STA 141

Homework 2

Report

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 Introduction

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 of this report is to find the relation between each variable and Cancer Incidence. Other interesting findings will also be stated. For detailed analysis, please refer to “HW_2 Codes and Analysis File”.

Noted that (Fig. 1) there are approximately 60% of missing values in BMI category. There also exist some missing values in "Transfererin", "Serum Iron", "TIBC", "Hemoglobin", "Albumin", and "Diet Iron". It is necessary to be cautious when dealing these missing values.

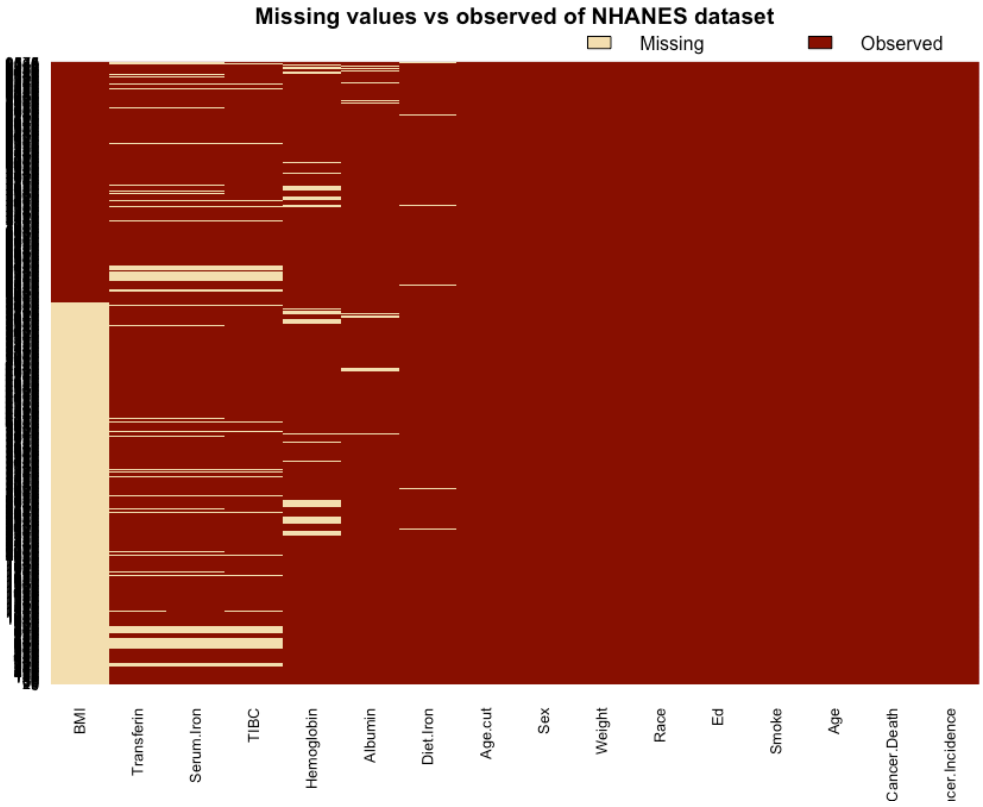

Figure 1 Missing values vs. Observed of NHANES dataset

2 Factors Affect Hemoglobin Level

2.1 Smoking Status

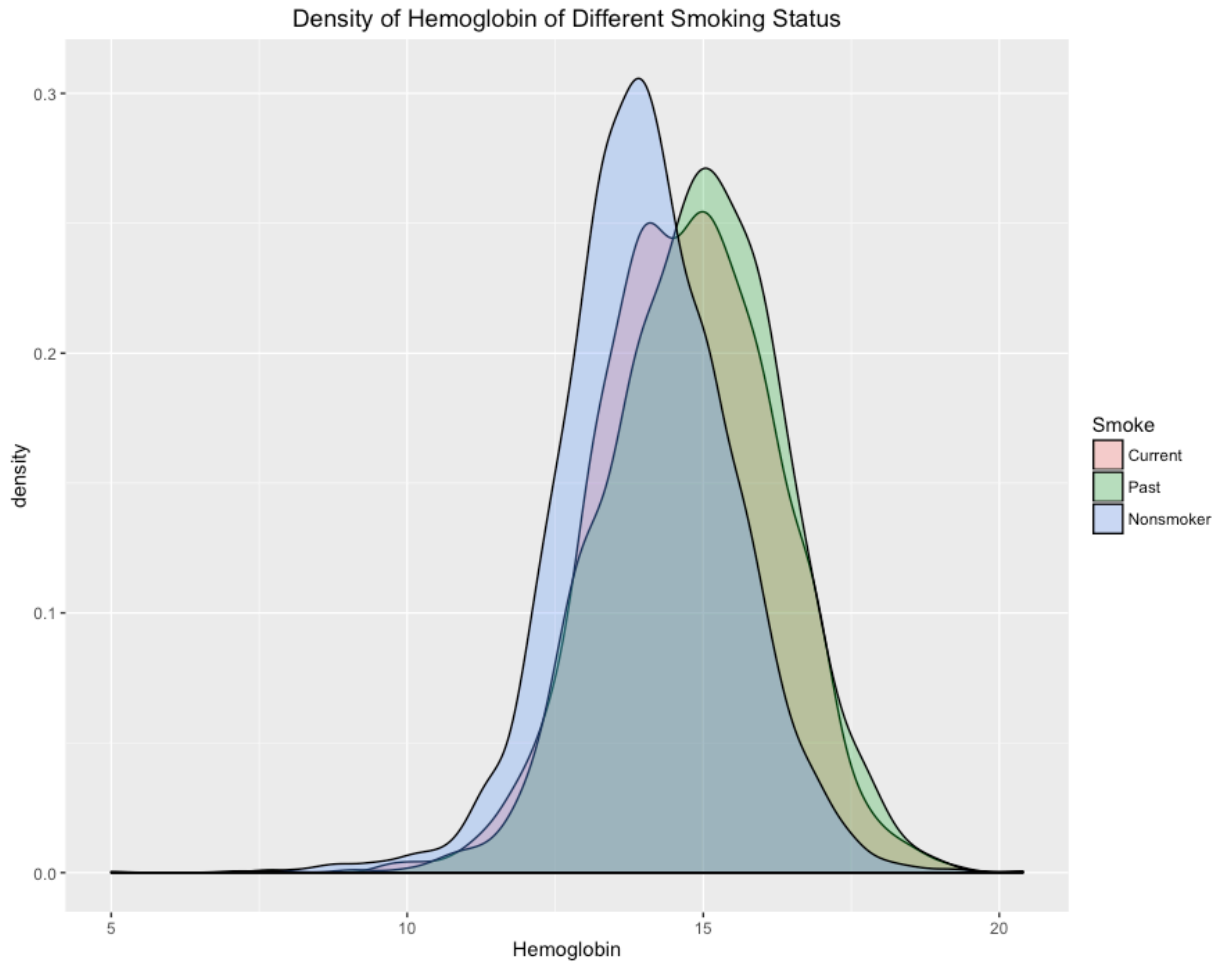

Figure 2.1-1 Density Plot of Hemoglobin Level of People with Different Smoking Status

Fig. 2.1-1 shows that smoking leads to a higher hemoglobin level. This conclusion is supported by Ref. 1. This trend holds when separating by other factors except gender, see Fig. 2.1-2.

