Hw 2

Ryan Brady

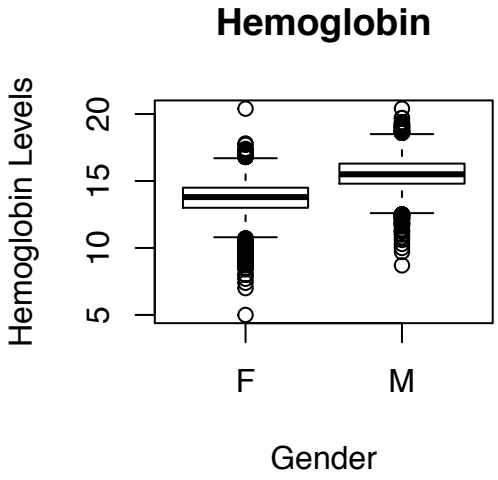
October 29, 2016

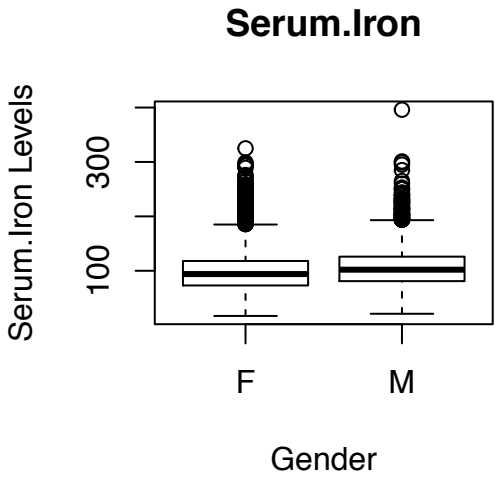
"The codes and results derived by using these codes constitute my own work. I have consulted the following resources regarding this assignment:" (Tzvi Gluck, Isaac Gluck)

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

## ResourceSelection 0.2-6   2016-02-15

## Warning: package 'corrgram' was built under R version 3.2.5
```

Cleaning The Data

The NHANES report contains many different variables, many of which have missing values. In order to run a statistical analysis of the data it is important to account for confounding variables. First we will look at sex to see if the different variables in the data will change for certain sex.





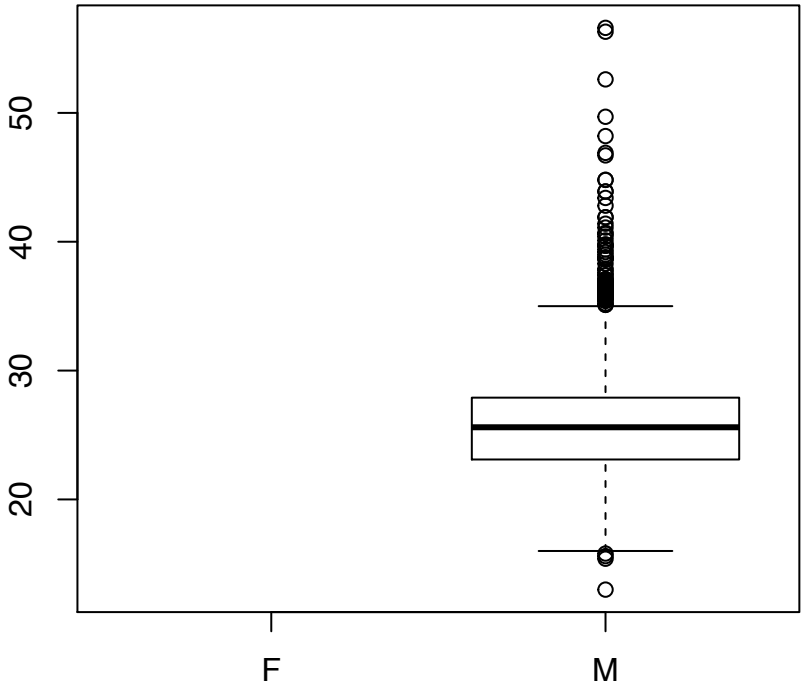
These two boxplots suggest that for males and females there are different levels for specific variables. When running a t.test we also see a significant difference for albumin levels.

```
##
##  Welch Two Sample t-test
##
## data:  male$Albumin and female$Albumin
## t = 11.845, df = 5948.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.07619993 0.10642432
## sample estimates:
## mean of x mean of y
##  4.429085  4.337773
```

On top of this there when looking at BMI levels All of the female BMI levels are listed as NA, therefore all predictions using the variable BMI must apply to males only.

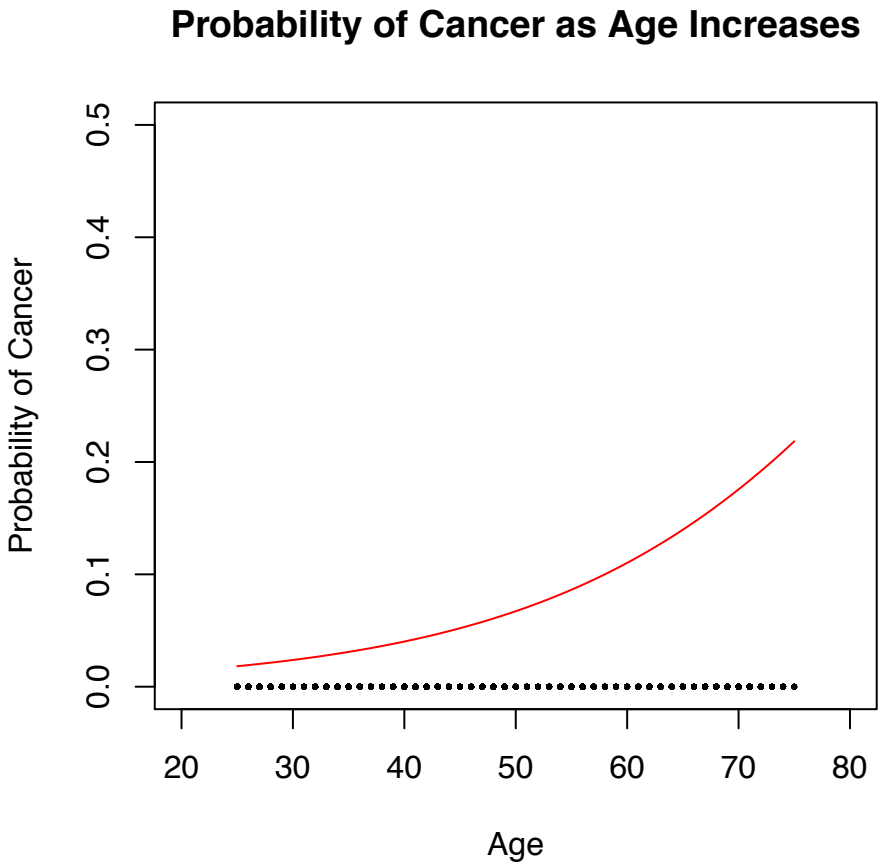
```
##
##            F    M
```

