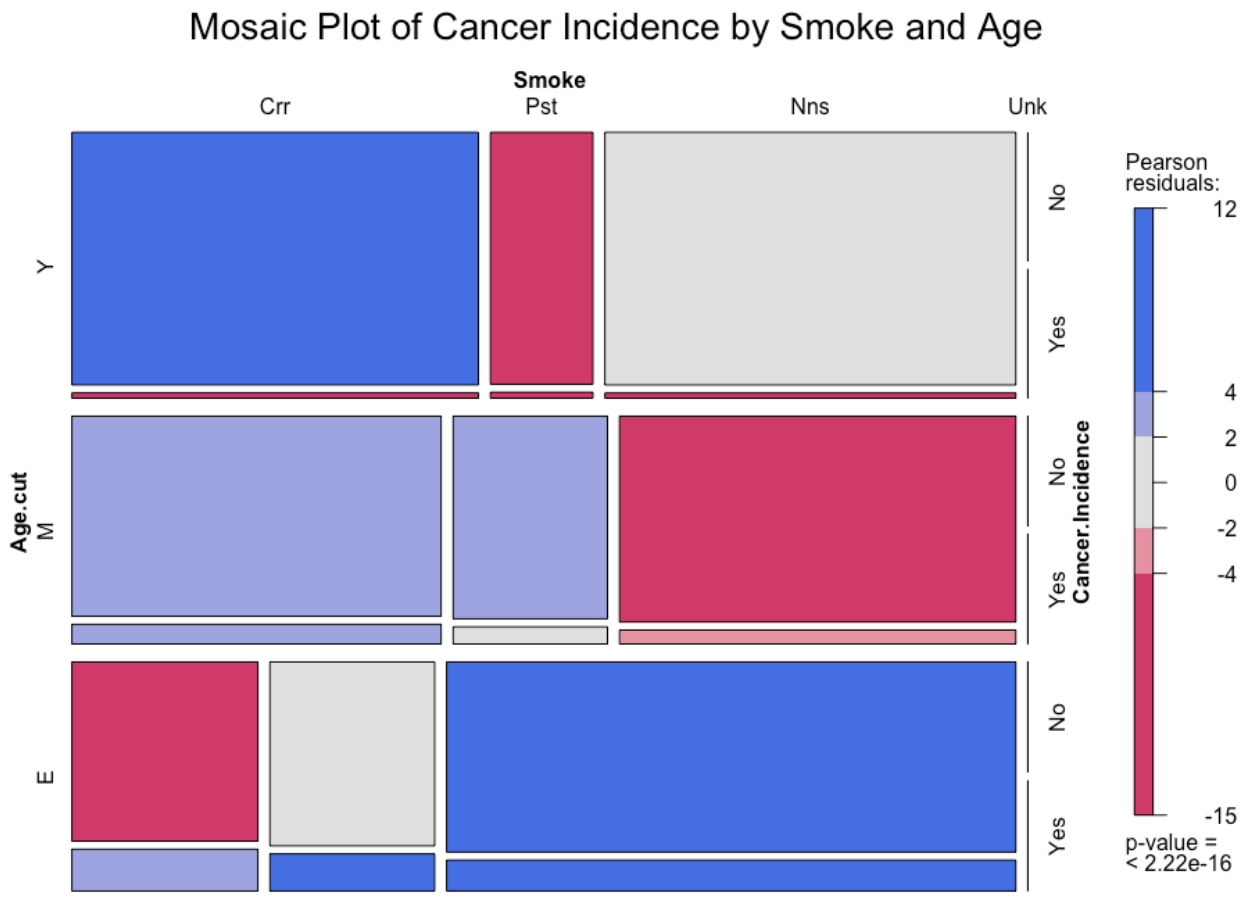
3.6 Relation between Cancer + Age across other factor

3.6.1 Cancer ~ Age Relation across Smoke

```
plt = ggplot( NHANES[NHANES$Smoke != "Unknown",], aes( x=Age.cut, fill=Cancer
.Incidence))
plt + geom_bar(position="fill", alpha = 0.5) + facet_grid(Smoke ~ . , labelle
r = label_both, scales = "free") +  labs(title = "Bar Plot of Cancer Incidenc
e by Age and Smoking Status")
```




```r
mosaic(~ Age.cut +  Smoke + Cancer.Incidence, data = NHANES[NHANES$Smoke != "
Unknown",], shade = TRUE, legend = TRUE, labeling_args=list(abbreviate=c(Smok
e = 3, Age.cut=1)), main = "Mosaic Plot of Cancer Incidence by Smoke and Age"
)
```



It can be seen from the first bar plot that, the dependence of cancer incidence on age does hold across different smoking status.

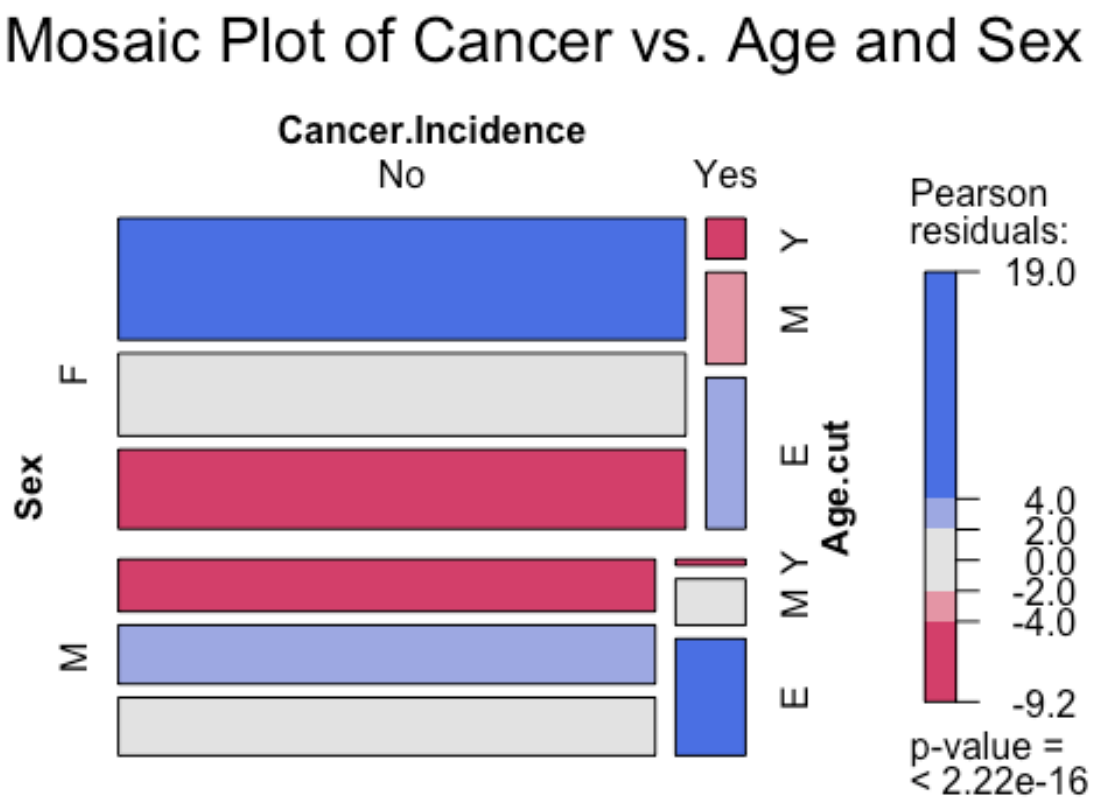
1) For elder people, the percentage of people who are diagnosed with cancer is the highest in "Current Smoker" category. Lower in "Past Smoker" group, and lowest in "Non-smoker" group. Here "Unknown" group is not considered. Thus, for elder people, smoking are more likely to cause the incidence of cancer.

2) Similar influence of smoking status on "Mid-aged" people can be seen. While for young people, there seems to be no influence.

So for older people, smoking might cause a higher chance to get cancer.

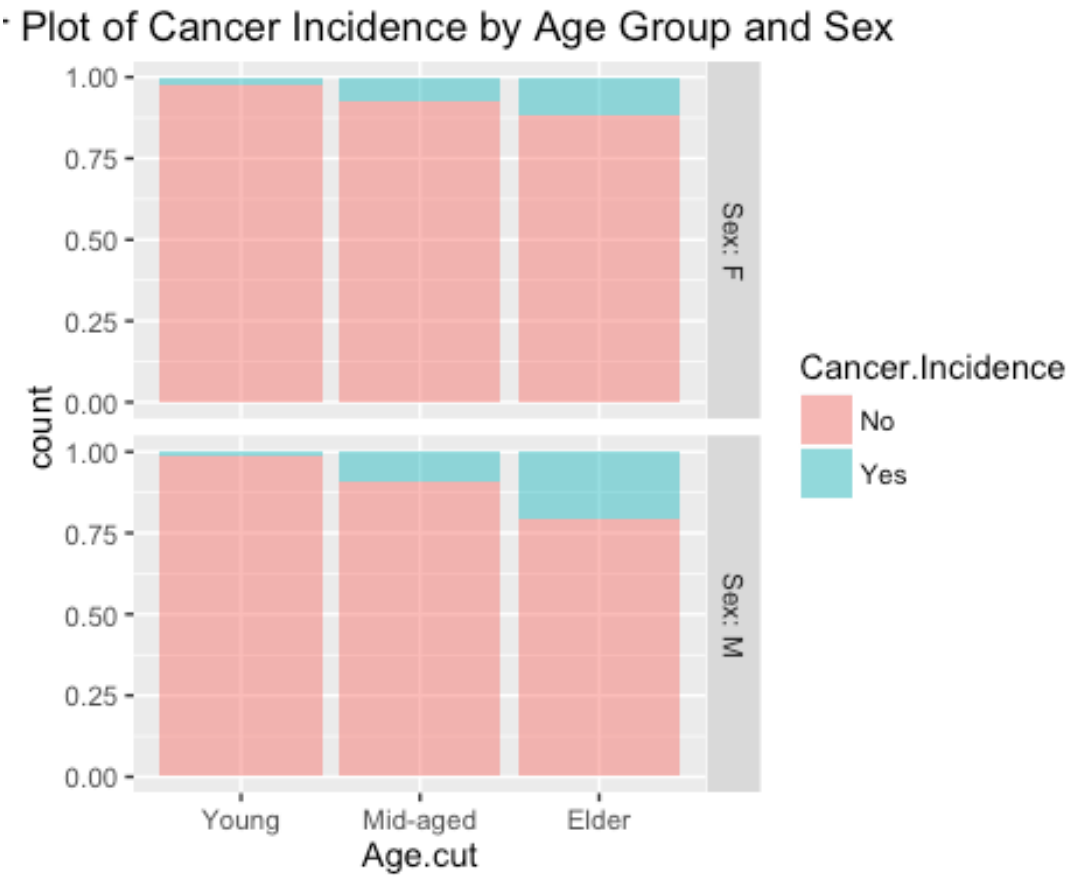
3.6.2 Cancer ~ Age Relation across Sex

```
mosaic(~ Sex + Cancer.Incidence + Age.cut , data = NHANES, shade = TRUE, lege
nd = TRUE,   labeling_args=list(abbreviate=c(Age.cut=1)), main = "Mosaic Plot
 of Cancer vs. Age and Sex")
```