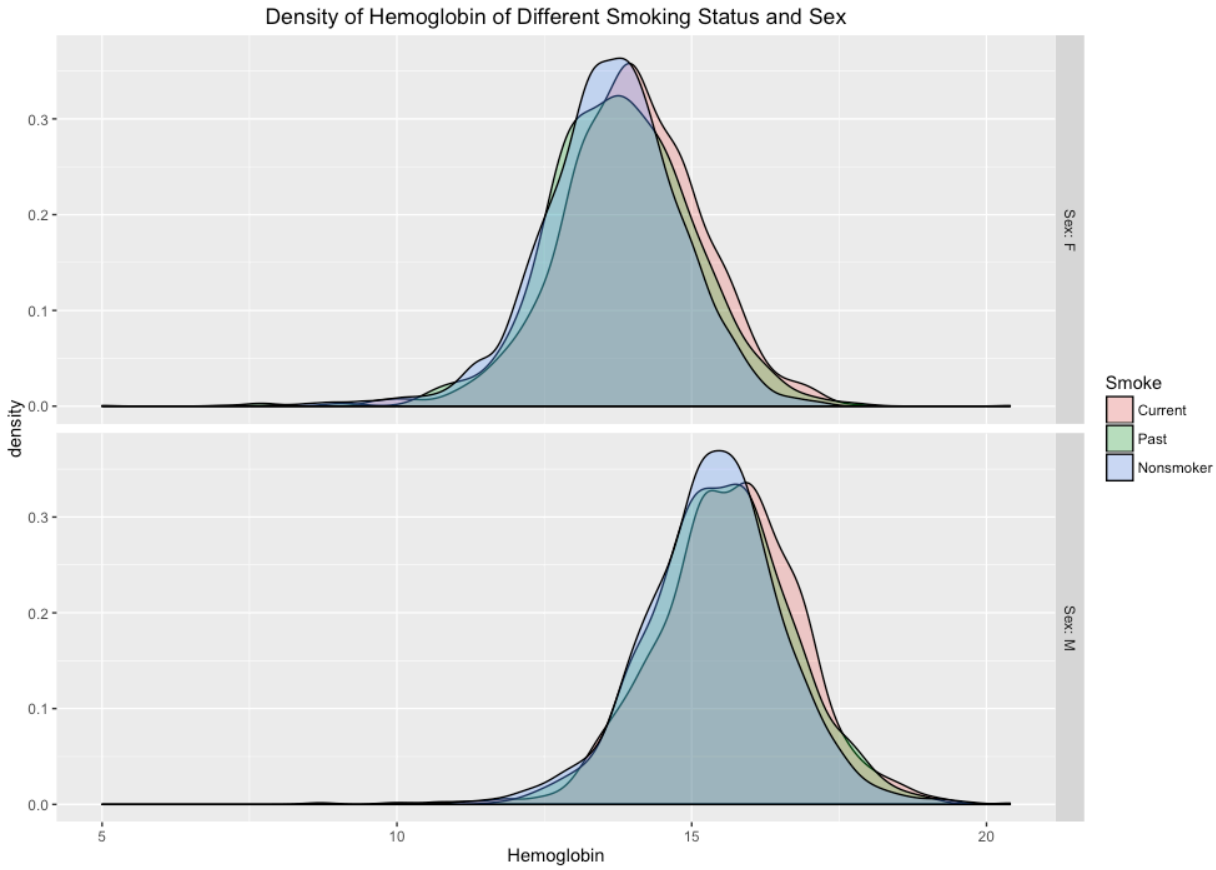

Figure 2.1-2 Density Plot of Hemoglobin Level of People with Different Smoking Status and Gender

2.2 Race

Fig. 2.2 shows that race has a significant influence on hemoglobin level. Caucasian in general have higher hemoglobin level than non-Caucasian. Same relation holds across other categories.

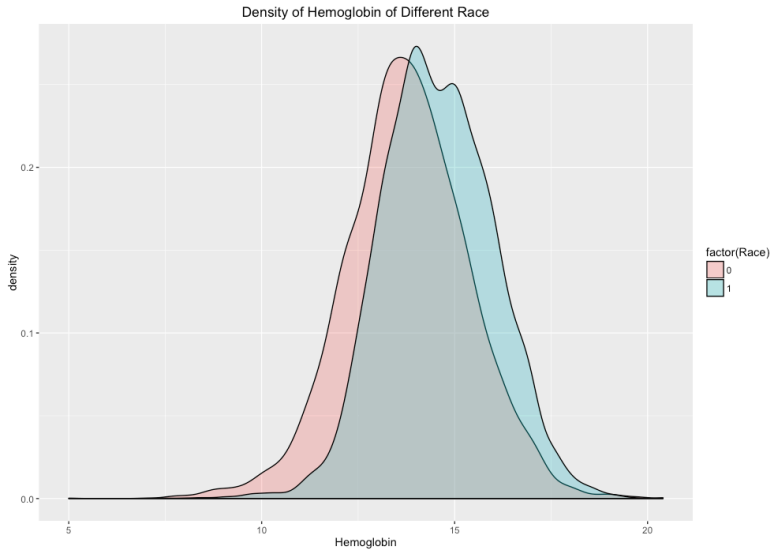

Figure 2.2 Density Plot of Hemoglobin Level of People with Different Race

2.3 Sex

Sex has a significant influence on hemoglobin level. Men in general have higher hemoglobin level than women. Same relation holds under other categories. (See Fig. 2.3) (Also supported by natural cluster in Fig. 5.4.)

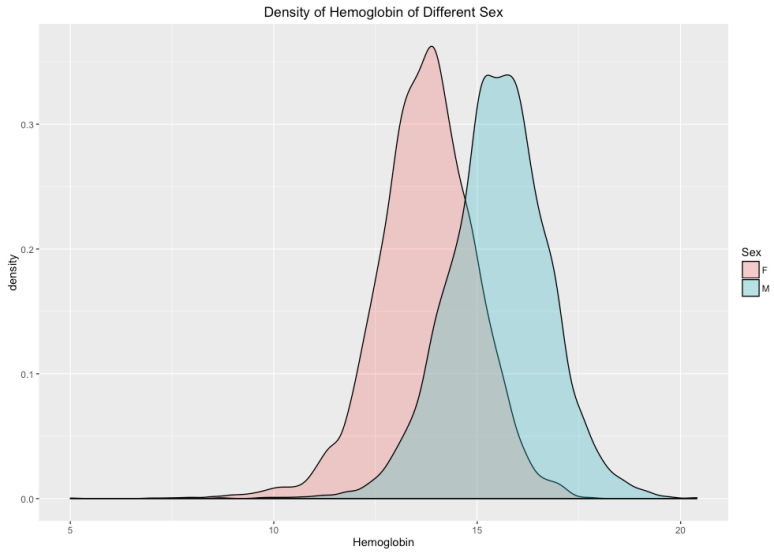

Figure 2.3 Density Plot of Hemoglobin Level of People with Different Gender

3 Factors Affect Cancer Incidence

3.1 Statistical Finding

It can be calculated 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.

3.2 Factors

3.2.1 Age

Below shows that older people are more likely to get cancer.

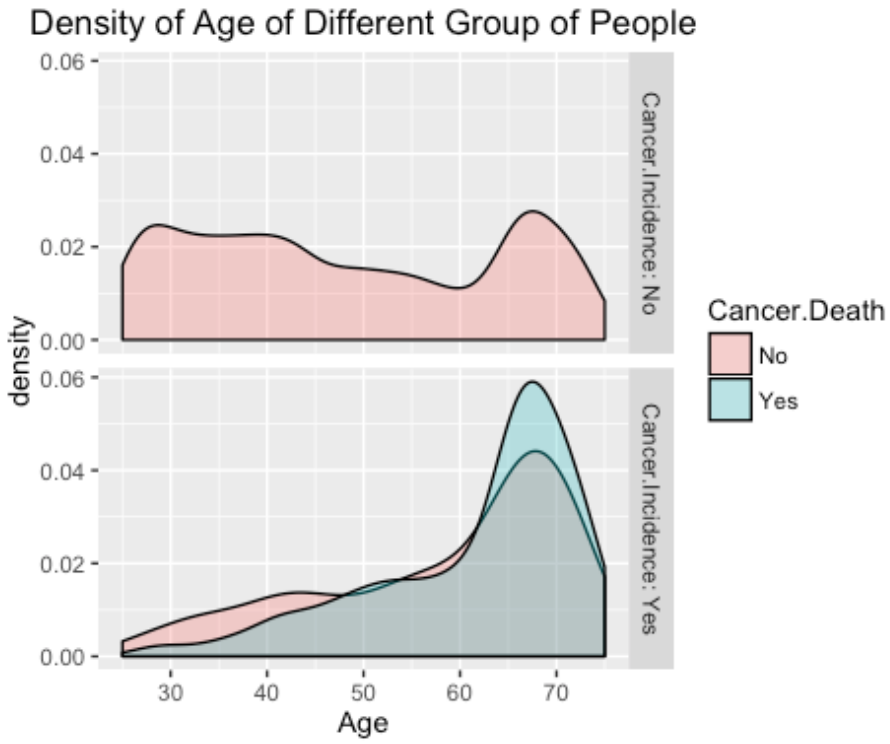

Figure 3.2.1-1 Density Plot of Age of Different Group of People based on Cancer Status