```
##   FALSE    0 3707
##   TRUE  5868    0
```



Splitting by Age Groups

In this report we also noticed that older people tended to obtain cancer more frequently and they also had a higher likilood of dieing from cancer.To show this, we split the age groups into three groups, young adults(20-40), middle aged(40-60), and elderly(60-80).

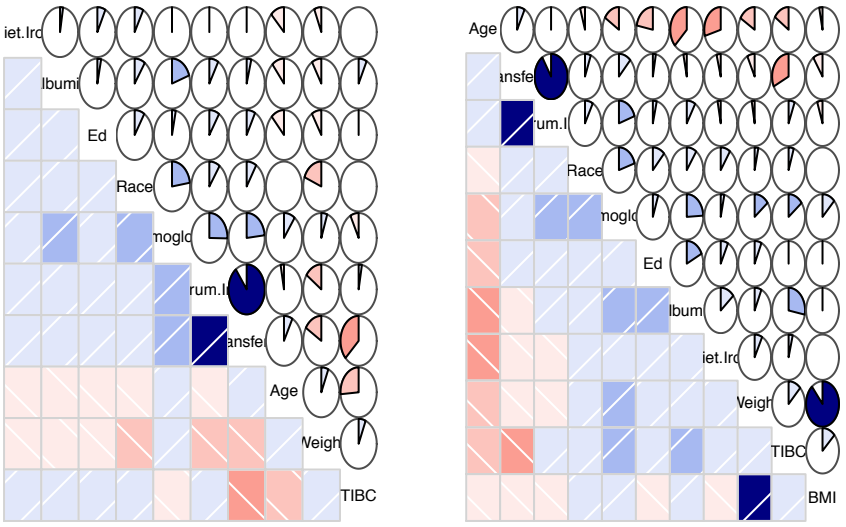


From the probability plot we see that as age increases there is an exponential growth in the probability that one attains cancer. Due to Age having a strong, linear influence on whether a person attains cancer or not, we found it useful to break Age into groups to hopefully lessen its affect as a confounding variable.

Analyzing the Data

First we will look at the distribution of the variables.

Correlogram Female Data Correlogram of Male Data



For each of the respective correelograms we see a strong relation between Transferin and Serum.Iron. Also for Males, it seems as though age has a stronger relation with the other variables than for females.

```
## Warning: Removed 1019 rows containing non-finite values (stat_bin).
```

From these plots, for each group, the mode for Transferin levels looks to be around 25 and most of the data is generally located around 25; however, there is a heavy skew right for each of the groups.

```
ggplot(NHANES,aes(x = Diet.Iron)) + geom_histogram(binwidth = 3,colour="#FF9999") + facet_wrap(~Age.cut
```