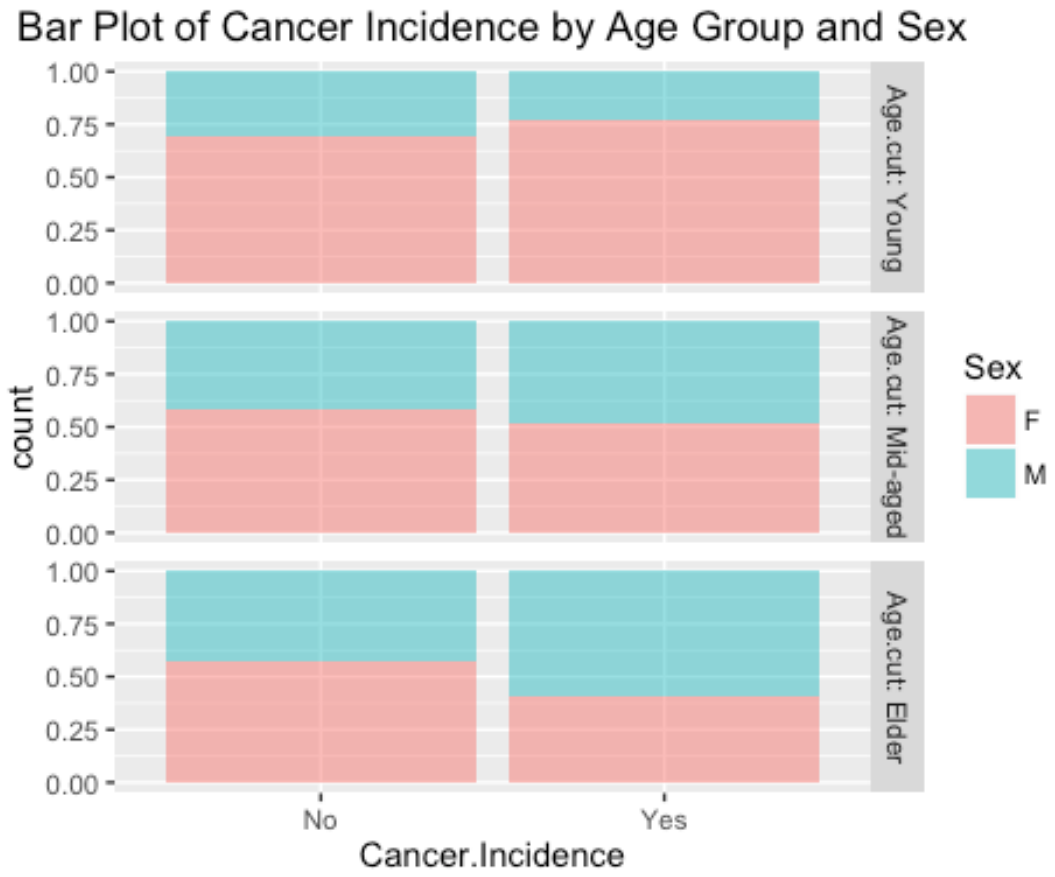
From this mosaic plot of Cancer vs. Age and Gender, it can be concluded that, 1) The dependence of cancer on age doesn't change across gender. 2) For both female and male, people diagnosed with cancer are in general have higher age than people who don't have cancer.

```
ggplot( NHANES, aes( x=Age.cut, fill=Cancer.Incidence )) + geom_bar(position=
"fill", alpha = 0.5) + facet_grid(Sex ~ . , labeller = label_both, scales = "
free") +  labs(title = "Bar Plot of Cancer Incidence by Age Group and Sex")
```



This bar plot shows that, the relation between cancer and age is the similar for two genders. However, the percentage of people diagnosed with cancer increases more for male, indicating that man have higer change to get cancer when getting old.

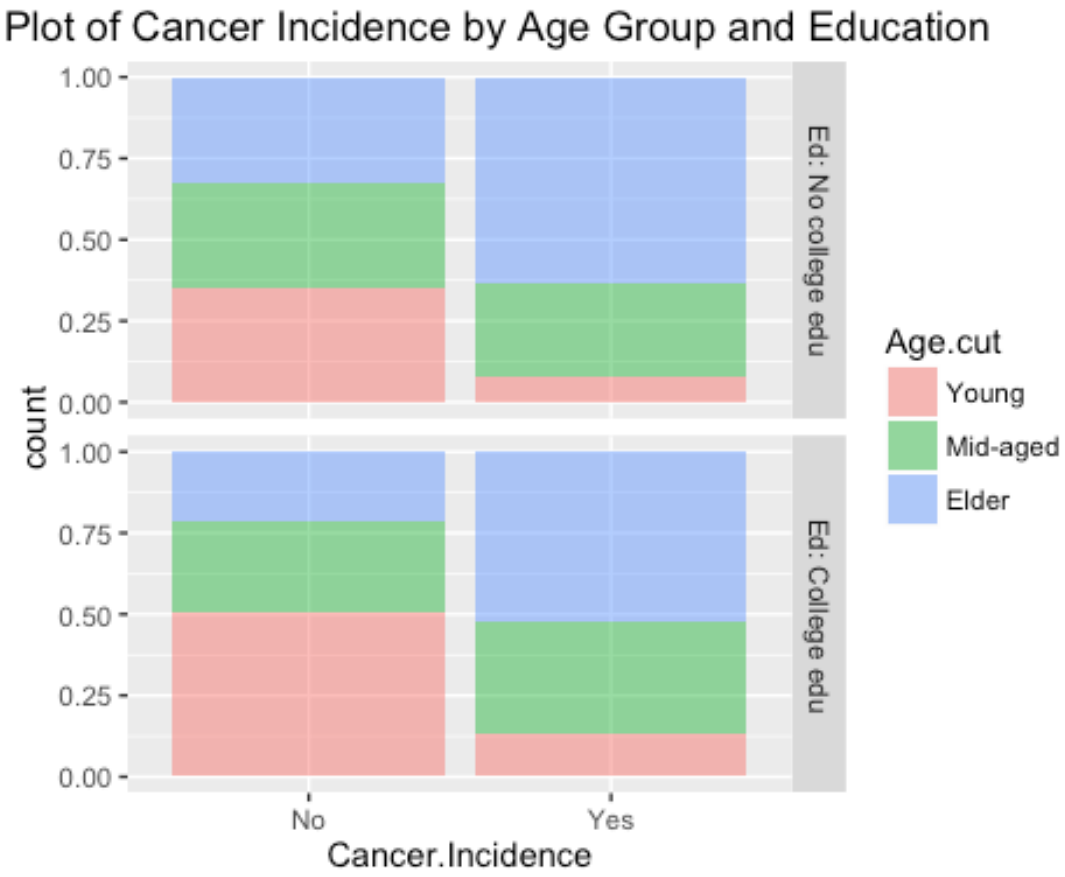
```
ggplot( NHANES, aes( x = Cancer.Incidence, fill = Sex )) + geom_bar(position=
"fill", alpha = 0.5) + facet_grid(Age.cut ~ . , labeller = label_both, scales
 = "free") +  labs(title = "Bar Plot of Cancer Incidence by Age Group and Sex
")
```



This graph is anther perspective of the relation among cancer, age and sex. It can be observed that at young age, female have higher chance to get cancer. While at old age, male have higher chance to get cancer.

3.6.3 Cancer ~ Age Relation across Education

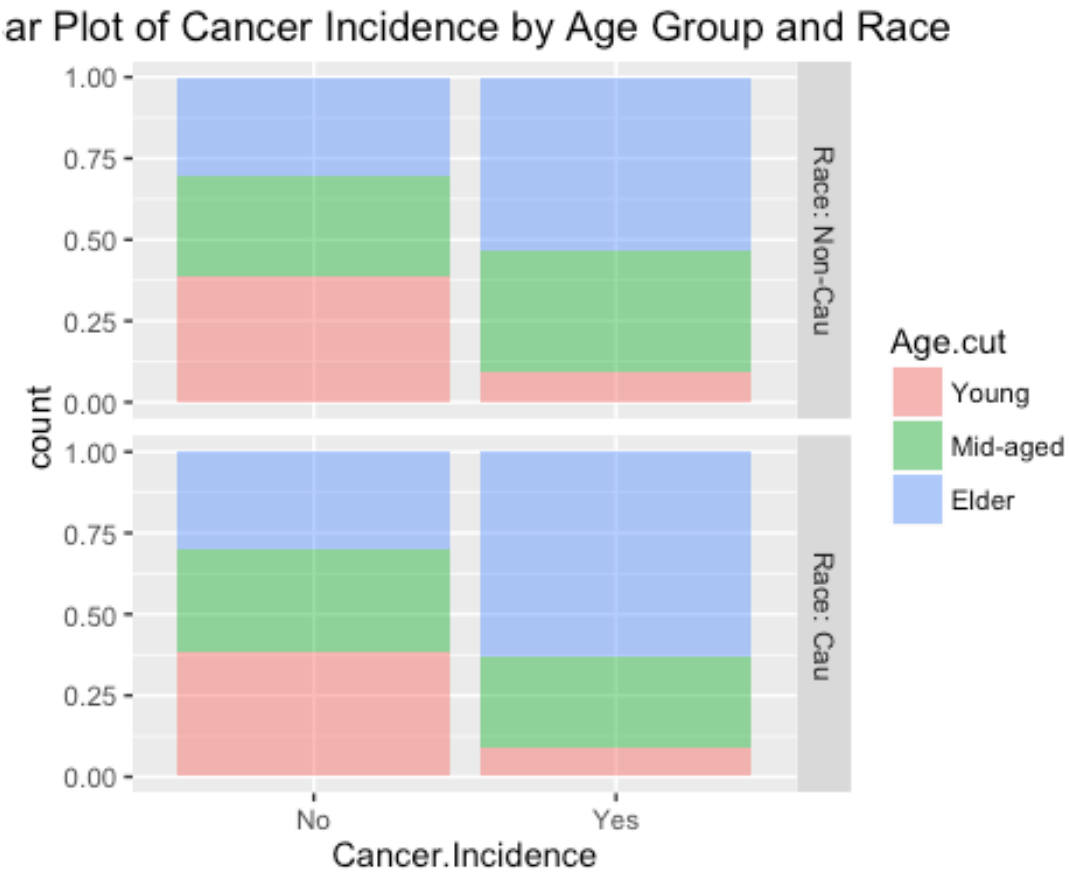
```
ggplot( NHANES, aes( x=Cancer.Incidence, fill=Age.cut )) + geom_bar(position="fill", alpha = 0.5) + facet_grid(Ed ~ . , labeller = label_both, scales = "free") +  labs(title = "Bar Plot of Cancer Incidence by Age Group and Education")
```



From this bar plot, it can be seen that the relation between Cancer and Age does hold across Education.

3.6.4 Cancer ~ Age Relation across Race

```
ggplot( NHANES, aes( x=Cancer.Incidence, fill=Age.cut )) + geom_bar(position=
"fill", alpha = 0.5) + facet_grid(Race ~ . , labeller = label_both, scales =
"free") +  labs(title = "Bar Plot of Cancer Incidence by Age Group and Race")
```

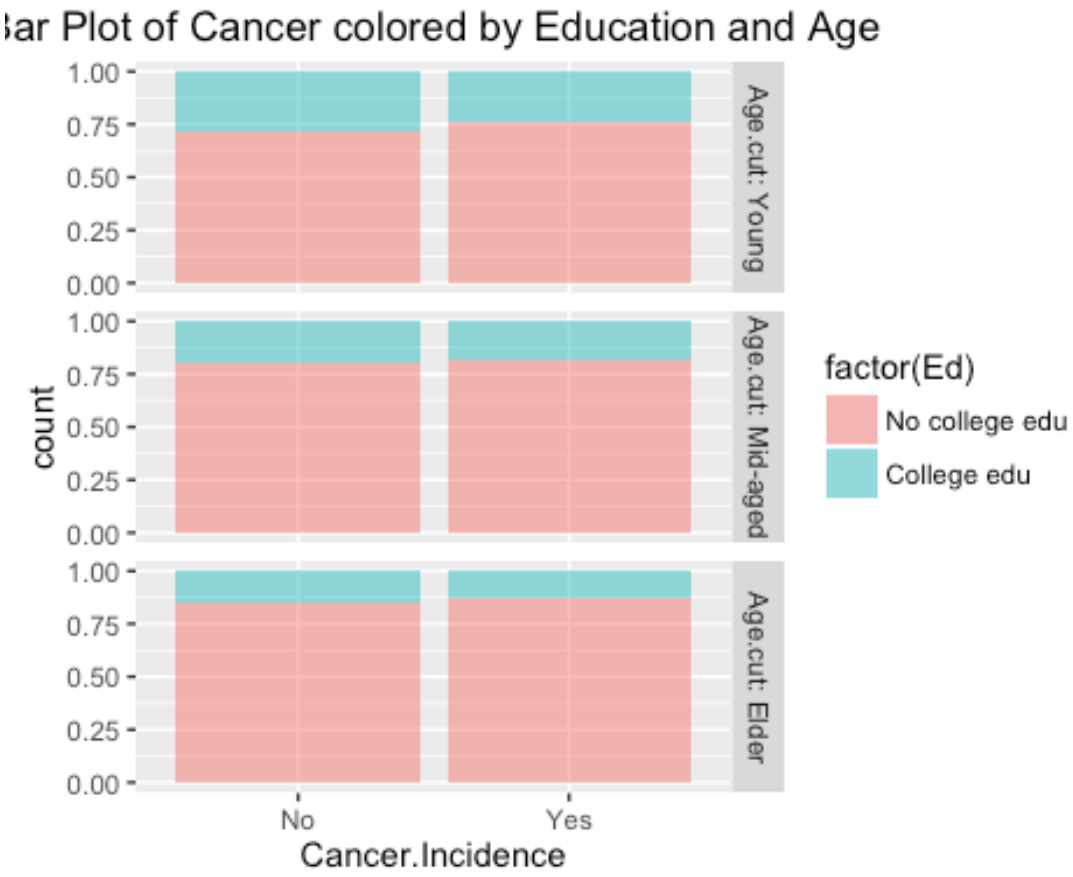


From this bar plot, it can be seen that the relation between Cancer and Age does hold for different races.