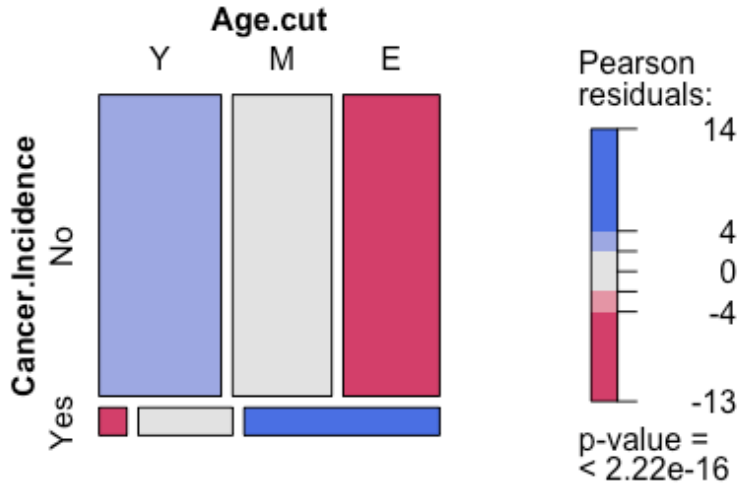

Figure 3.2.1-2 Mosaic Plot of Age and Cancer Incidence

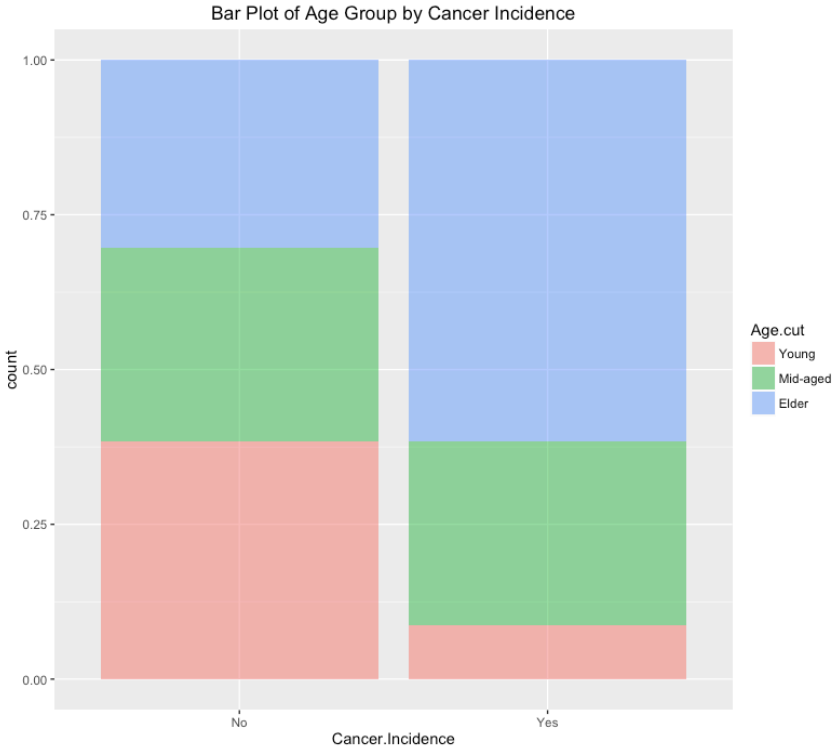

Figure 3.2.1-3 Bar Plot of Cancer Incidence of Different Age Group

Over 60% of people who get cancer falls into the elder category. People who are over 40 years old takes up 90% of the people who get cancer.

3.2.2 Age + Smoke

For mid-aged and elder people, the percentage of cancer incidence is the highest in "Current Smoker" category, and lowest in "Non-smoker" group. So for mid-aged and older people, smoking can cause a higher cancer incidence rate.

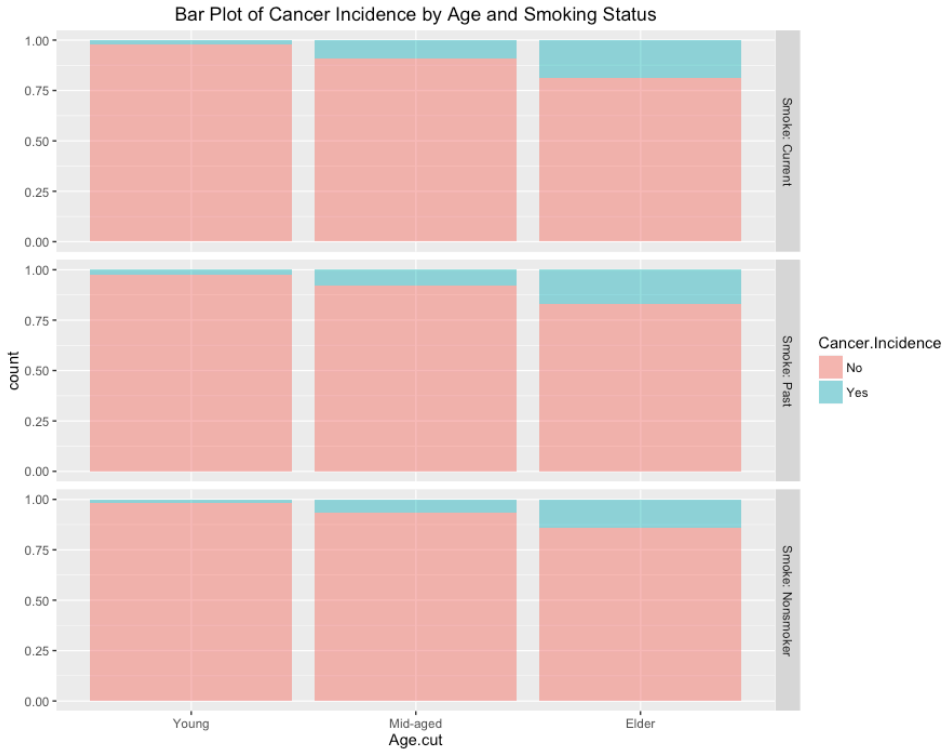

Figure 3.2.2 Bar Plot of Age Groups Separated by Cancer Status and Cancer Incidence

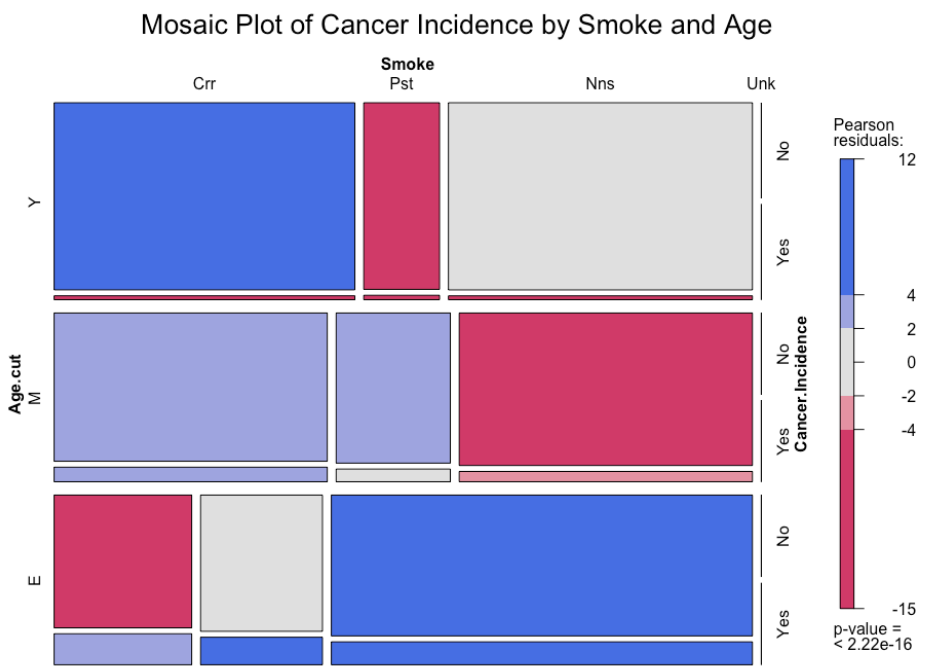

Figure 3.2.2-2 Mosaic Plot of Cancer Incidence, Smoking Status and Age

3.2.3 Age + Sex

It can be observed from Fig. 3.2.3-1 and -2 that at young age, female have higher chance to get cancer. While at old age, male have higher chance to get cancer.