```
## Warning: Removed 141 rows containing non-finite values (stat_bin).
```


```
median(young.female$Diet.Iron,na.rm = TRUE)
```

```
## [1] 9.3
```

```
median(young.male$Diet.Iron,na.rm = TRUE)
```

```
## [1] 15.1
```

```
median(middle.age.f$Diet.Iron,na.rm = TRUE)
```

```
## [1] 9.2
```

```
median(middle.age.man$Diet.Iron,na.rm = TRUE)
```

```
## [1] 13.2
```

```
median(elder.female$Diet.Iron,na.rm = TRUE)
```

```
## [1] 8
```

```
median(elder.male$Diet.Iron,na.rm = TRUE)
```

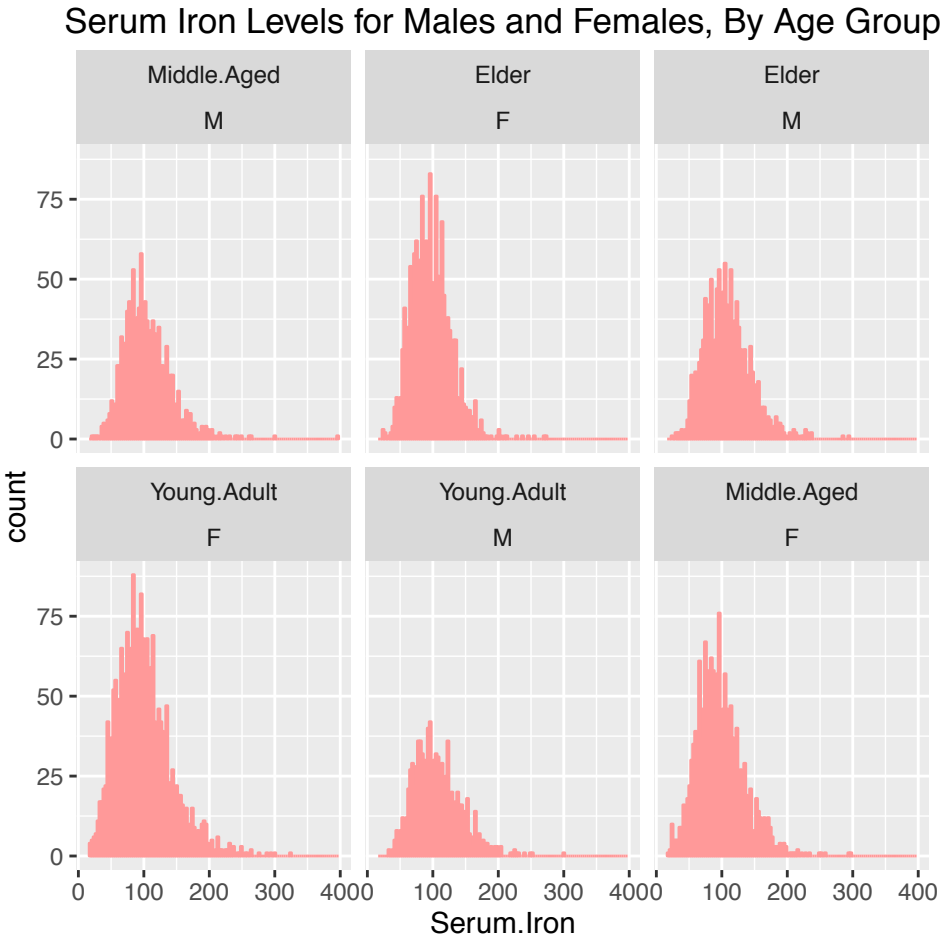
```
## [1] 10.8
```

For iron diet many of the graphs look fairly normal around the 10 mg mark; however young.males have a significantly wider spread than the rest of the groups. They have a median of 15.1 mgs, whereas the other groups are all about 11 mgs or lower.

```
## Warning: Removed 141 rows containing non-finite values (stat_bin).
```

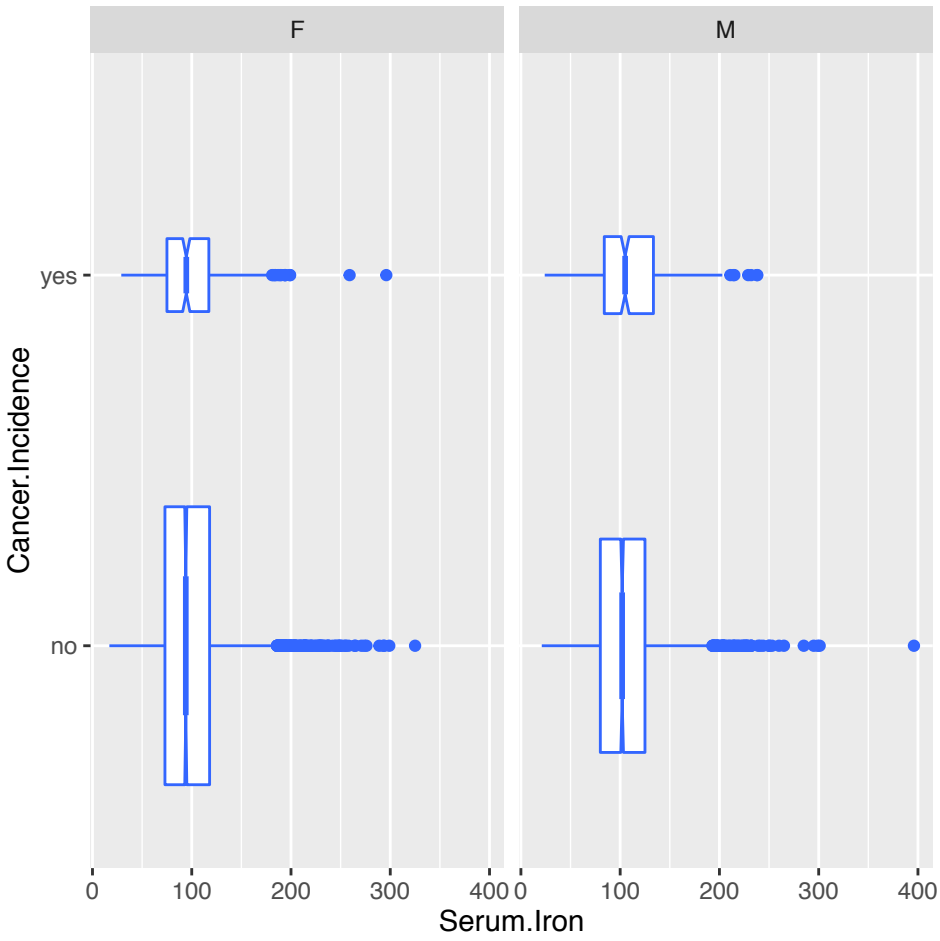
Next we look at Serum Iron levels.

