

Statistical Analysis on NHANES

STA 141A HW#2

Hiu Man(Stephanie) Lam ID: 913358632

HONOR CODE

The analysis derived by using the codes attached in appendix constitutes my own work. I have consulted the following resources regarding this assignment:

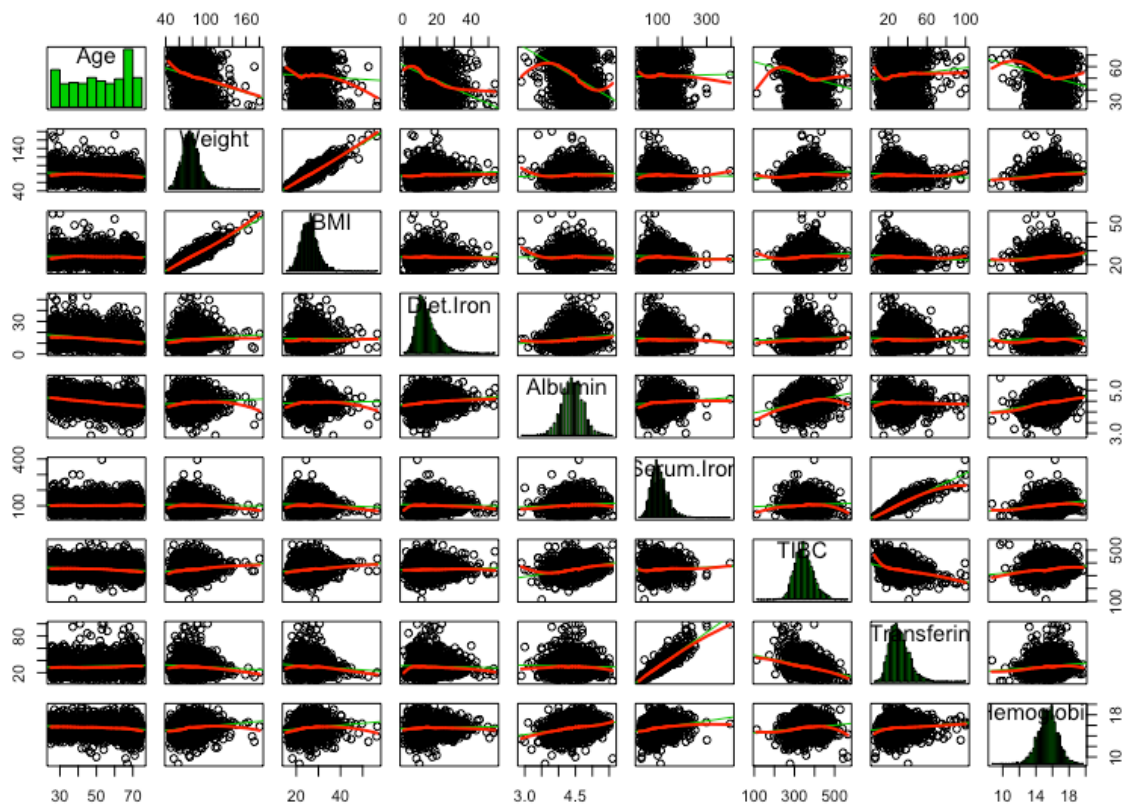
Samuel Kadin, Ryan Kuan, Janet Loyola, and Visualization of imputed values using the R-package VIM: <https://cran.r-project.org/web/packages/VIMGUI/vignettes/VIM-Imputation.pdf>