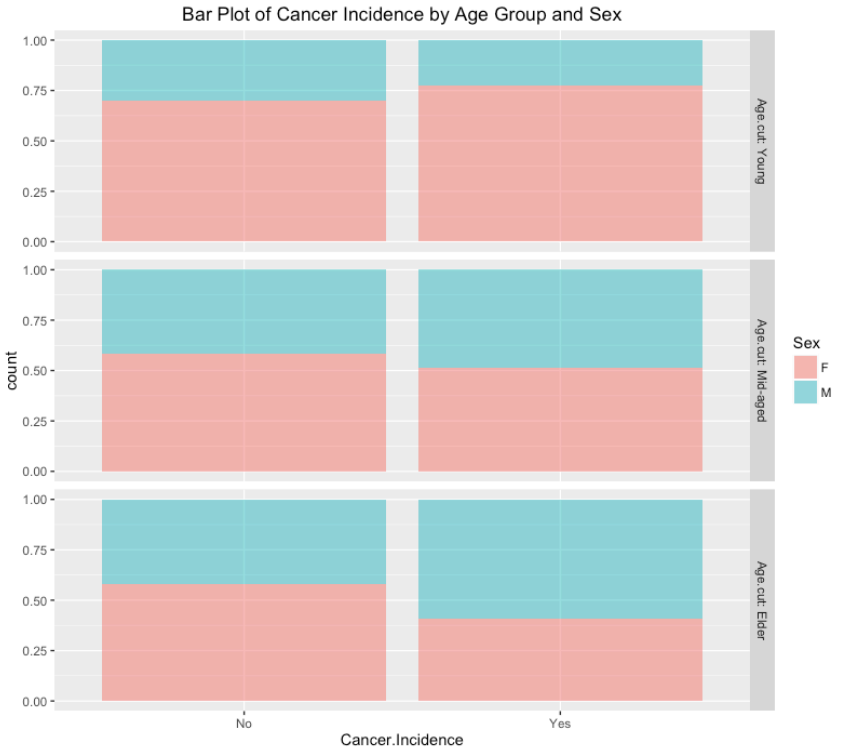

Figure 3.2.3-1 Bar Plot of Cancer Incidence Separated by Age Group and Sex

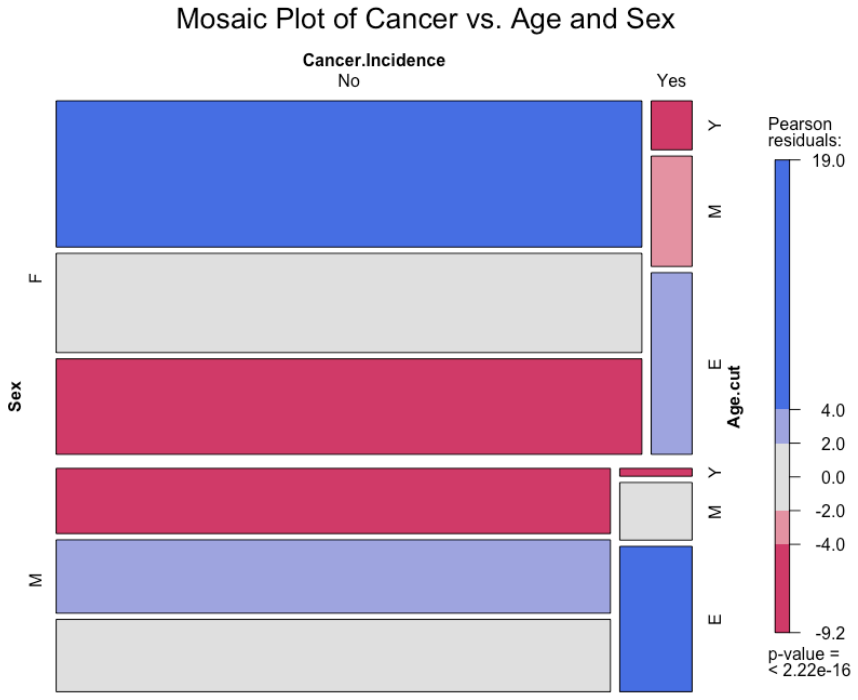

Figure 3.2.3-2 Mosaic Plot of Cancer Incidence, Age Group and Sex

3.2.4 Sex + Smoke

Below shows that male who are currently smoking have higher cancer incidence.

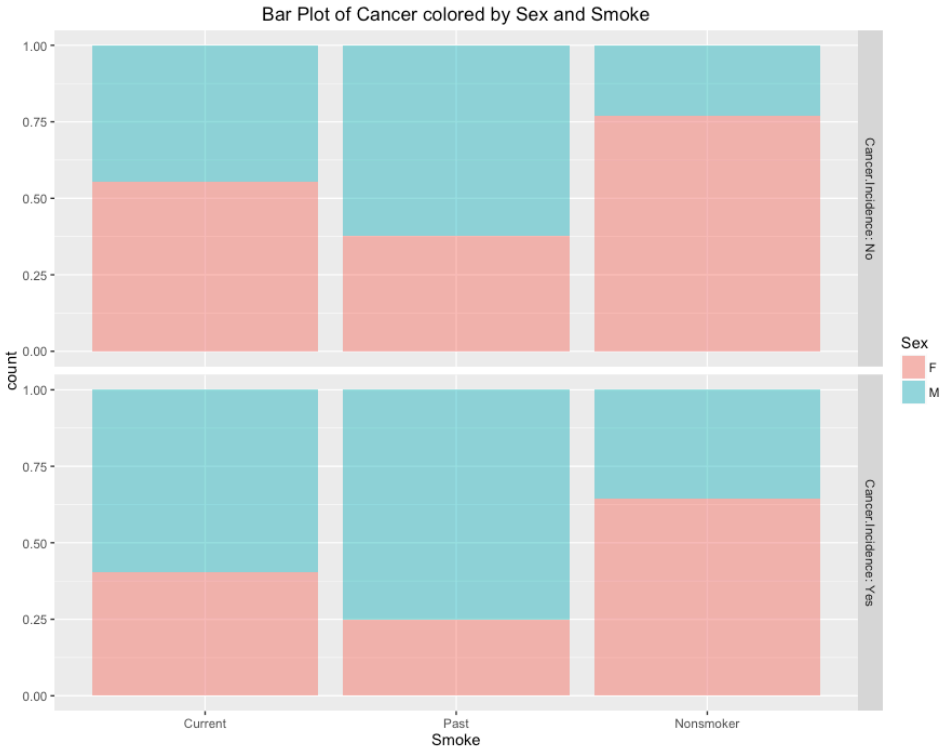

Figure 3.2.4-1 Bar Plot of Smoke Status Separated by Cancer Incidence and Sex

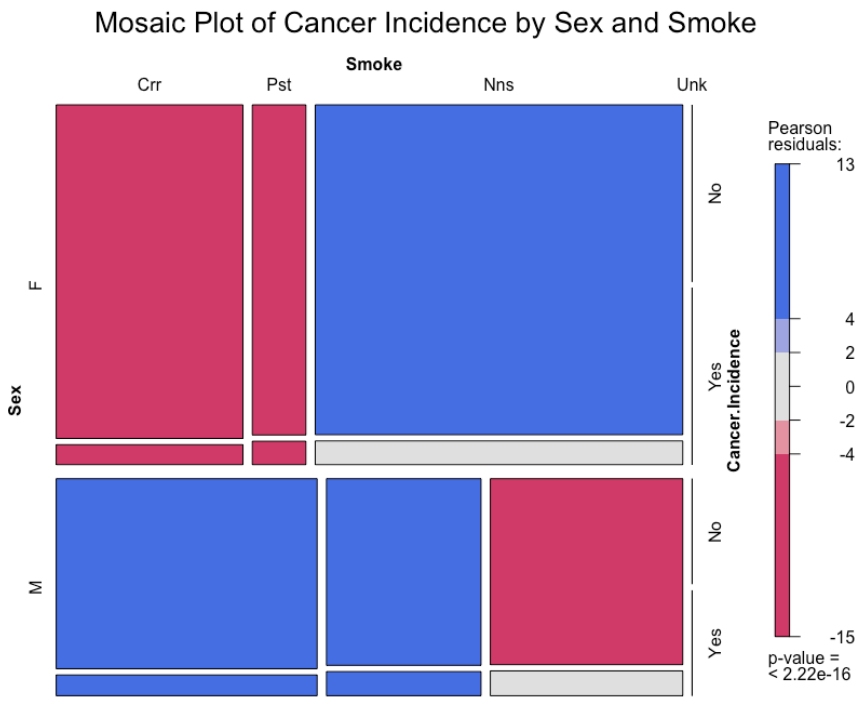

Figure 3.2.4-2 Mosaic Plot of Smoke Status, Cancer Incidence and Sex

4 Prediction of BMI

4.1 BMI – Weight Model

BMI value can be linear fitted on Weight. The diagnosis plot shows that there is slightly violation on the normality and equal variance assumption.

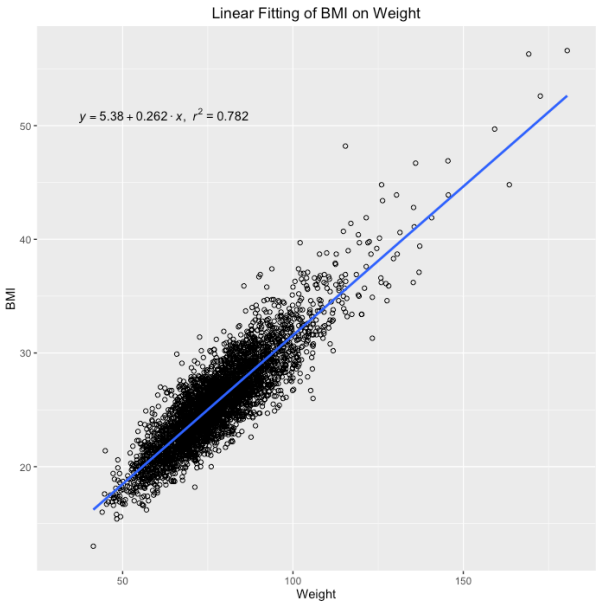

Figure 4.1-1 Linear Fitting of BMI on Weight

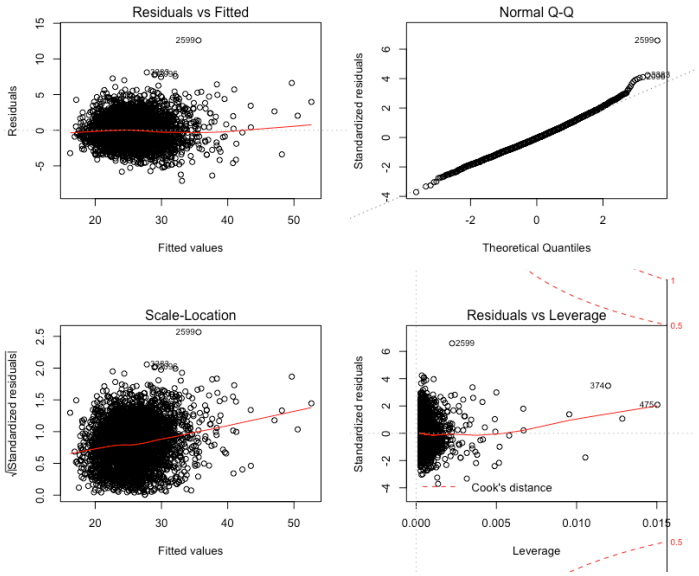

Figure 4.1-2 Diagnosis Plot of Linear Model of BMI on Weight

4.2 Factors affect Linear Model

Education and Race are two most significant factors that influence the linear model of BMI~Weight due to the observable change in slope. The density of BMI change with Education and Race too.