3.7 Relation between Cancer + Ed across other Factors

3.7.1 Cancer ~ Ed Relation across Age Group

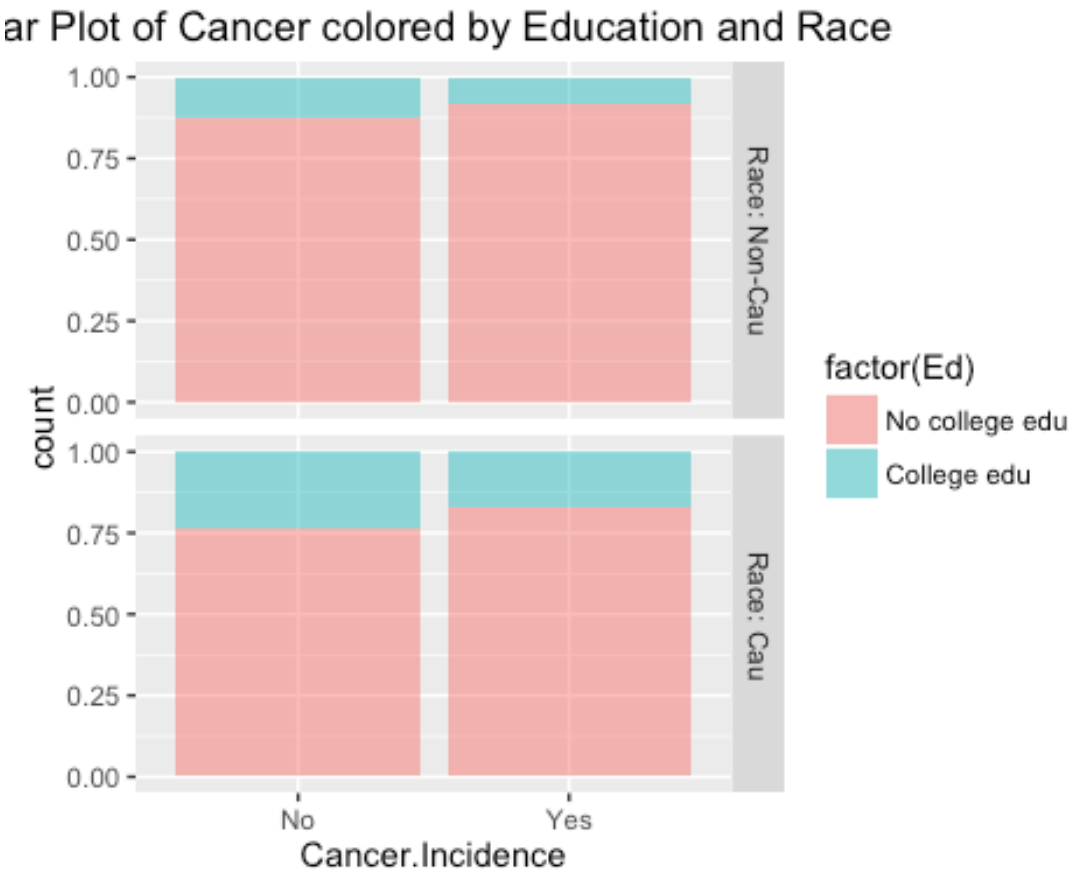
```
plt = ggplot( NHANES, aes( x = Cancer.Incidence, fill = factor(Ed) ))
plt + geom_bar(position="fill", alpha = 0.5) + facet_grid( Age.cut ~ . , labe
ller = label_both, scales = "free") + labs(title = "Bar Plot of Cancer colore
d by Education and Age")
```



It can be observed that for each age range, the relation between cancer and education still hold. More obvious trend can be seen in younger generation, which might be attributed to the fact that the ratio of educated people are higher in younger generation.

3.7.2 Cancer ~ Ed Relation across Race

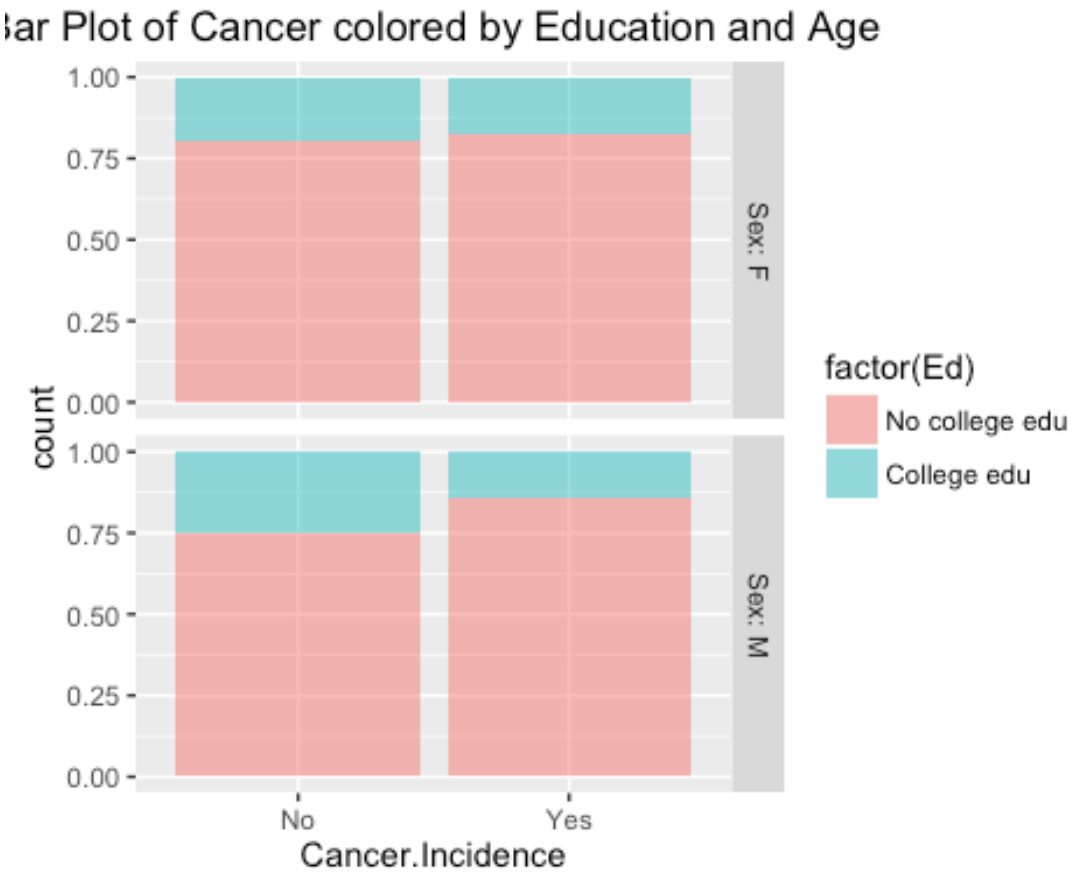
```
plt = ggplot( NHANES, aes( x = Cancer.Incidence, fill = factor(Ed) ))
plt + geom_bar(position="fill", alpha = 0.5) + facet_grid( Race ~ . , labelle
r = label_both, scales = "free") + labs(title = "Bar Plot of Cancer colored b
y Education and Race")
```



From this bar plot, it can be seen that the relation between Cancer and Education does hold for both races.

3.7.3 Cancer ~ Ed Relation across Sex

```
plt = ggplot( NHANES, aes( x = Cancer.Incidence, fill = factor(Ed) ))
plt + geom_bar(position="fill", alpha = 0.5) + facet_grid( Sex ~ . , labeller
 = label_both, scales = "free") + labs(title = "Bar Plot of Cancer colored by
 Education and Age")
```

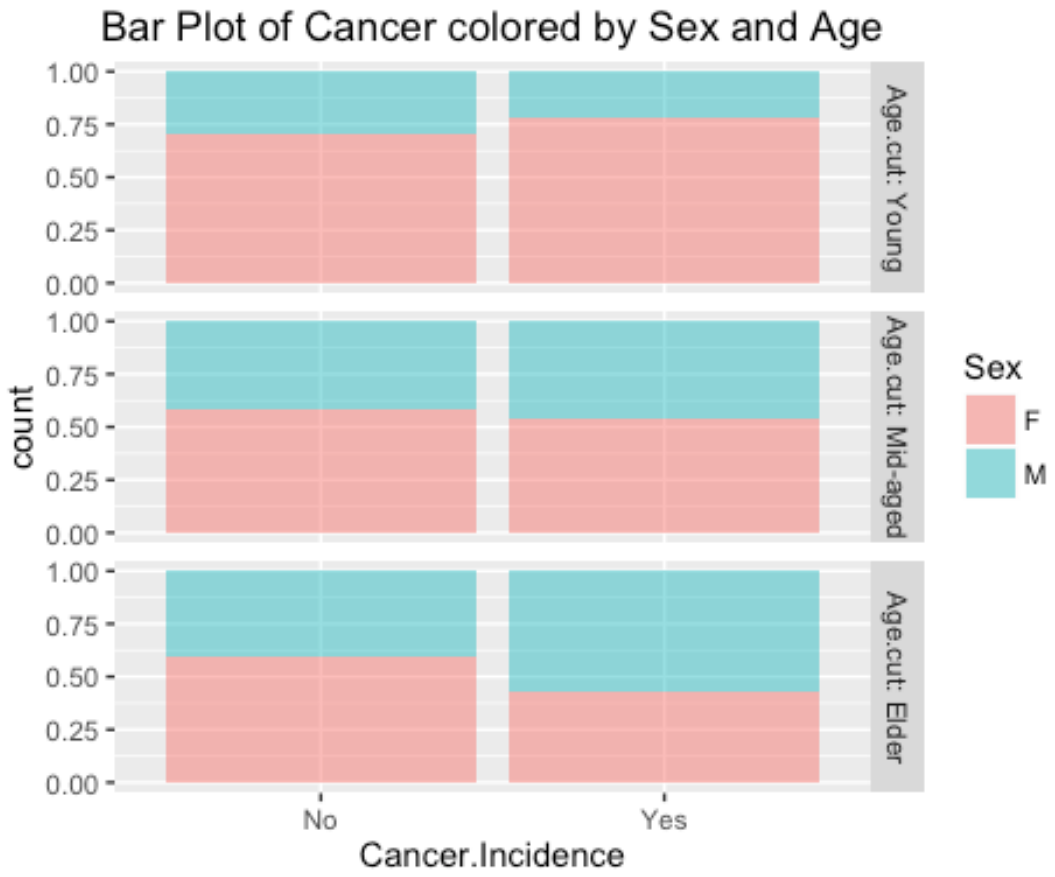


It can be seen from this bar plot that relation between cancer and education is independent of gender.

3.8 Relation between Cancer + Sex across other Factors

3.8.1 Cancer ~ Sex Relation across Age Group

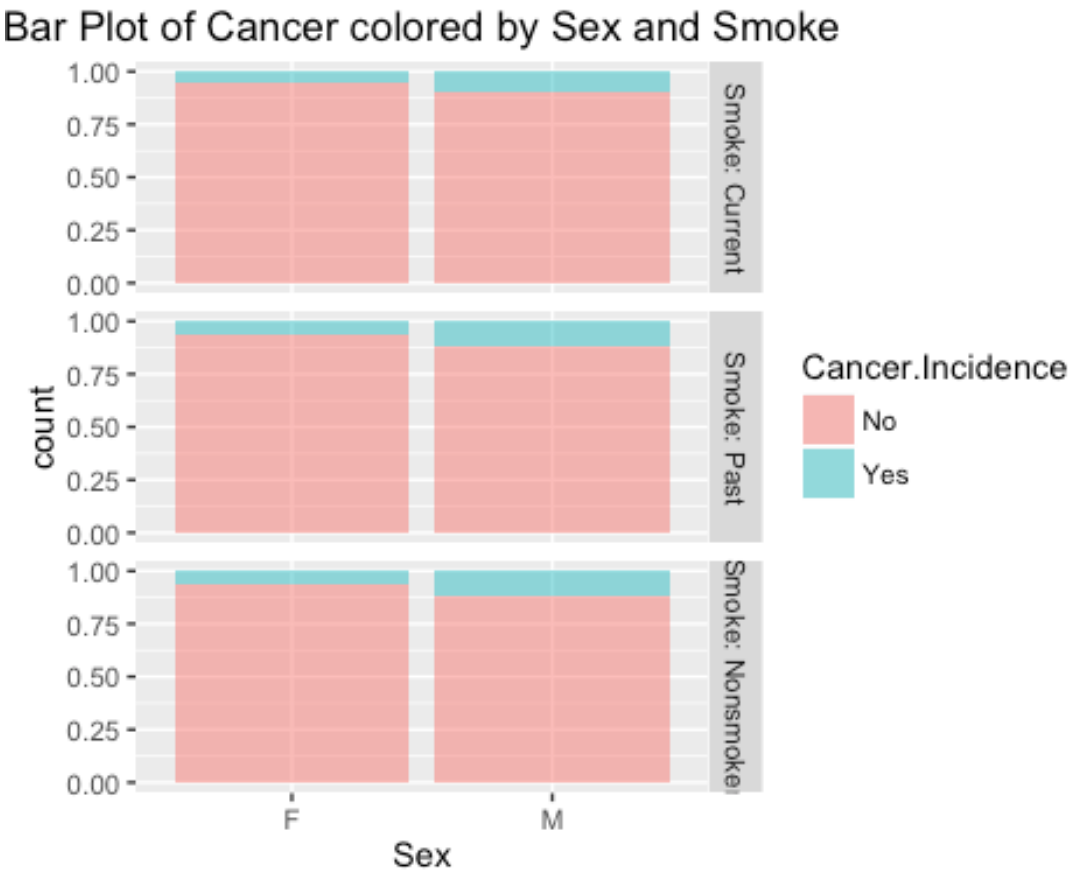
```
plt = ggplot( NHANES[NHANES$Smoke != "Unknown",], aes( x = Cancer.Incidence, fill = Sex ))
plt + geom_bar(position="fill", alpha = 0.5) + facet_grid( Age.cut ~ . , labeller = label_both, scales = "free") + labs(title = "Bar Plot of Cancer colored by Sex and Age")
```