```
## Warning: Removed 1008 rows containing non-finite values (stat_bin).
```



Similar to Transferin plots, the data is skewed to the right slightly but most of the data revolves around the median point of approximately 100. However, when looking at the extreme outliers for Serum.Iron there was no significant trend for people having higher serum levels and cancer.

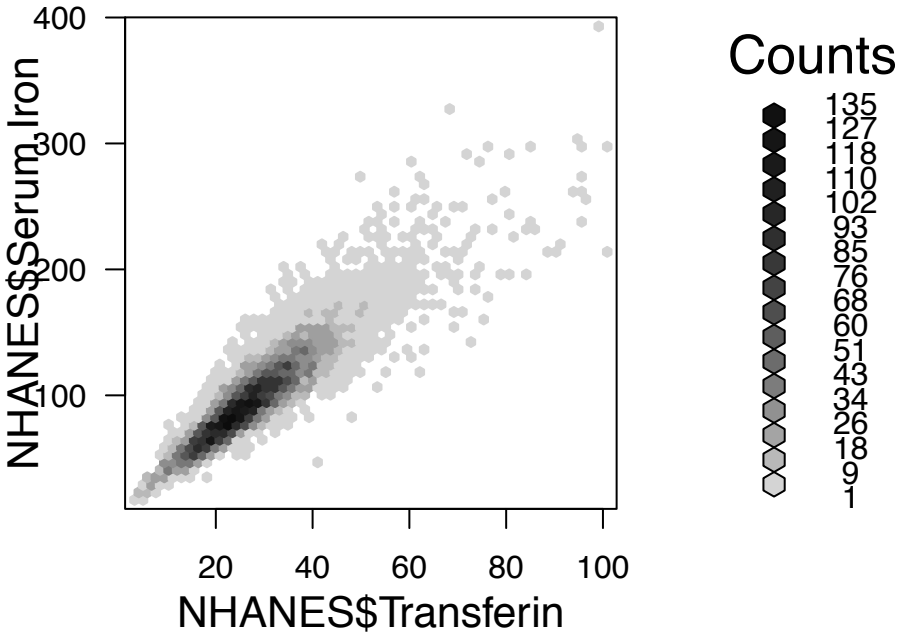
```
## Warning: Removed 1008 rows containing non-finite values (stat_boxplot).
```



Answering the Questions

1.

As seen in the corrleograms at the beginning of the report, there seems to be a strong relationship between Transferin and Serum.Iron for both males and females.

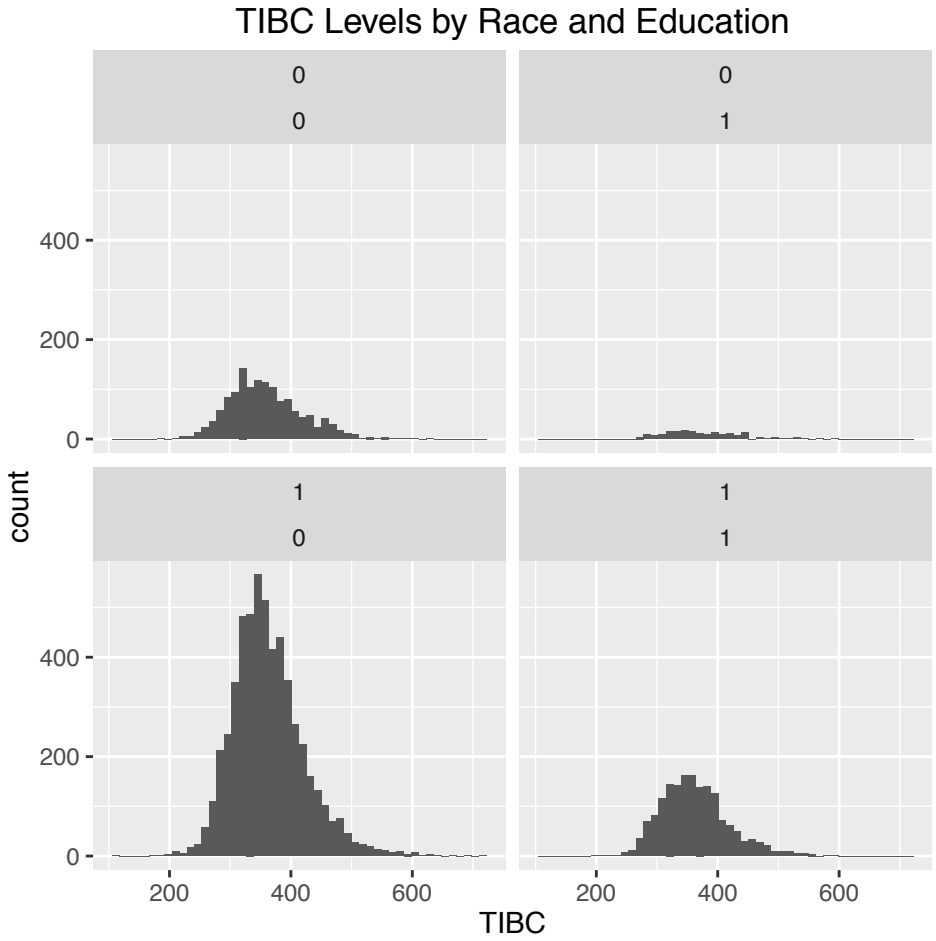


From the plot it seems there is a strong linear trend between the two variables as both are small, but as both variables increase, the variance of points increases. Also from the plots above we saw a small relationship between age and cancer.incidence as well as BMI for males and Weight for Males.

2.

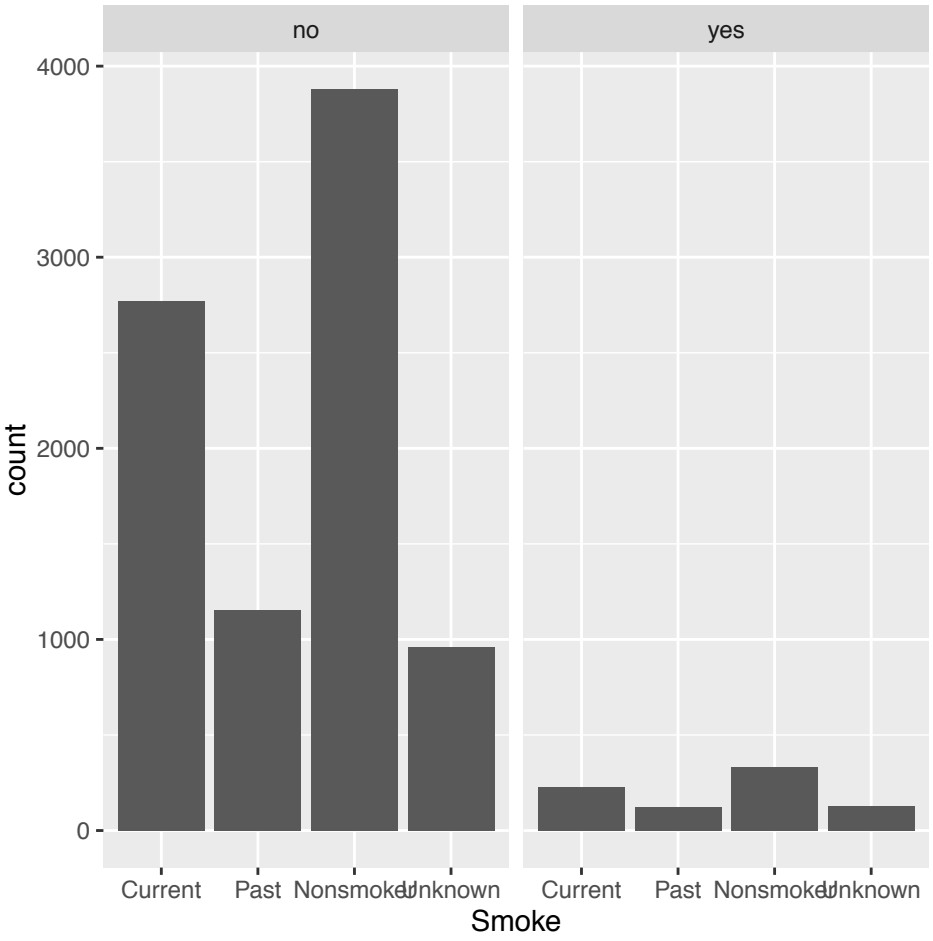
In the plots in the analyzing of data part of report, we see that sex has affects on variables such as Diet.Iron; however, for Serum Iron and Transferin there seems to be no relation to sex.