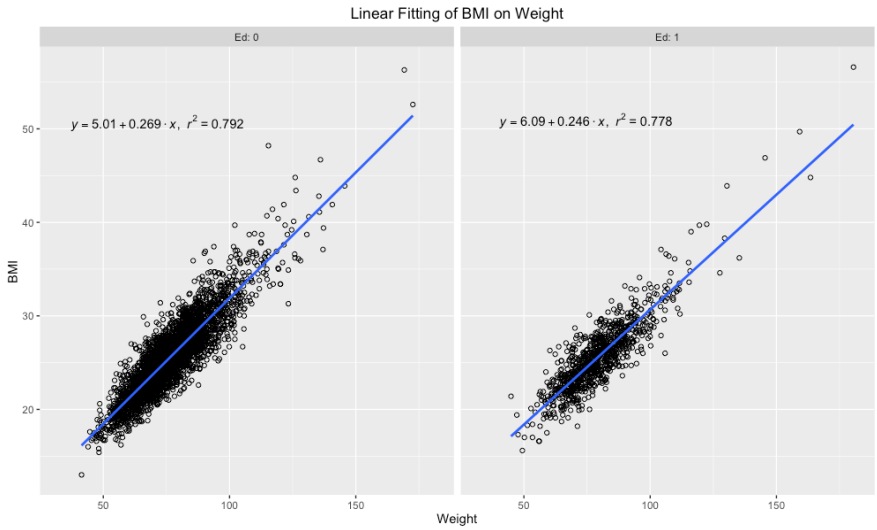

Figure 4.2-1 Linear Fitting of BMI on Weight Faceted by Education

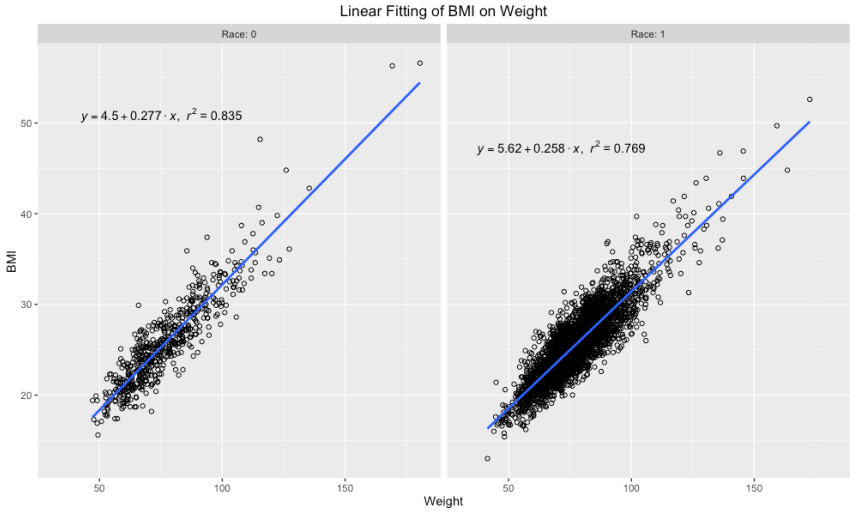

Figure 4.2-2 Linear Fitting of BMI on Weight Faceted by Race

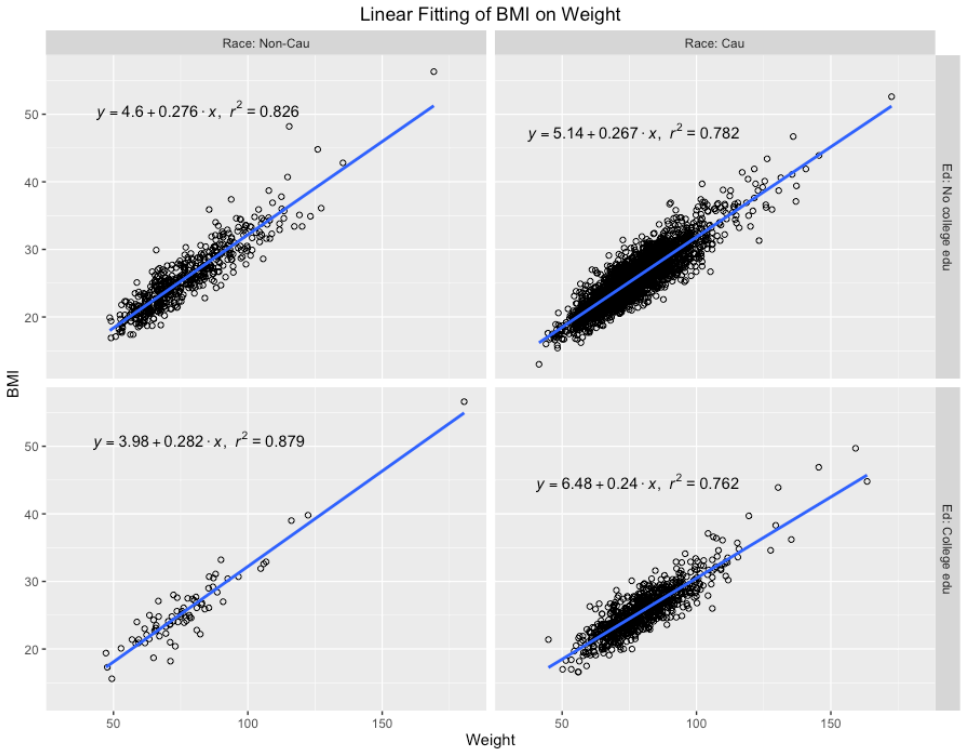

Figure 4.2-3 Linear Fitting of BMI on Weight Faceted by Race and Education

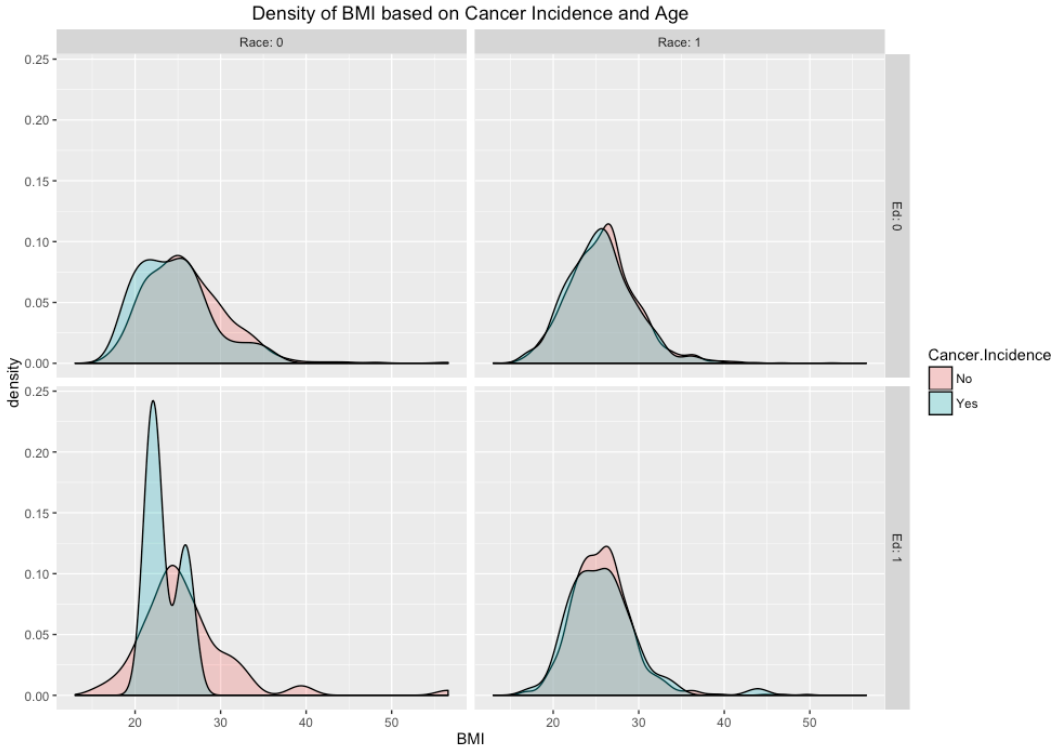

Figure 4.2-4 Density Plot of BMI Faceted by Race and Education colored by Cancer Incidence