The Cancer ~ Sex relation hold across Mid-aged and Elder group. However, in young group, the percentage of male who are diagnosed with cancer is lower than female. This is a similar conclusion as Section 3.6.2

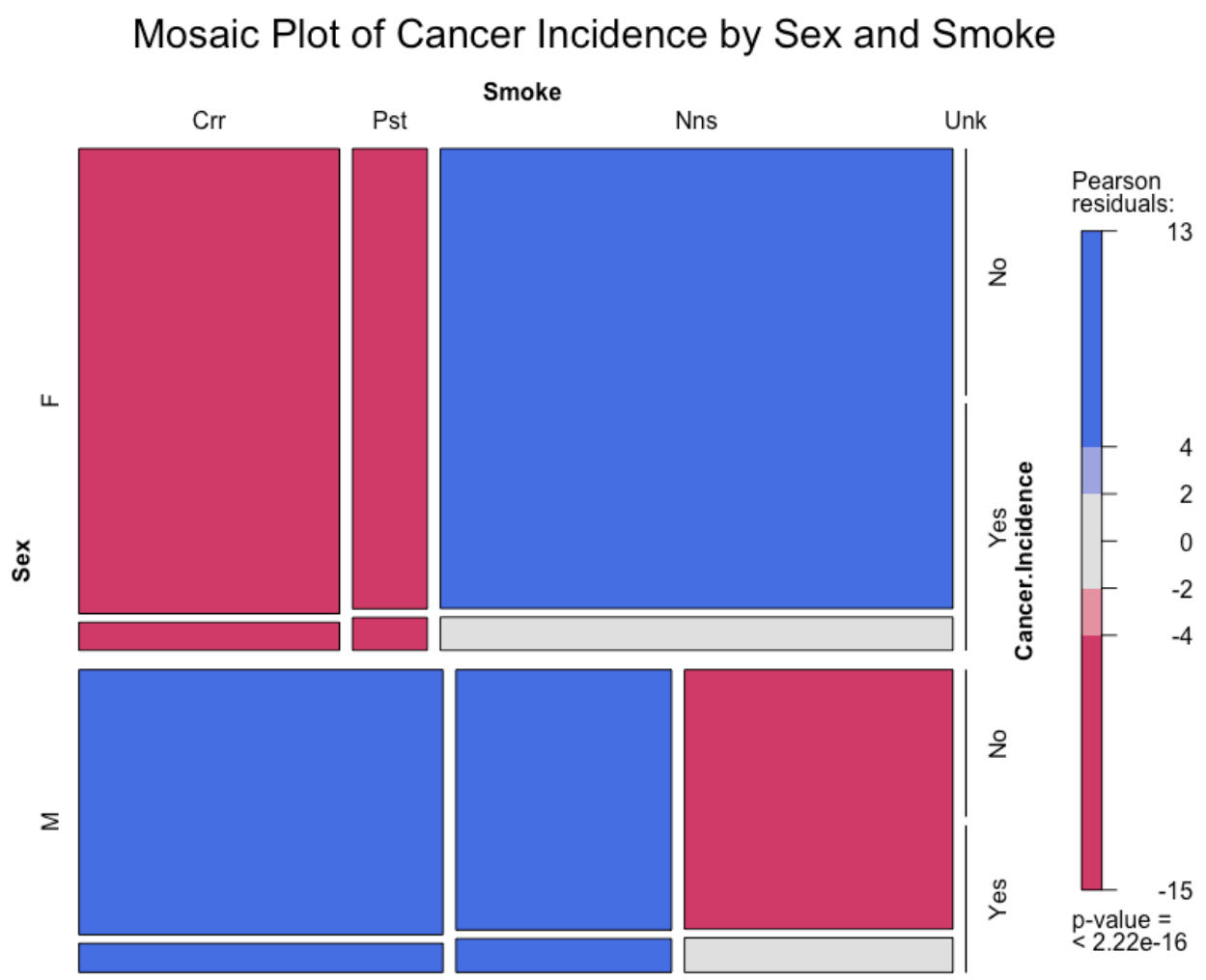
3.8.2 Cancer ~ Sex Relation across Smoke

```
plt = ggplot( NHANES[NHANES$Smoke != "Unknown",], aes( x = Sex, fill = Cancer
.Incidence ))
plt + geom_bar(position="fill", alpha = 0.5) + facet_grid( Smoke ~ . , labell
er = label_both, scales = "free") + labs(title = "Bar Plot of Cancer colored
by Sex and Smoke")
```

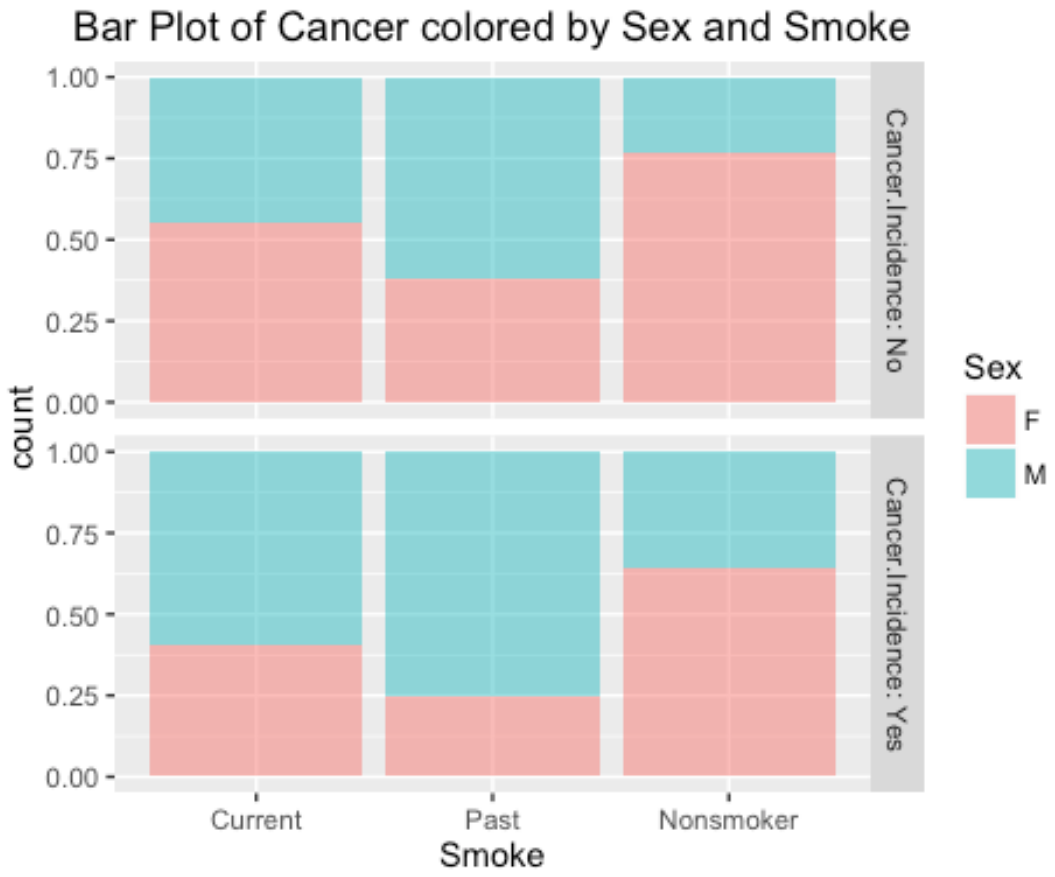


These graphs show that the fact that male are more likely to get cancer holds across different smoking groups.

```
mosaic(~ Sex + + Smoke + Cancer.Incidence, data = NHANES[NHANES$Smoke != "Unk
nown",], shade = TRUE, legend = TRUE, labeling_args=list(abbreviate=c(Age.cut
=1)), main = "Mosaic Plot of Cancer vs. Age and Smoke")
```



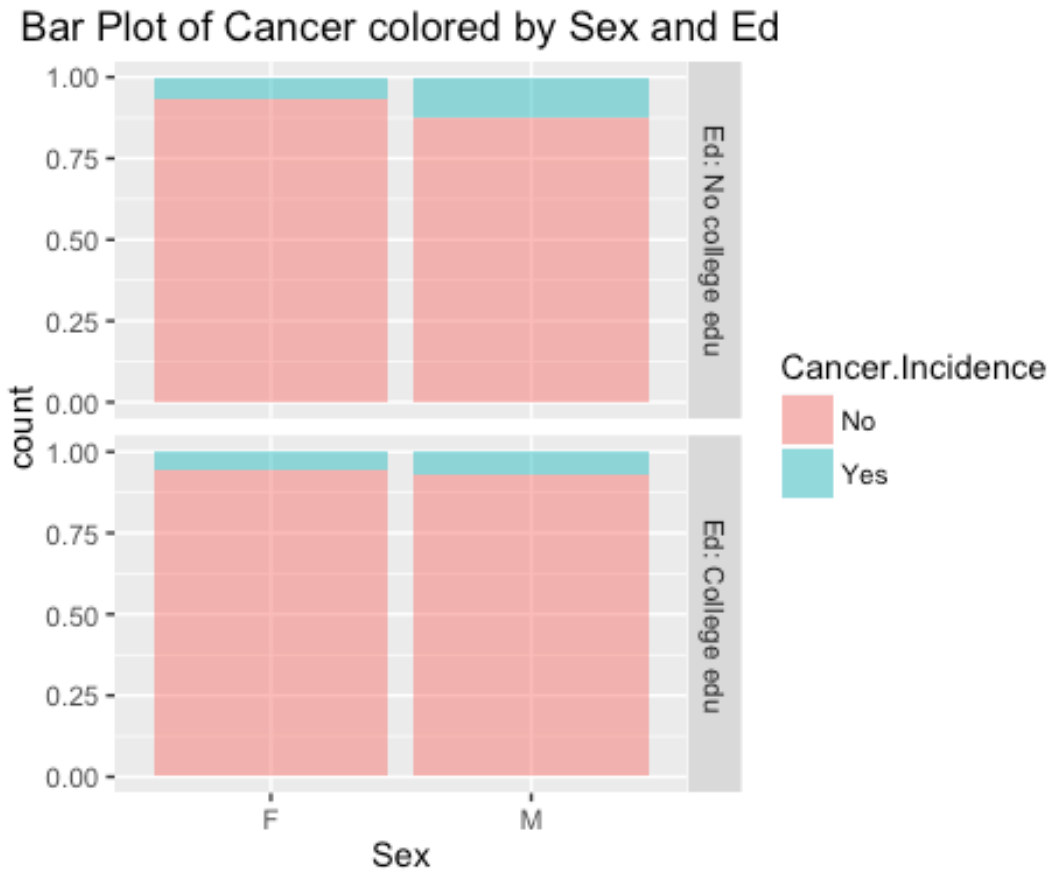

```
plt = ggplot( NHANES[NHANES$Smoke != "Unknown",], aes( x = Smoke, fill = Sex ))
plt + geom_bar(position="fill", alpha = 0.5) + facet_grid( Cancer.Incidence~ . , labeller = label_both, scales = "free") + labs(title = "Bar Plot of Cancer colored by Sex and Smoke")
```



From this plot, it can be seen that male who are currently smoking have higher cancer incidence.