```
## Warning: Removed 853 rows containing non-finite values (stat_bin).
```



When looking at TIBC levels for people split into groups by education and race there seems to be little difference between the groups. All of the graphs have similar spread in their distribution and each has a unimodal peak at around 355.

3



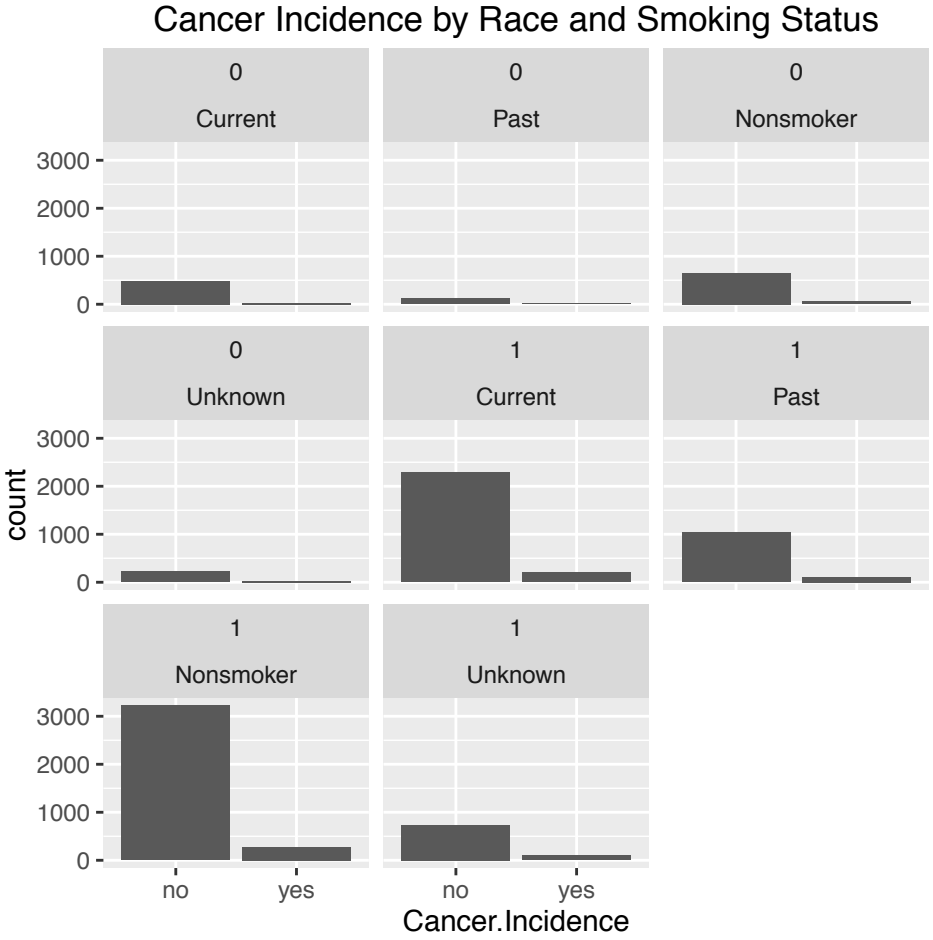
It appears that current smokers have about a 2.2 higher likelihood of having cancer than nonsmokers do. While this looks to be a strong relation, when applying a generalized linear model to see how smoking affects a persons likelihood of obtaining cancer,we see that the model is only effective at predicting the cancer rates for males almost 55 percent of the time, and females about 53 percent of the time.