5 Other Findings

5.1 Correlation

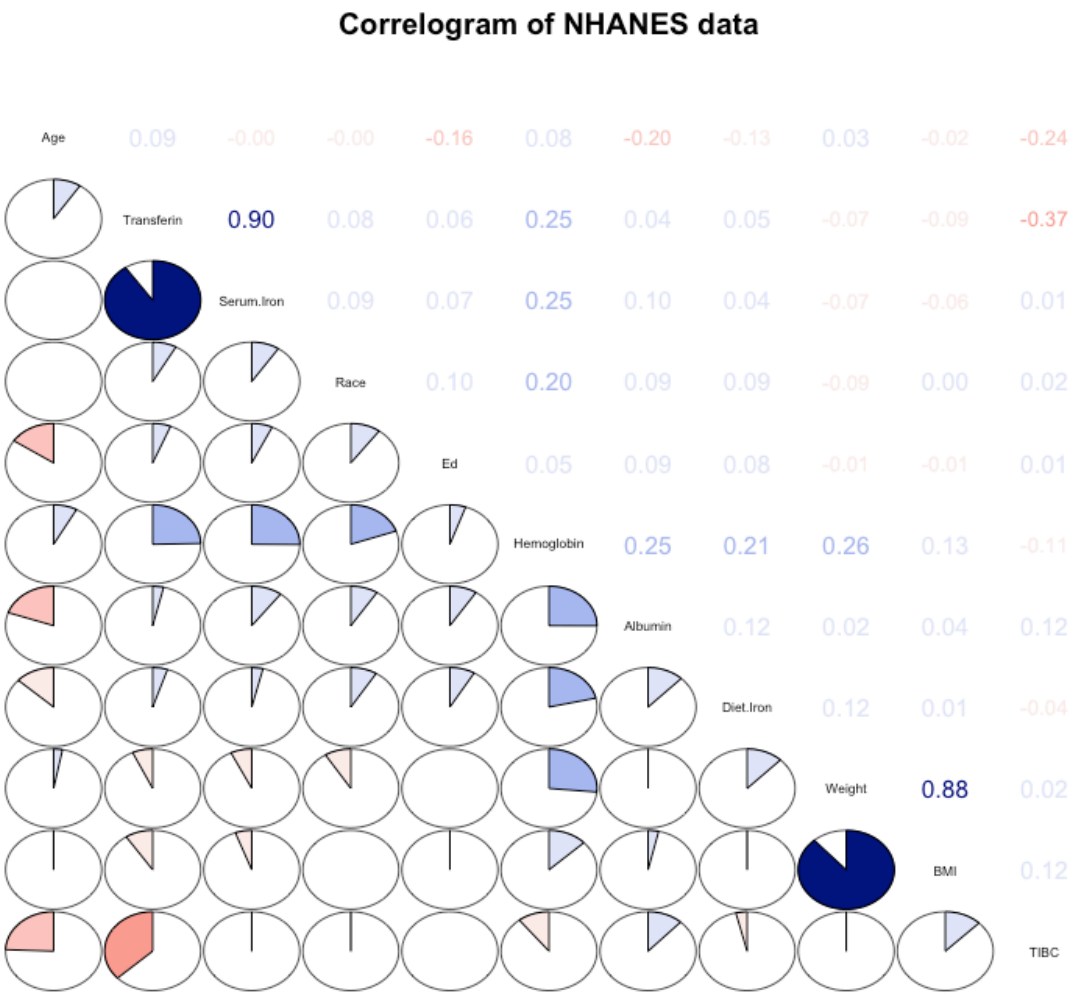

Figure 5.1 Correlogram of NHANES Data Set

Strong correlation can be seen on BMI ~ Weight, Transferin ~ Serum.Iron and Transferin ~ TIBC pairs. Since Trasnferin is defined by Serum.Iron/TIBC, the observation of high covariance among Serum.Iron, TIBC and Transferin is not surprising.