3.8.3 Cancer ~ Sex Relation across Ed

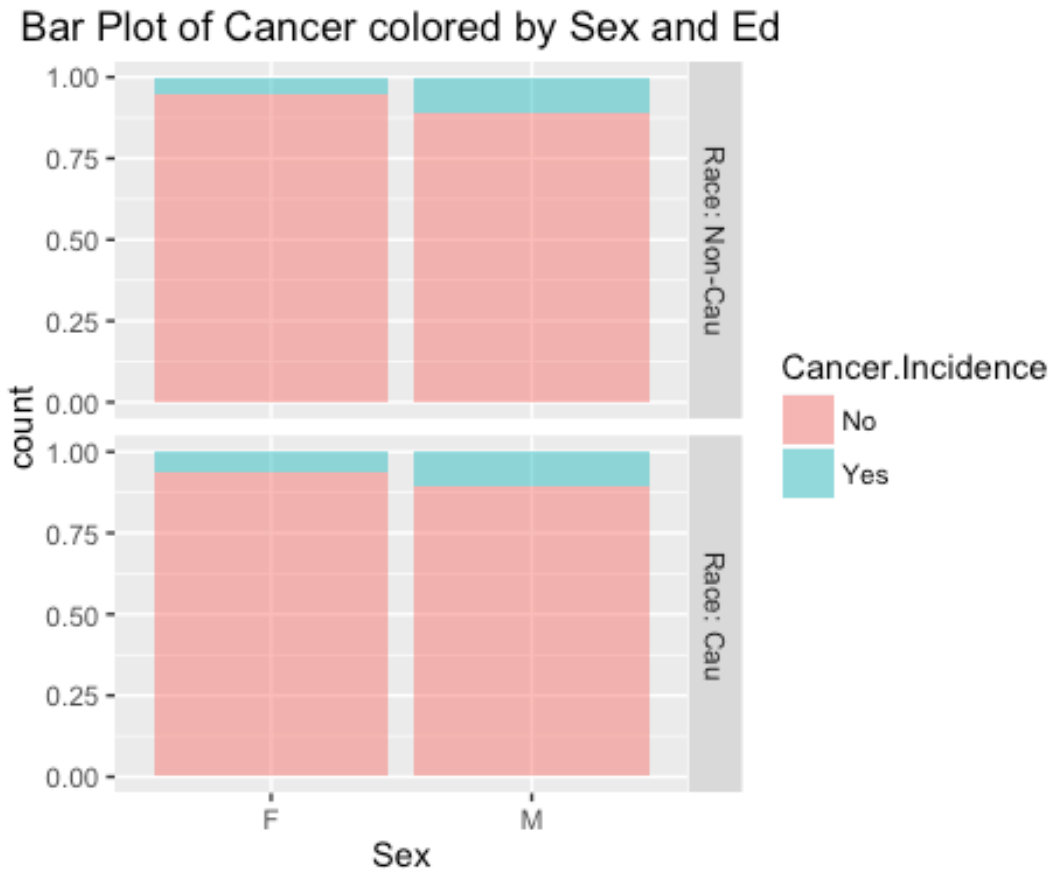
```
plt = ggplot( NHANES[NHANES$Smoke != "Unknown",], aes( x = Sex, fill = Cancer
.Incidence ))
plt + geom_bar(position="fill", alpha = 0.5) + facet_grid( Ed ~ . , labeller
= label_both, scales = "free") + labs(title = "Bar Plot of Cancer colored by
Sex and Ed")
```



From this graph, it seems that male people who are educated has lower chance to get cancer.

3.8.4 Cancer ~ Sex Relation across Race

```
plt = ggplot( NHANES[NHANES$Smoke != "Unknown",], aes( x = Sex, fill = Cancer
.Incidence ))
plt + geom_bar(position="fill", alpha = 0.5) + facet_grid( Race ~ . , labelle
r = label_both, scales = "free") + labs(title = "Bar Plot of Cancer colored b
y Sex and Ed")
```



Race doesn't have influence on the relation between cancer and sex.

3.9 Highlight of Findings

1. Age has a significant influence on cancer. Large amount of people who get cancer falls into the elder category.
2. The cancer incidence is independent of smoking status.
3. Education has an influence on Cancer. Lowest percentage of educated people are observed for people died from cancer. Higher percentage are observed for people who don't have cancer.
4. Gender is an important factor in cancer. Possible conclusion is that male are more likely to get cancer and die from cancer.
5. Smoking has an influence on the relation between age and cancer. For mid-aged and elder people, smoking habbit is likely to cause incidence of cancer.
6. Sex has an influence on the relation between age and cancer. For elder people, male have higher chance to get cancer. For young people, it is female who have higher chance to get cancer.

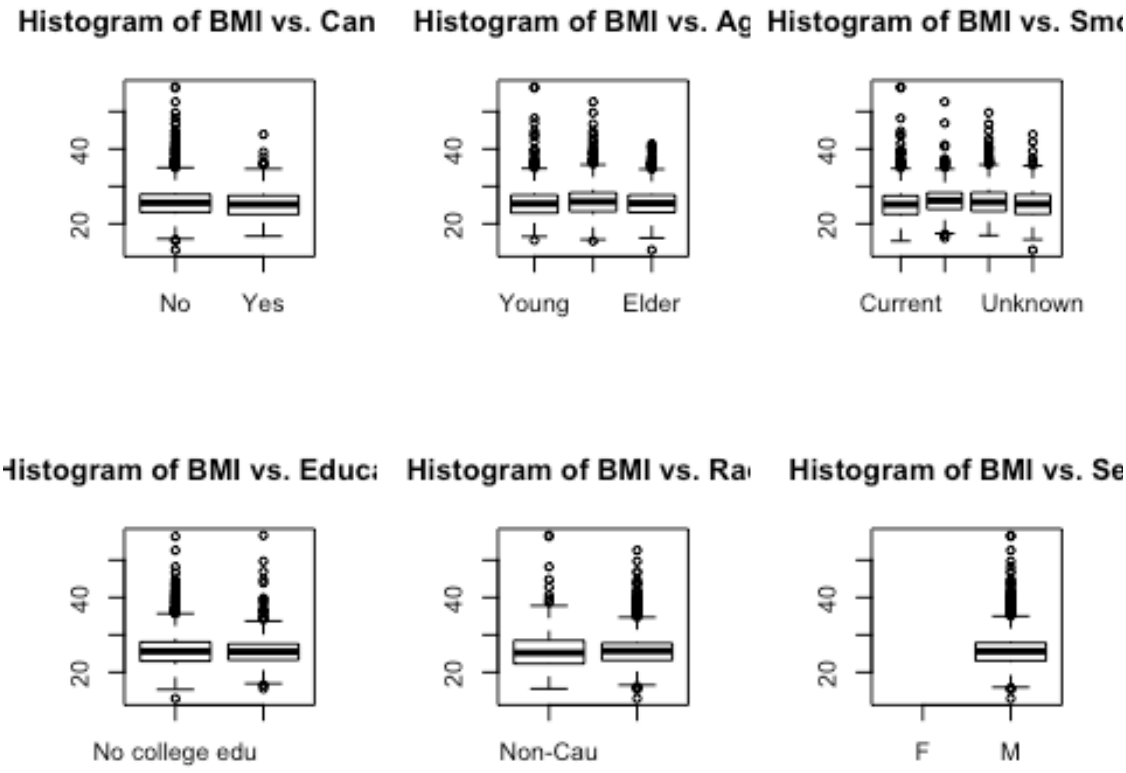
7. For male, it seems current male smokers have higher chance to die when they are diagnosed with cancer.

4 Prediction of BMI

4.1 Missing Values in BMI

Recall that over 60% of BMI value is missing.

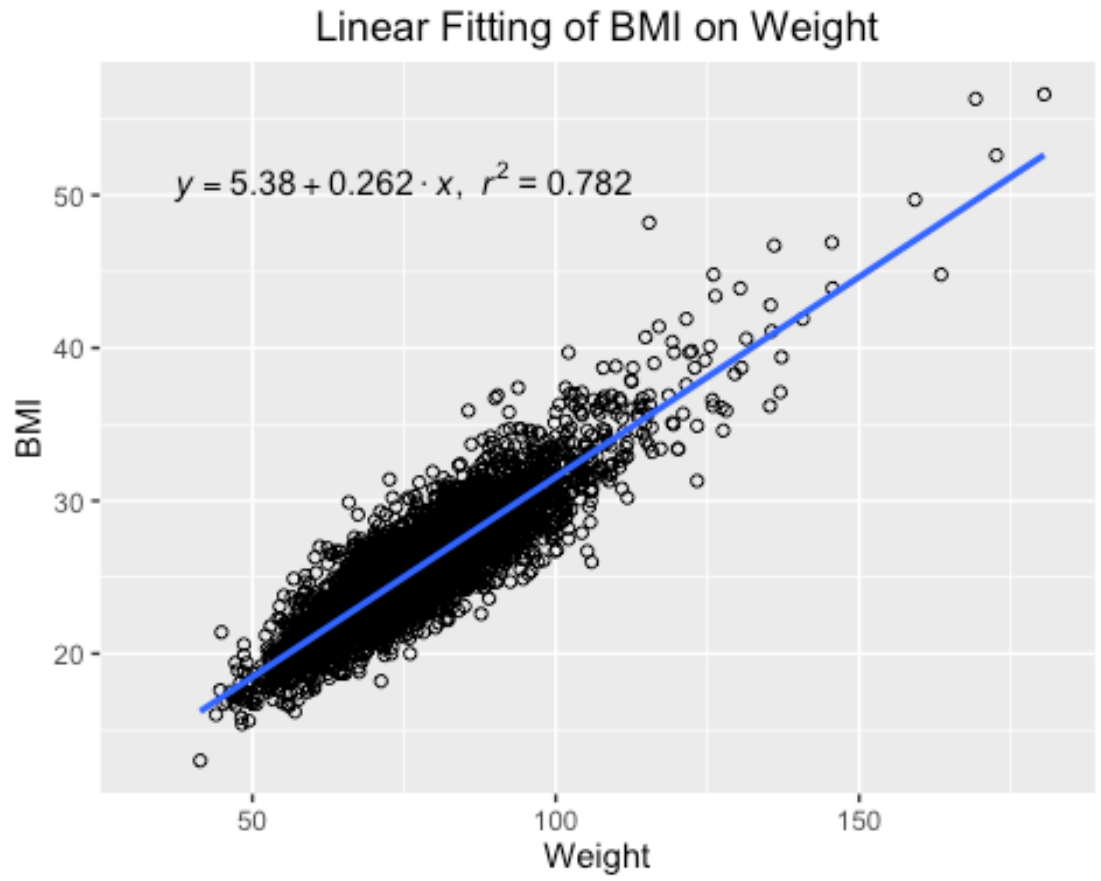
```
par(mfrow=c(2,3))
boxplot(BMI ~ Cancer.Incidence, data = NHANES, main = "Histogram of BMI vs. C
ancer")
boxplot(BMI ~ Age.cut, data = NHANES, main = "Histogram of BMI vs. Age")
boxplot(BMI ~ Smoke, data = NHANES, main = "Histogram of BMI vs. Smoke")
boxplot(BMI ~ Ed, data = NHANES, main = "Histogram of BMI vs. Education")
boxplot(BMI ~ Race, data = NHANES, main = "Histogram of BMI vs. Race")
boxplot(BMI ~ Sex, data = NHANES, main = "Histogram of BMI vs. Sex")
```



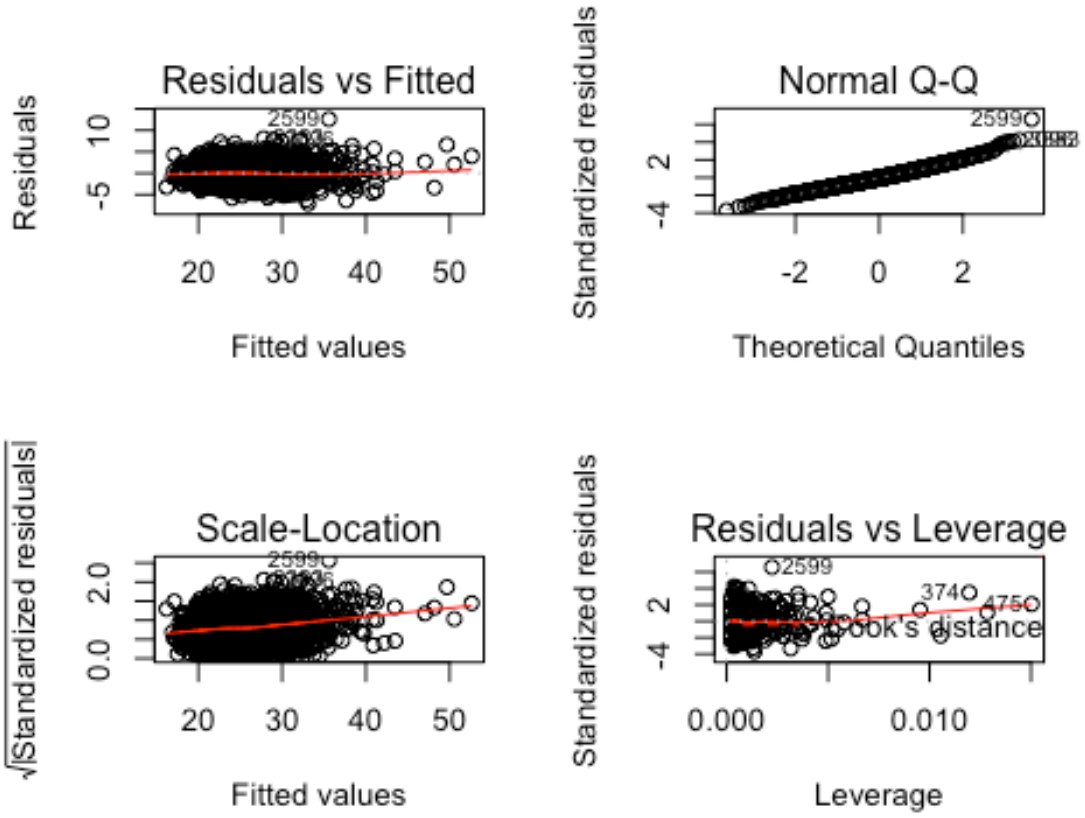
From these boxplot above, it can be observed that the BMI value of female are completely missing. This could because of the reason that the heights of female are not recorded. So here a possible way to predict the missing BMI of female is by using the almost linear relation between BMI and weight from male.