```
##
## Call:
## roc.default(response = smoke.mod.m$y, predictor = smoke.mod.m$fitted.values,     auc = TRUE, ci = TRU
##
## Data: smoke.mod.m$fitted.values in 2571 controls (smoke.mod.m$y 0) < 324 cases (smoke.mod.m$y 1).
## Area under the curve: 0.5488
## 95% CI: 0.5164-0.5812 (DeLong)
```

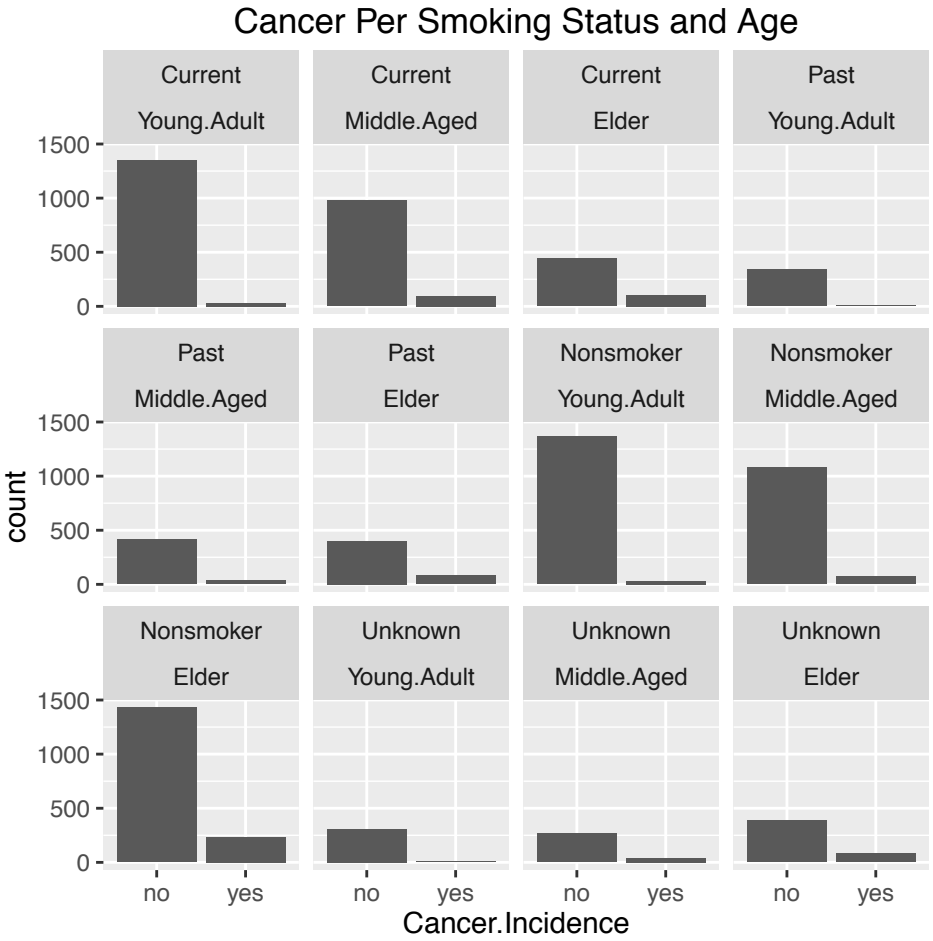
```
##
## Call:
## roc.default(response = smoke.mod.f$y, predictor = smoke.mod.f$fitted.values,     auc = TRUE, ci = TRU
##
## Data: smoke.mod.f$fitted.values in 4337 controls (smoke.mod.f$y 0) < 296 cases (smoke.mod.f$y 1).
## Area under the curve: 0.5334
## 95% CI: 0.5028-0.5639 (DeLong)
```

4

Another key interest is to see if Race will affect how Smoking Status affects Cancer.Incidence. From the bar plot though it is hard to see if there is a signficant difference in the proportions of whether someone has Cancer or not.



Next we will look to see if age groups affects how smoking status predicts Cancer Rates.

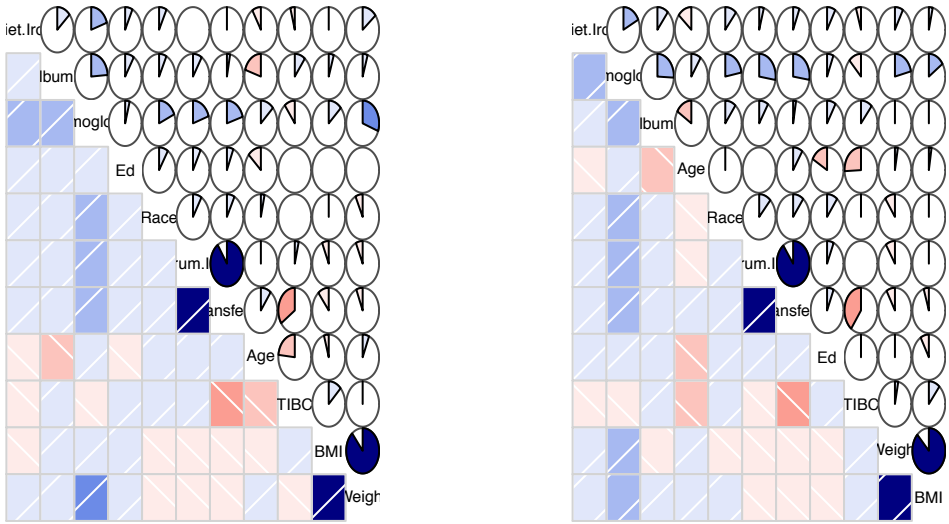


These bar plots suggest Young adults have a much lower likelihood of obtaining cancer than older adults. Also these bar plots reveal that there appears a relationship between smoking, age, and cancer rates; but as stated above in the report, these factors are not strong enough to predict whether somone will have cancer or not.

5

For this problem we will subset groups by race, age, and smoking status and look at there corrleogram to see how variable relationsions change over different groups.

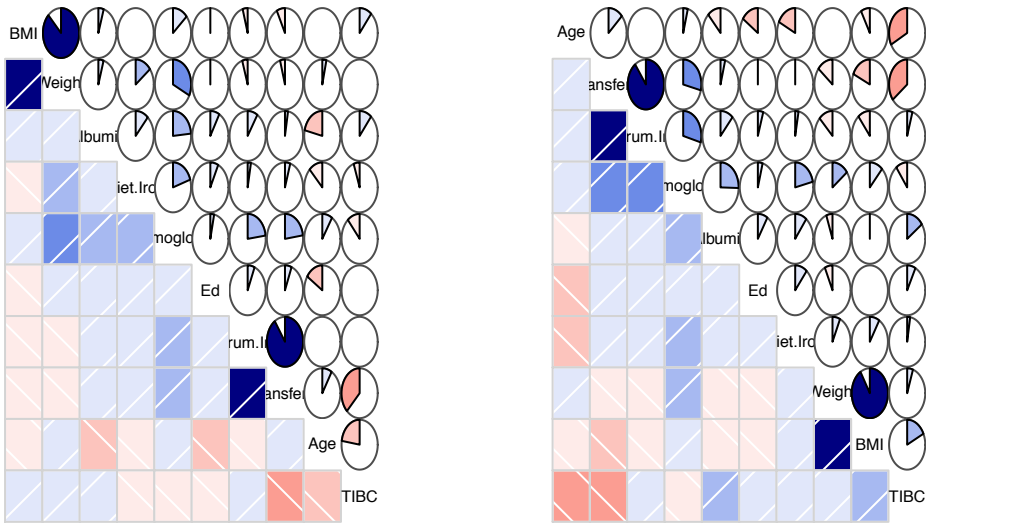
elogram for past or Current Smo Correlogram for non–Smokers



For smokers versus non smokers it seems that for Smokers Hemoglobin and Weight have a higher correlation than that of the general population and also transferin and TIBC for smokers has a higher correlation. For non smokers it seems hemoglobins relationship with all variables is stronger than that of the general populations as well as the Transferin and TIBC levels having a stronger correlation.