5.2 Health Index vs. Age

1. Weight of young people is lower than the others.
2. People who is younger tend to have slightly higher Albumin.

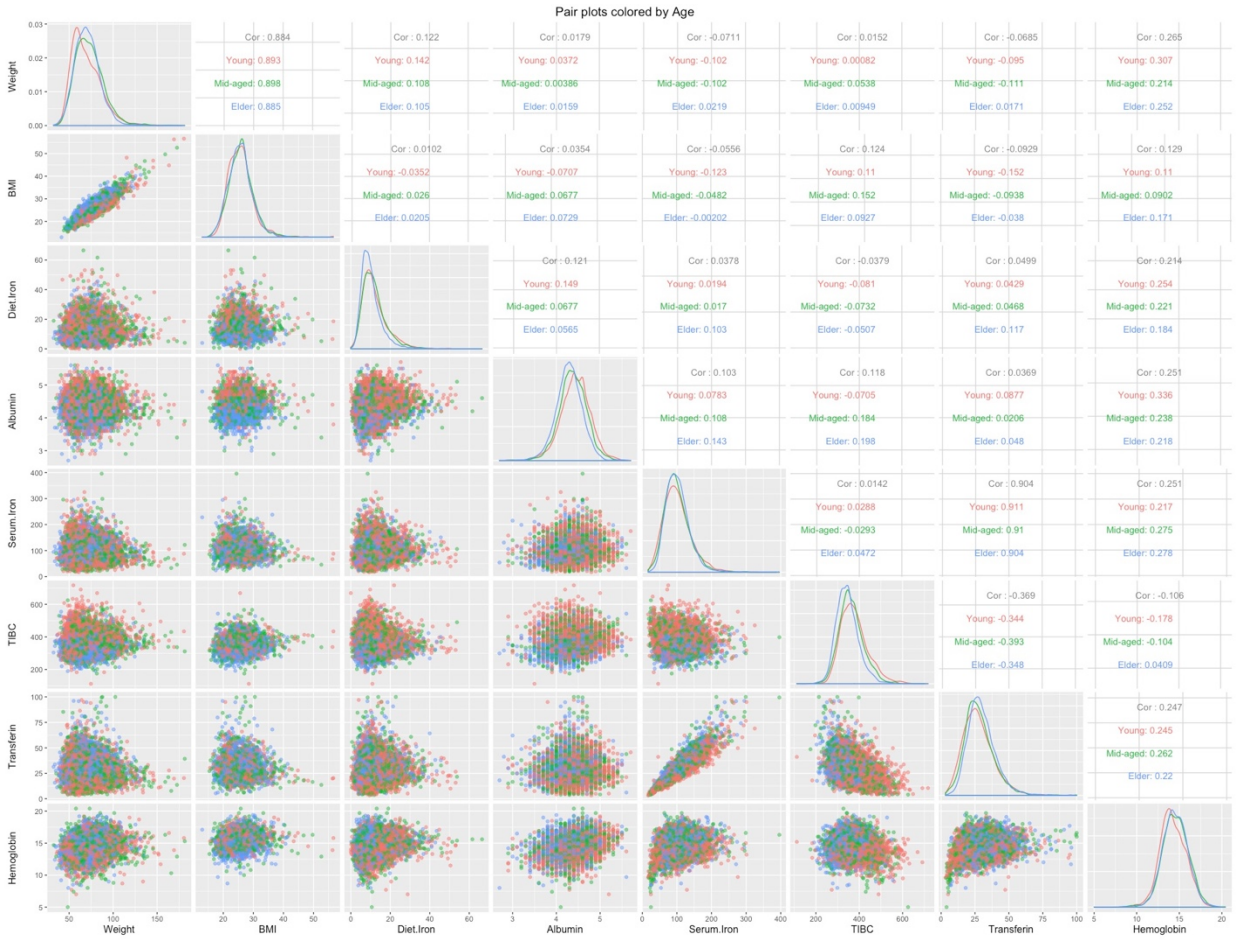

Figure 5.2-1 Pair Plots Colored by Age

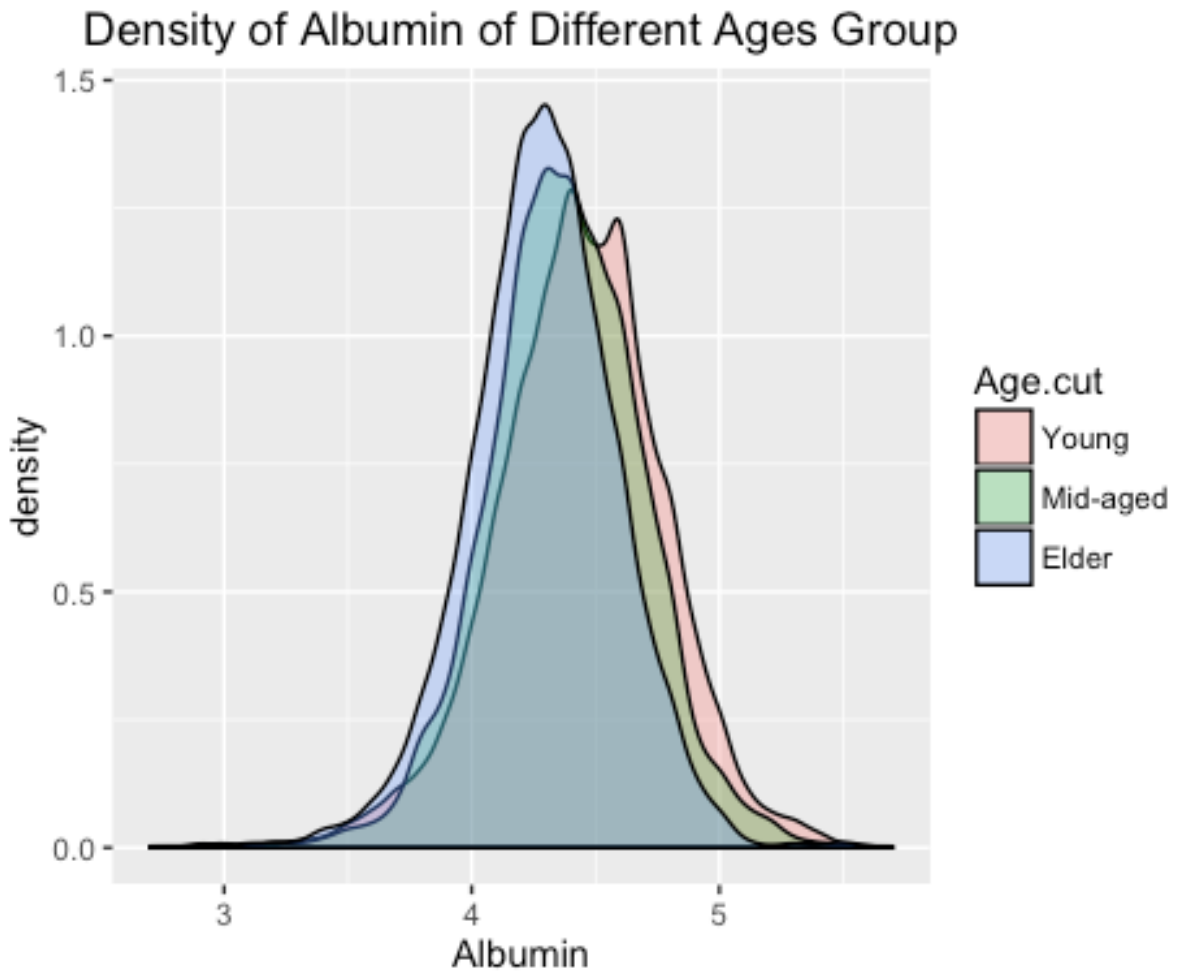

Figure 5.2-2 Density Plot of Albumin Colored by Age

5.3 Health Index vs. Smoke

People who are currently smoking or are past smoker tend to have higher diet iron. Non-smoker have lower diet iron.

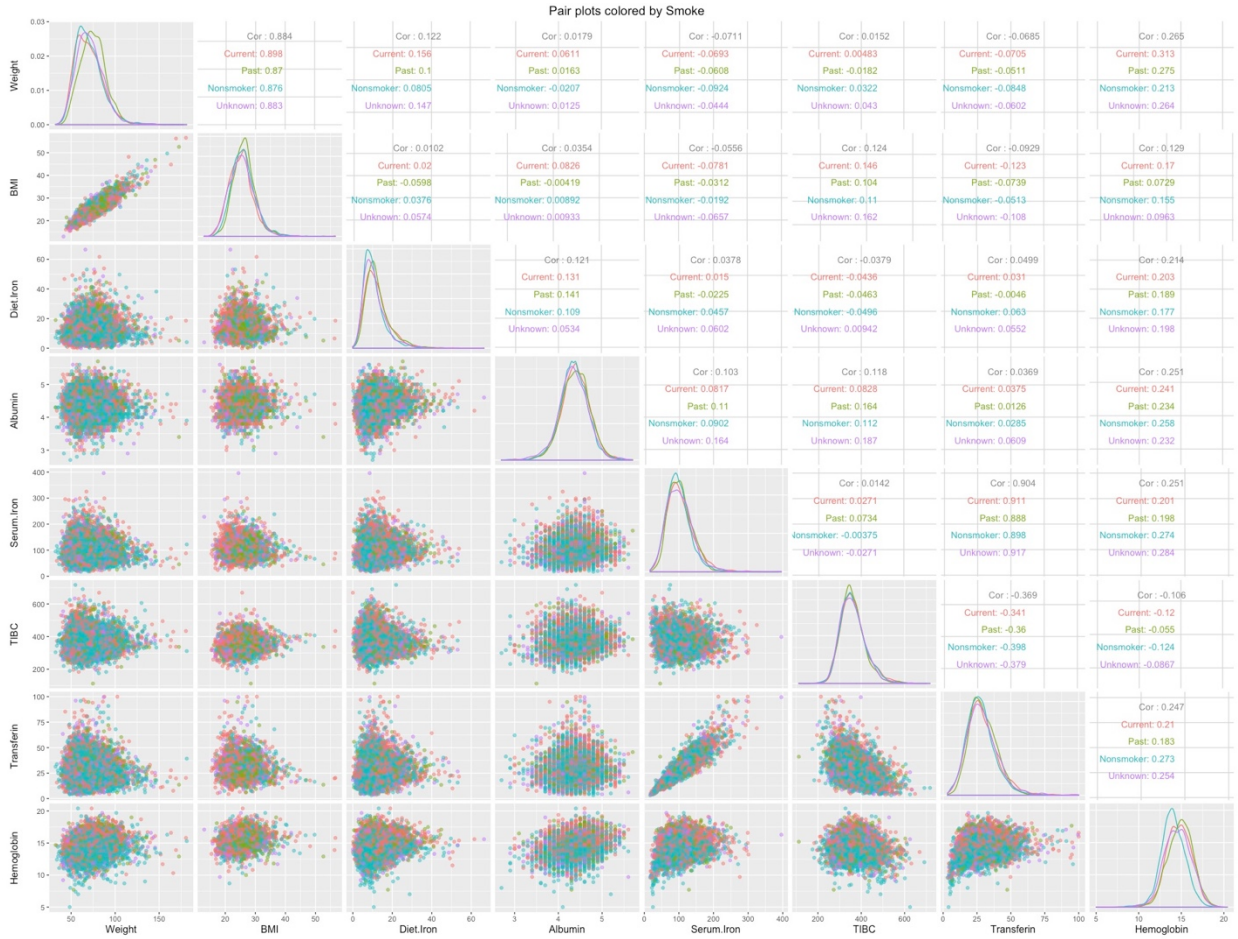

Figure 5.3 Pair Plots Colored by Smoke

5.4 Health Index vs. Race

Caucasian tend to have higher level in both Serum Iron and Transferin.