4.2 Linear Model on BMI ~ Weight

```
# install.packages("devtools")
library(devtools)
ggplot(NHANES, aes(x = Weight, y = BMI)) +
    geom_point(shape=1) +
    stat_smooth_func(geom="text",method="lm",hjust=0,parse=TRUE) +
    geom_smooth(method="lm", se=FALSE) + labs(title = "Linear Fitting of BMI
on Weight")
```



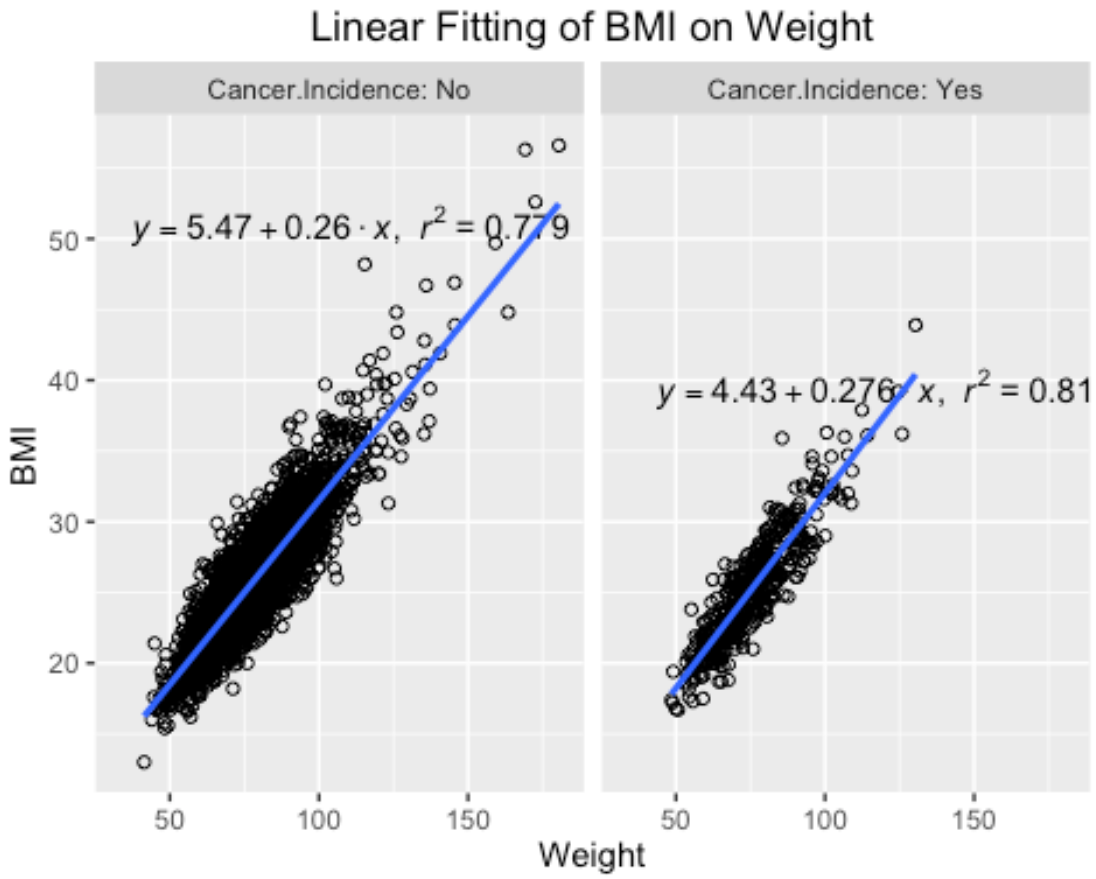

```
model_lm = lm(BMI ~ Weight)
par(mfrow = c(2,2))
plot(model_lm)
```



BMI value can be linear fitted on Weight. Linearity relation holds very well in the Weight range. However, the diagnosis plot shows that there is slightly violation on the normality and equal variance assumption. Since there are about 4000 sample, so this violation seems ok. There is no outlier in this sub-data set.

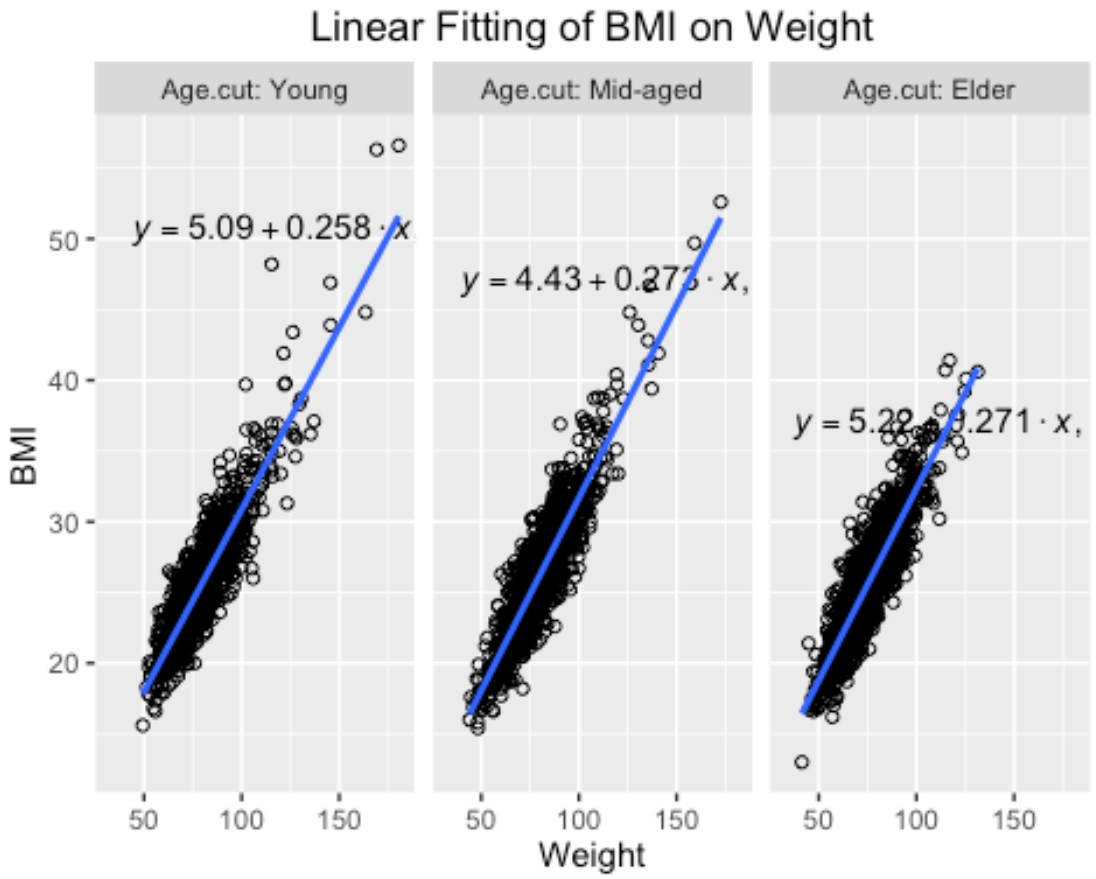
4.2.1 Linear Model on BMI ~ Weight seperated by Cancer

```
# install.packages("devtools")
library(devtools)
ggplot(NHANES, aes(x = Weight, y = BMI)) +
    geom_point(shape=1) +
    stat_smooth_func(geom="text",method="lm",hjust=0,parse=TRUE) +
    geom_smooth(method="lm", se=FALSE) +  facet_grid( . ~ Cancer.Incidence ,
labeller = label_both) + labs(title = "Linear Fitting of BMI on Weight")
```



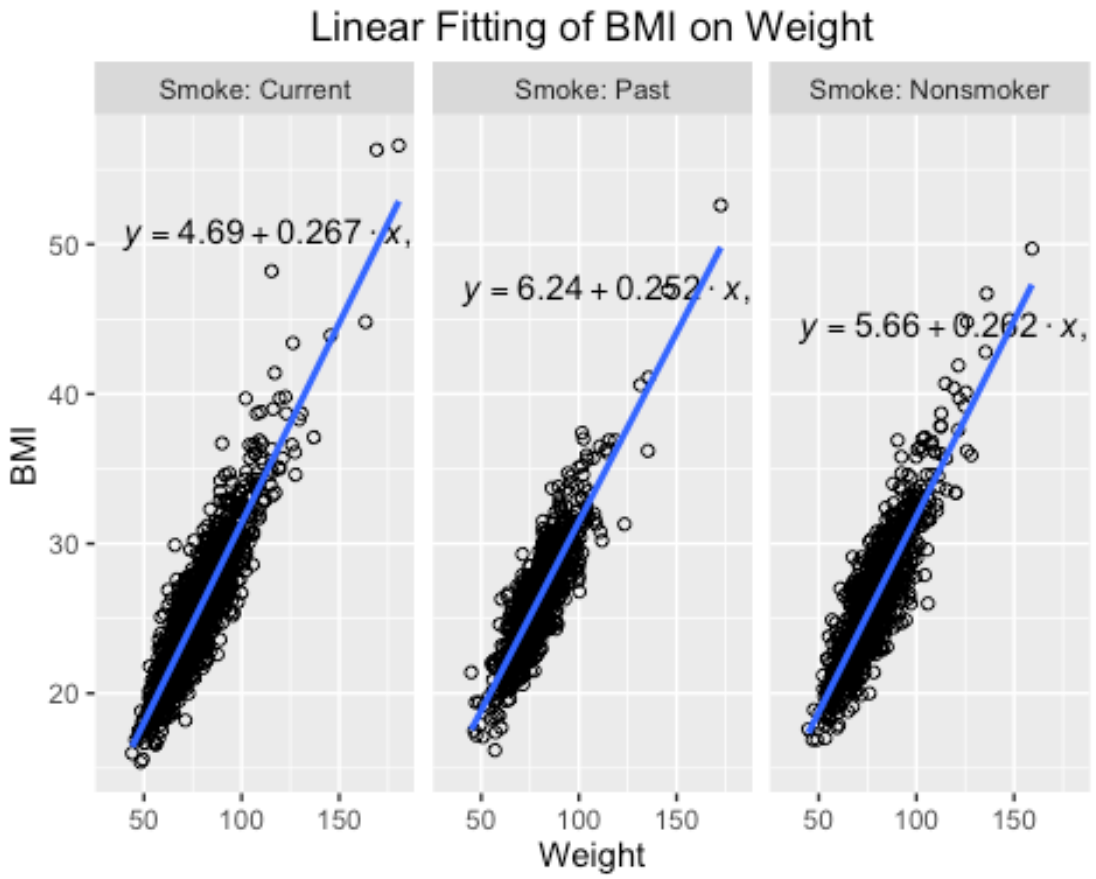
4.2.2 Linear Model on BMI ~ Weight seperated by Age

```
ggplot(NHANES, aes(x = Weight, y = BMI)) +
    geom_point(shape=1) +
    stat_smooth_func(geom="text",method="lm",hjust=0,parse=TRUE) +
    geom_smooth(method=lm, se=FALSE) +
    facet_grid( . ~ Age.cut , labeller = label_both) + labs(title = "Linear F
itting of BMI on Weight")
```