```
## [1] "numeric"
```

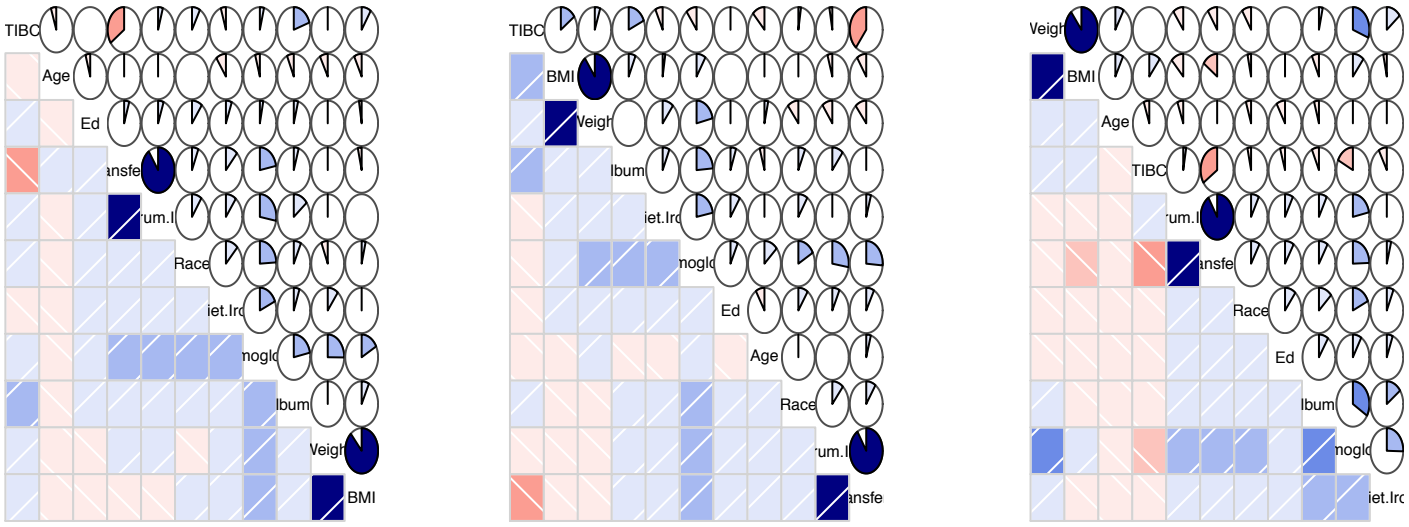
Correlogram for Caucasians orrelogram for Non–white Peopl



Comparing non white people to white people we see that Caucasians tend to have a stronger relationship between hemoglobin and Weight than that of non whites. Also nonwhite people have a slightly strong relationship between hemoglobin and Transferin than white people.

Next we will look at the corrleograms for different age groups.

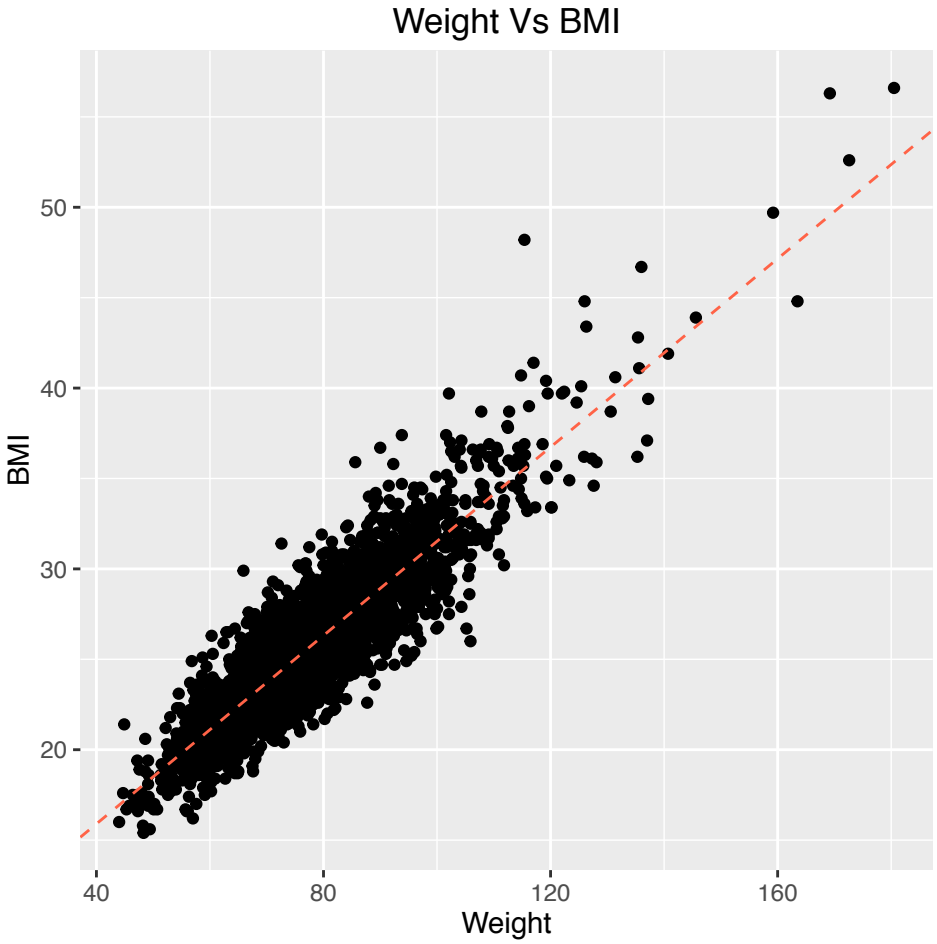
Correlogram for Elderly **orrleogram for Middle Aged Peop** **Young Adult Corrleogran**



From these Corrleograms we see that middle aged people tend to have a strong transferin and TIBC relationship than the other age groups. Other than that most of the factors are relatively similar across the groups.

6

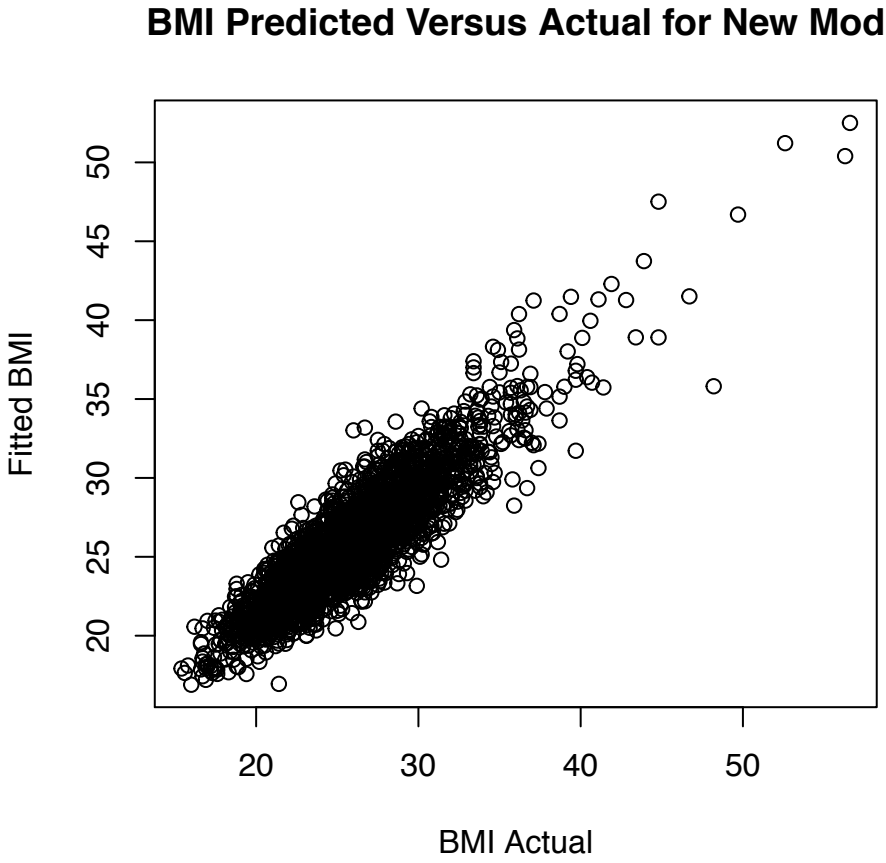
For this part of the project we will work on seeing what affects BMI. We already previously saw that only Males have non-na BMI levels. We also know from the Corrleograms that BMI has a strong corrleation with Weight.They have an r-squared of .89.



As seen above the linear model appears to fit better, next we will run a stepwise AIC model selection process to see if there is a better predictor for BMI.

```
## [1] 11789.63
```

```
## [1] 11931.99
```



This new model is slightly better than the model with just Weight. It's AIC value is about 200 lower than the old model and R-squared for the new model is atad larger than the model with just weight at .895. The new model of Transferin, Weight, TIBC, Race, Education, Serum Iron Levels, and Diet Iron levels is a lot more costly than the other model, and due to the surplus of predictors it would be valuable to look at the BIC levels and try other types of predicting such as k-nearest neighbor and Lasso to see if it is okay to drop the variables.

Conclusion

This report mainly evidences that besides smoking and age, no other factor has a major affect on Cancer Rates. While age and smoking are definitely correlated with Cancer, they do not have a significant enough of an influence to correctly predict whether a person will actually obtain cancer. Many other factors are at play when predicting cancer and to further this report would require many more variables.