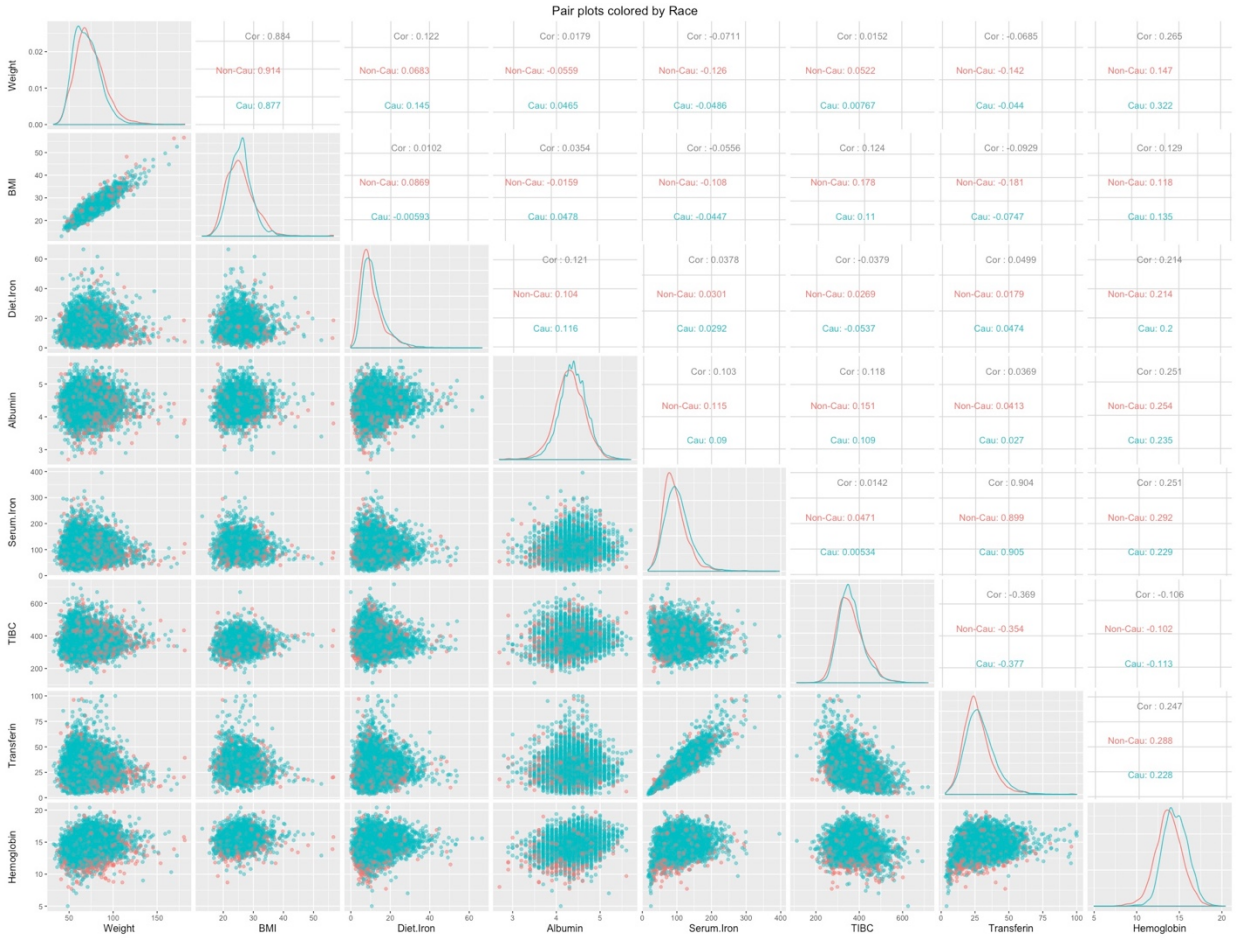

Figure 5.4 Pair Plots Colored by Race

5.5 Health Index vs. Sex

1. Men in general have higher weight than female. Clear cluster can be observed.

2. The dietary iron value is different for two genders.

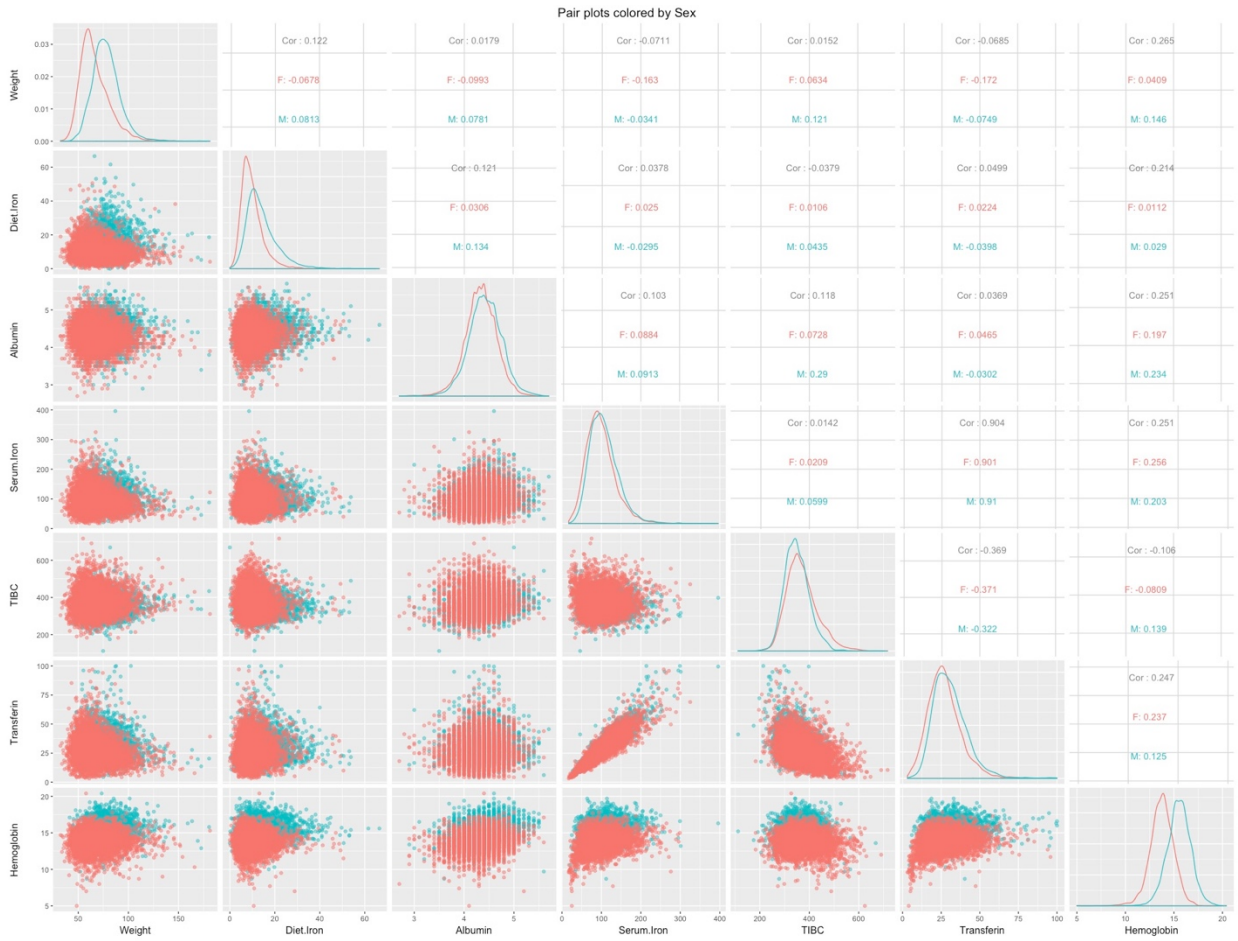

Figure 5.4 Pair Plots Colored by Sex

6 Error Analysis

6.1 Distribution of Categorical Variable

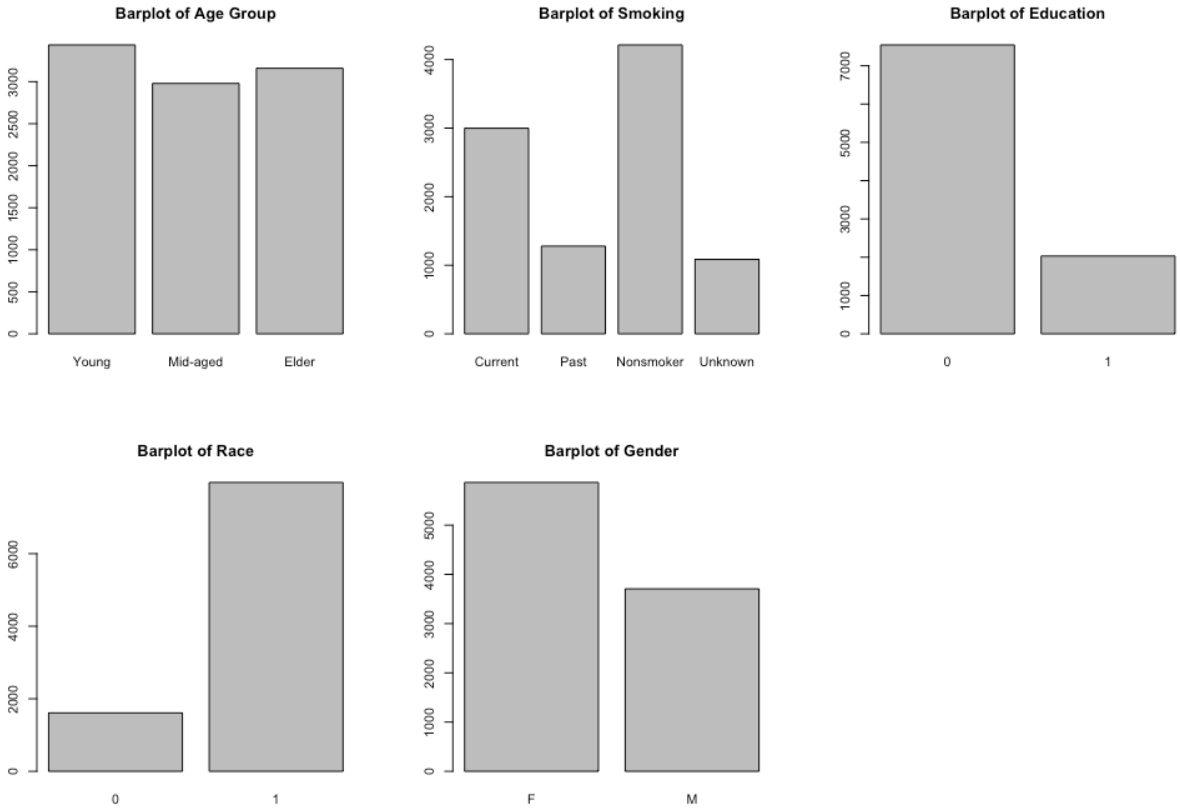

Figure 6.1 Barplot of Each Categorical Variables

Note that, the count of different category in these factor variables are not evenly distributed. This might introduce possible source of error to the analysis.

6.2 Missing values

Reference:

1. Light, A., et al. "Carboxyhemoglobin levels in smokers vs. non-smokers in a smoking environment." Respir Care 52.11 (2007): 1576.