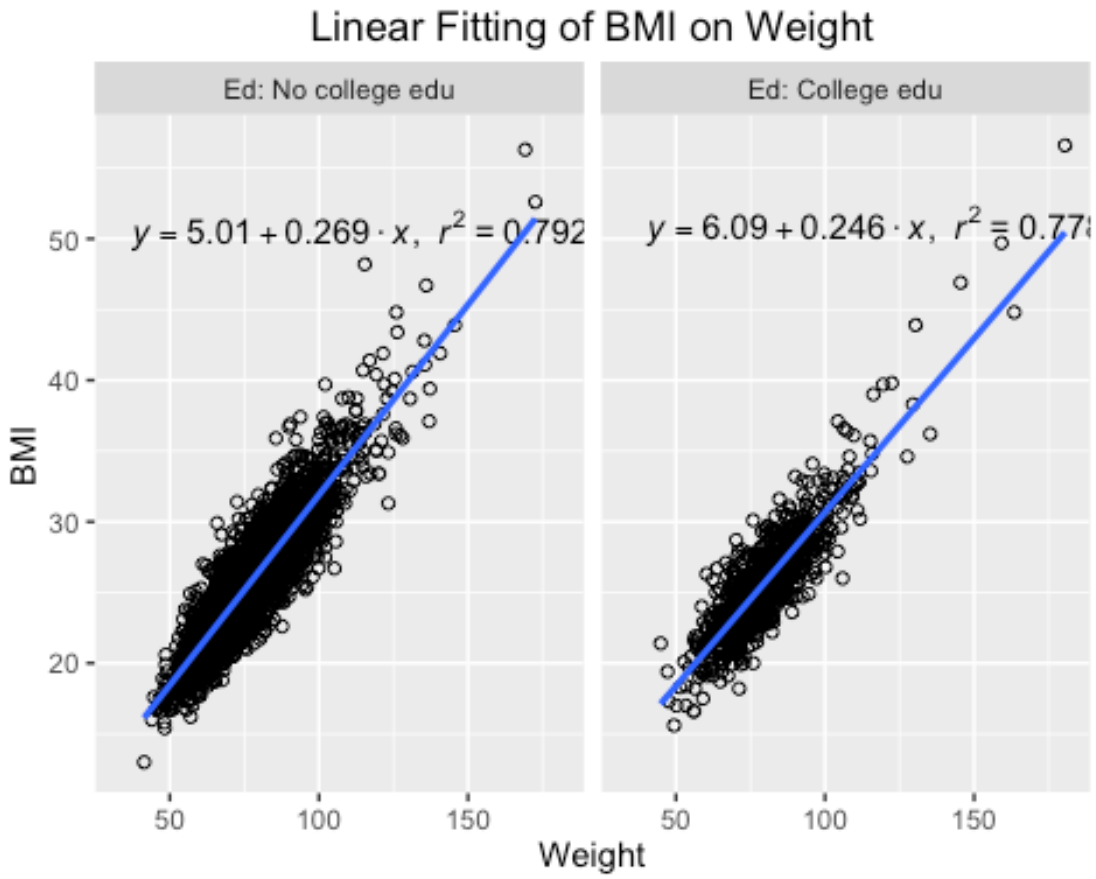
4.2.3 Linear Model on BMI ~ Weight seperated by Smoke

```
ggplot(NHANES[NHANES$Smoke != "Unknown",], aes(x = Weight, y = BMI)) +
    geom_point(shape=1) +
    stat_smooth_func(geom="text",method="lm",hjust=0,parse=TRUE) +
    geom_smooth(method=lm, se=FALSE) +
    facet_grid( . ~ Smoke , labeller = label_both) + labs(title = "Linear Fit
ting of BMI on Weight")
```



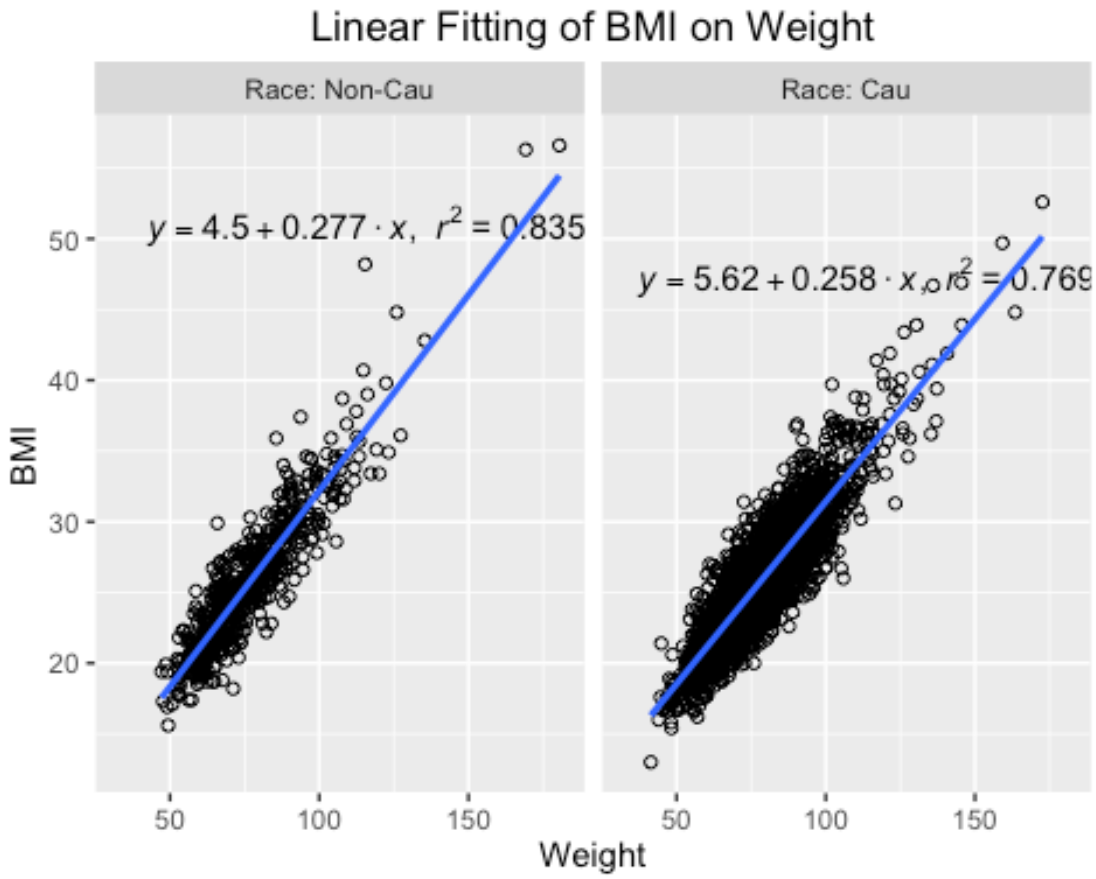
4.2.4 Linear Model on BMI ~ Weight seperated by Ed

```
ggplot(NHANES, aes(x = Weight, y = BMI)) +
    geom_point(shape=1) +
    stat_smooth_func(geom="text",method="lm",hjust=0,parse=TRUE) +
    geom_smooth(method=lm, se=FALSE) +
    facet_grid( . ~ Ed , labeller = label_both) + labs(title = "Linear Fittin
g of BMI on Weight")
```



4.2.5 Linear Model on BMI ~ Weight seperated by Race

```
ggplot(NHANES, aes(x = Weight, y = BMI)) +
    geom_point(shape=1) +
    stat_smooth_func(geom="text",method="lm",hjust=0,parse=TRUE) +
    geom_smooth(method=lm, se=FALSE) +
    facet_grid( . ~ Race , labeller = label_both) + labs(title = "Linear Fitt
ing of BMI on Weight")
```

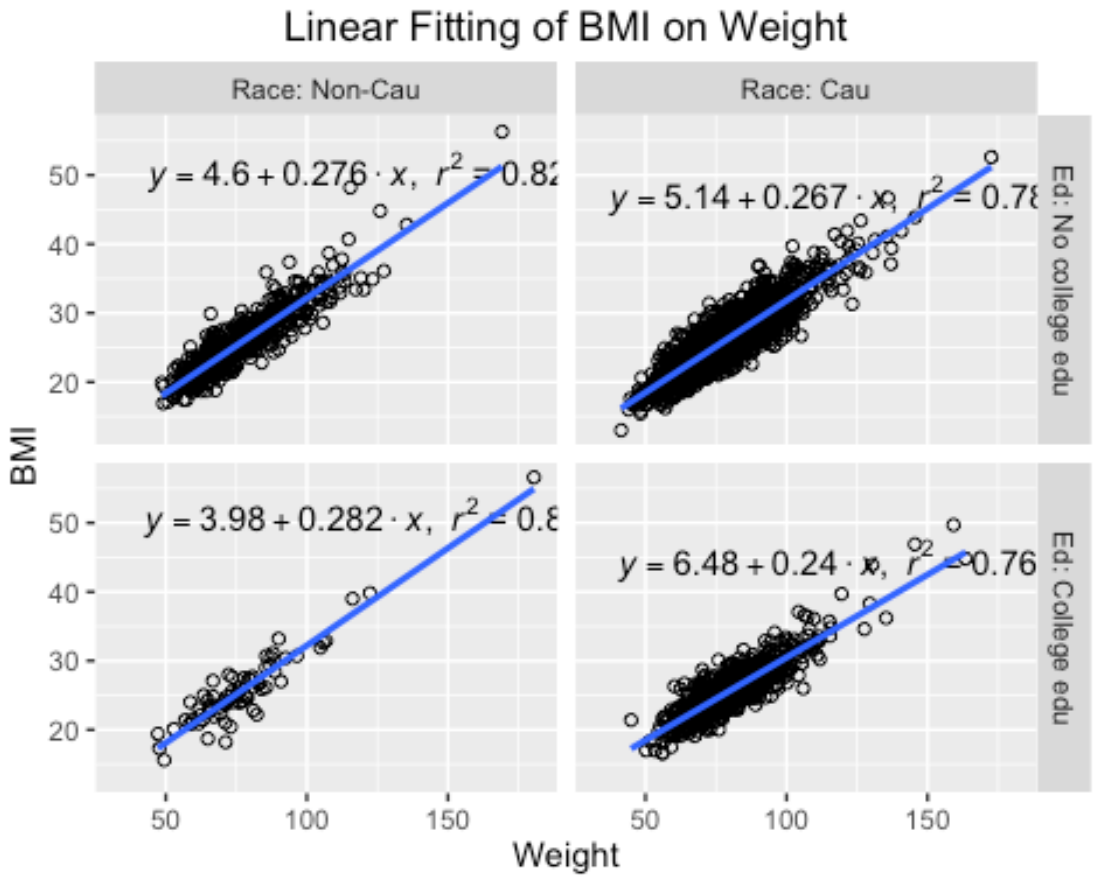


4.2.6 Linear Model on BMI ~ Weight seperated by One Factor

Seperated by one categorical factor, it can be observed there is not much difference in the slope when categorized by Cancer, Age or Smoke. However, when seperated by Education and Race, there are observable change in the slope of the linear model. Thus education and race are two important factors influencing the slope of this linear model (BMI ~ Weight). In the next section, the effect of education and race will be compared, because they are two most significant variables that can change the slope.

4.2.7 Linear Model on BMI ~ Weight seperated by Education and Race

```
ggplot(NHANES, aes(x = Weight, y = BMI)) +
    geom_point(shape=1) +
    stat_smooth_func(geom="text",method="lm",hjust=0,parse=TRUE) +
    geom_smooth(method=lm, se=FALSE) +
    facet_grid( Ed ~ Race , labeller = label_both) + labs(title = "Linear Fit
ting of BMI on Weight")
```



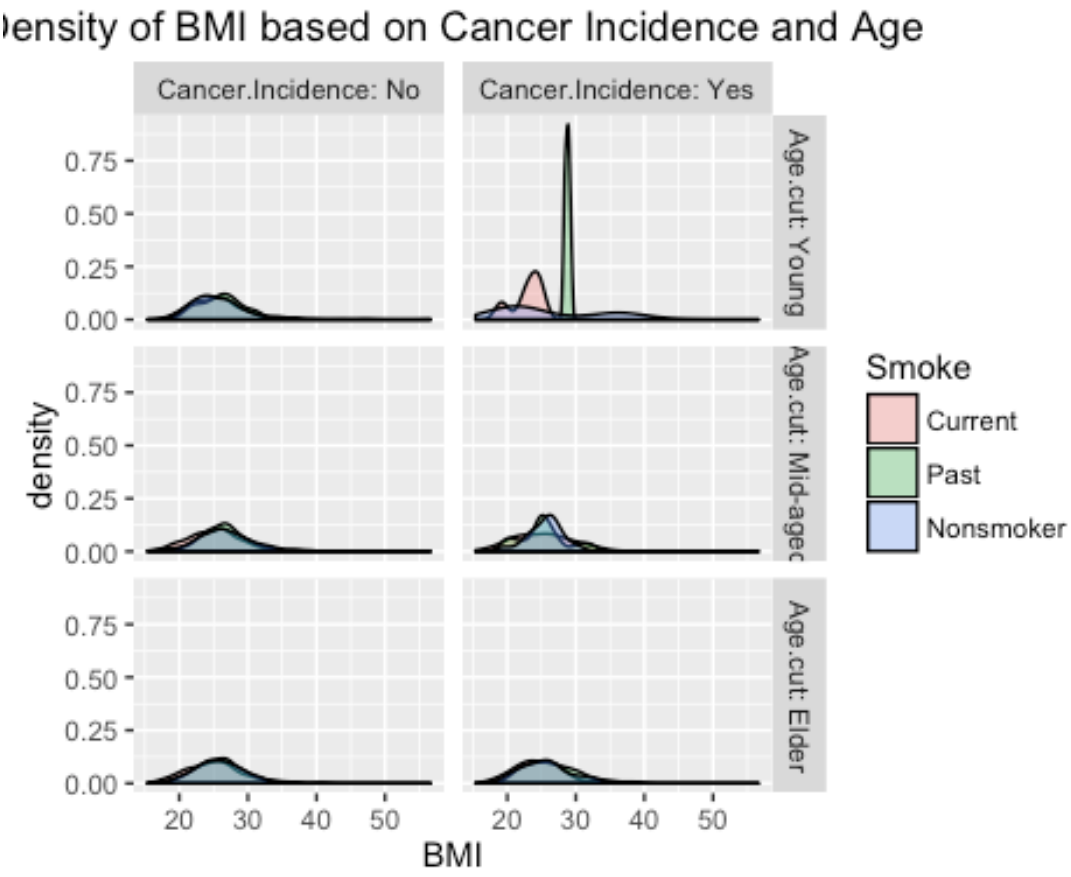
It can be seen from the graph that caucasian group generally have smaller slope than non-caucasian. People with college-degree as a whole have lower slope than those who don't have college-degree. However, for non-caucasian people, education factor is actually increasing the slope. This may be due to the reason that there are few people in non-caucasian/education group. Small sample size may cause the deviation from the linear model.

4.3 Density Plot of BMI

Another way to explore the factor of BMI is by looking at the spread and distribution of the density plots of BMI under different category. Below is one example to find this.

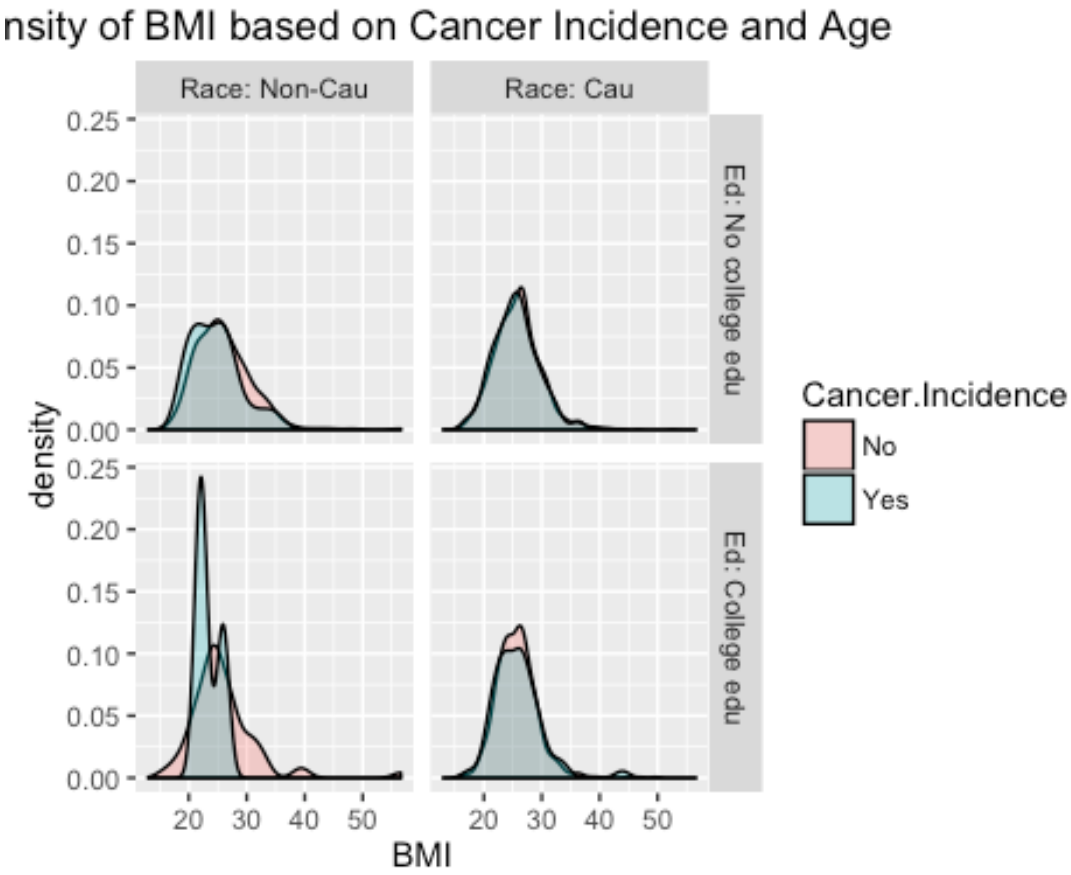
4.3.1 Density Plot of BMI by Cancer Age

```
ggplot(NHANES[NHANES$Smoke != "Unknown",], aes(BMI, fill = Smoke)) + geom_den
sity(alpha = 0.3)+ facet_grid(Age.cut ~ Cancer.Incidence , labeller = label_b
oth) + labs( title = "Density of BMI based on Cancer Incidence and Age" )
```



It can be observed that young non-smoker who are diagnosed with cancer are more narrowly distributed and the peak of this category falls at higher BMI value.

4.3.2 Density Plot of BMI by Cancer Ed Race

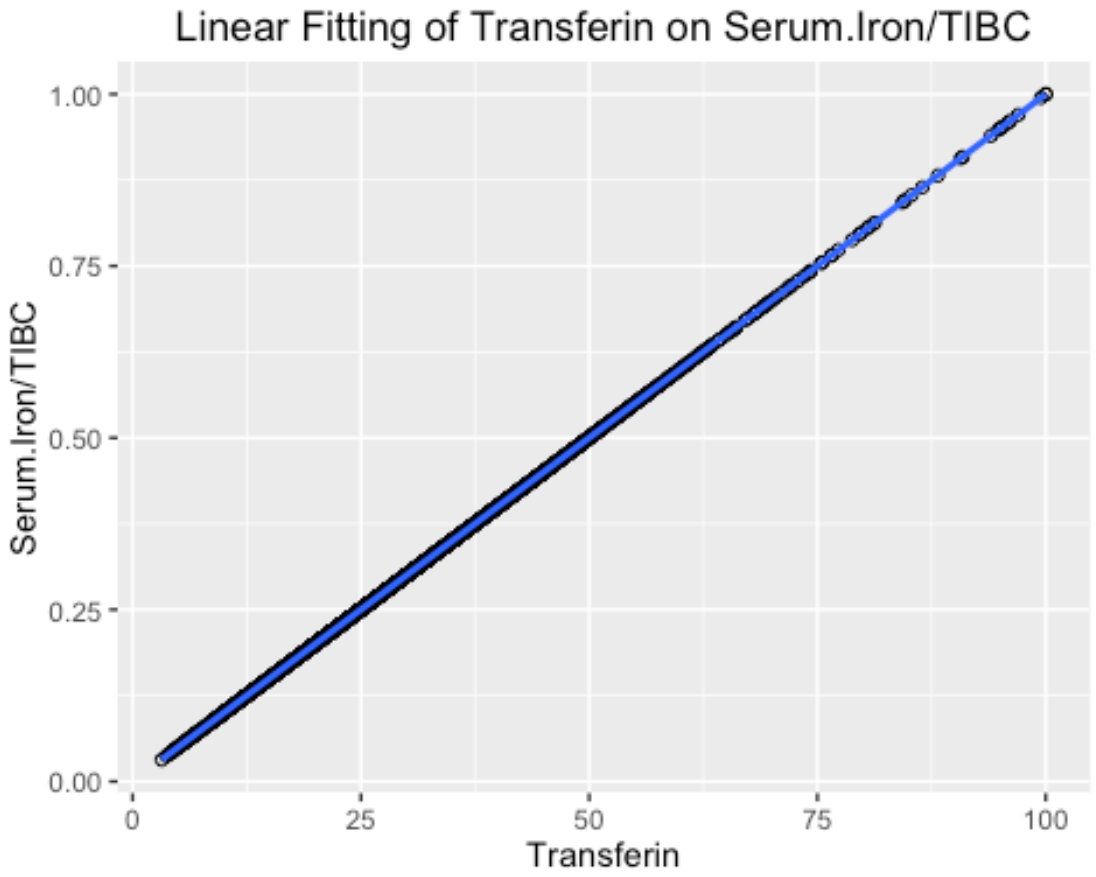


It can be observed that the BMI for people with college-degree or caucasian race are more narrowly distributed. This indicates that Education and Race are two important factors in prediction of BMI value.

5 Relation between Transferin and Serum.Iron/TIBC

5.1 Generating Linear Model on Transferin and Serum.Iron/TIBC

```
ggplot(NHANES, aes(x = Transferin, y = Serum.Iron/TIBC)) +
    geom_point(shape=1) +    # Use hollow circles
    geom_smooth(method=lm,
                se=FALSE) +    # No confidence interval
    labs(title = "Linear Fitting of Transferin on Serum.Iron/TIBC")
```



It can be seen clearly that Transferin follows equation "100 * Serum.Iron/TIBC" strictly.

Appendix

1. Source - https://gist.github.com/kdauria/524eade46135f6348140

```
library(devtools)
library(ggplot2)
stat_smooth_func <- function(mapping = NULL, data = NULL,
                         geom = "smooth", position = "identity",
                         ...,
                         method = "auto",
                         formula = y ~ x,
                         se = TRUE,
                         n = 80,
                         span = 0.75,
                         fullrange = FALSE,
                         level = 0.95,
                         method.args = list(),
                         na.rm = FALSE,
                         show.legend = NA,
                         inherit.aes = TRUE,
                         xpos = NULL,
                         ypos = NULL) {
  layer(
    data = data,
    mapping = mapping,
    stat = StatSmoothFunc,
    geom = geom,
    position = position,
    show.legend = show.legend,
    inherit.aes = inherit.aes,
    params = list(
      method = method,
      formula = formula,
      se = se,
      n = n,
      fullrange = fullrange,
      level = level,
      na.rm = na.rm,
      method.args = method.args,
      span = span,
      xpos = xpos,
      ypos = ypos,
      ...
    )
  )
}


StatSmoothFunc <- ggproto ("StatSmooth", Stat,
```

```
                        setup_params = function(data, params) {
                          # Figure out what type of smoothing to do: loess for
small datasets,
                          # gam with a cubic regression basis for large data
                          # This is based on the size of the _largest_ group.
                          if (identical(params$method, "auto")) {
                            max_group <- max(table(data$group))

                            if (max_group < 1000) {
                              params$method <- "loess"
                            } else {
                              params$method <- "gam"
                              params$formula <- y ~ s(x, bs = "cs")
                            }
                          }
                          if (identical(params$method, "gam")) {
                            params$method <- mgcv::gam
                          }

                          params
                        },

                        compute_group = function(data, scales, method = "auto",
 formula = y~x,
                                                 se = TRUE, n = 80, span = 0.75
, fullrange = FALSE,
                                                 xseq = NULL, level = 0.95, met
hod.args = list(),
                                                 na.rm = FALSE, xpos=NULL, ypos
=NULL) {
                          if (length(unique(data$x)) < 2) {
                            # Not enough data to perform fit
                            return(data.frame())
                          }

                          if (is.null(data$weight)) data$weight <- 1

                          if (is.null(xseq)) {
                            if (is.integer(data$x)) {
                              if (fullrange) {
                                xseq <- scales$x$dimension()
                              } else {
                                xseq <- sort(unique(data$x))
                              }
                            } else {
                              if (fullrange) {
                                range <- scales$x$dimension()
                              } else {
                                range <- range(data$x, na.rm = TRUE)
```

```
                            }
                            xseq <- seq(range[1], range[2], length.out = n)
                          }
                        }
                        # Special case span because it's the most commonly us
ed model argument
                        if (identical(method, "loess")) {
                          method.args$span <- span
                        }

                        if (is.character(method)) method <- match.fun(method)

                        base.args <- list(quote(formula), data = quote(data),
 weights = quote(weight))
                        model <- do.call(method, c(base.args, method.args))

                        m = model
                        eq <- substitute(italic(y) == a + b %.% italic(x)*","
~~italic(r)^2~"="~r2,
                                         list(a = format(coef(m)[1], digits =
 3),
                                              b = format(coef(m)[2], digits =
 3),
                                              r2 = format(summary(m)$r.square
d, digits = 3)))
                        func_string = as.character(as.expression(eq))

                        if(is.null(xpos)) xpos = min(data$x)*0.9
                        if(is.null(ypos)) ypos = max(data$y)*0.9
                        data.frame(x=xpos, y=ypos, label=func_string)

                      },

                      required_aes = c("x", "y")
)
```