I. Introduction

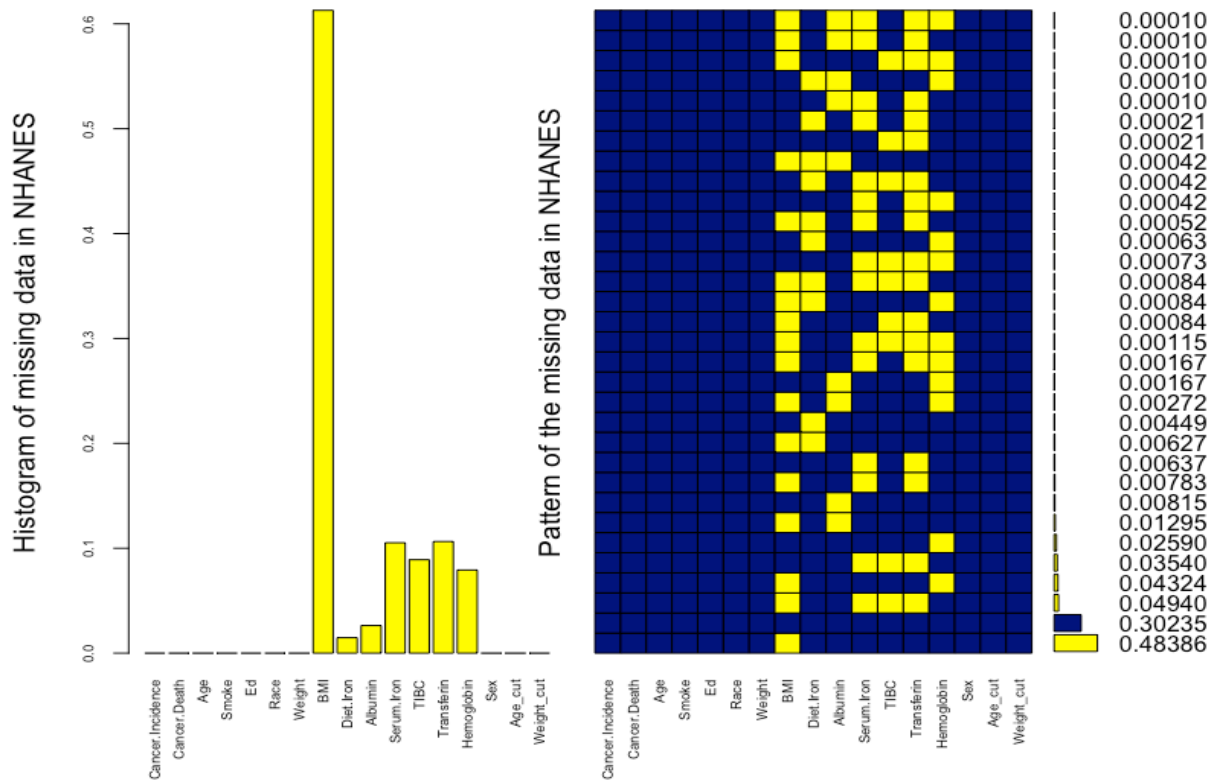
This report presents a data analysis using the dataset NHANES that aims to discover interesting patterns in the data that may help in building predictive models. For analysis, Ed will be defined as No college education and College education while Race as Non-Caucasian and Caucasian.

II. Data Analysis

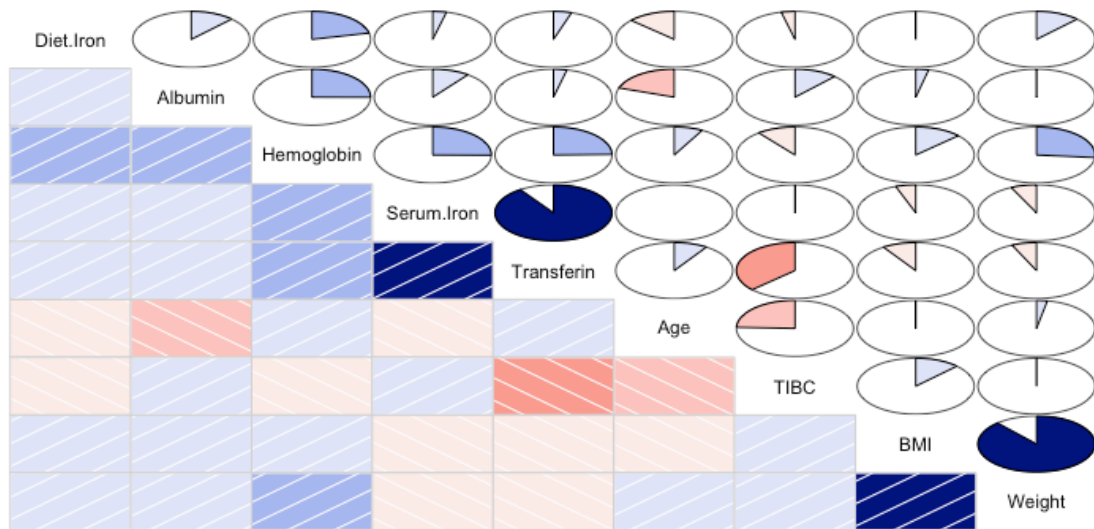
Basic scatter plot matrix of NHANES data



Based on scatterplot matrices, most of the continuous data have skewed distributions and outliers. The data contains many missing values that might have made the data skewed. By interpreting the pattern of the histogram of the missing data below, the missing BMI are not random since none of the BMI for female is recorded. And if the value of Serum.Iron is missing, there is a great chance that Transferin is missing too. The missing patterns of other variables appear to be random.



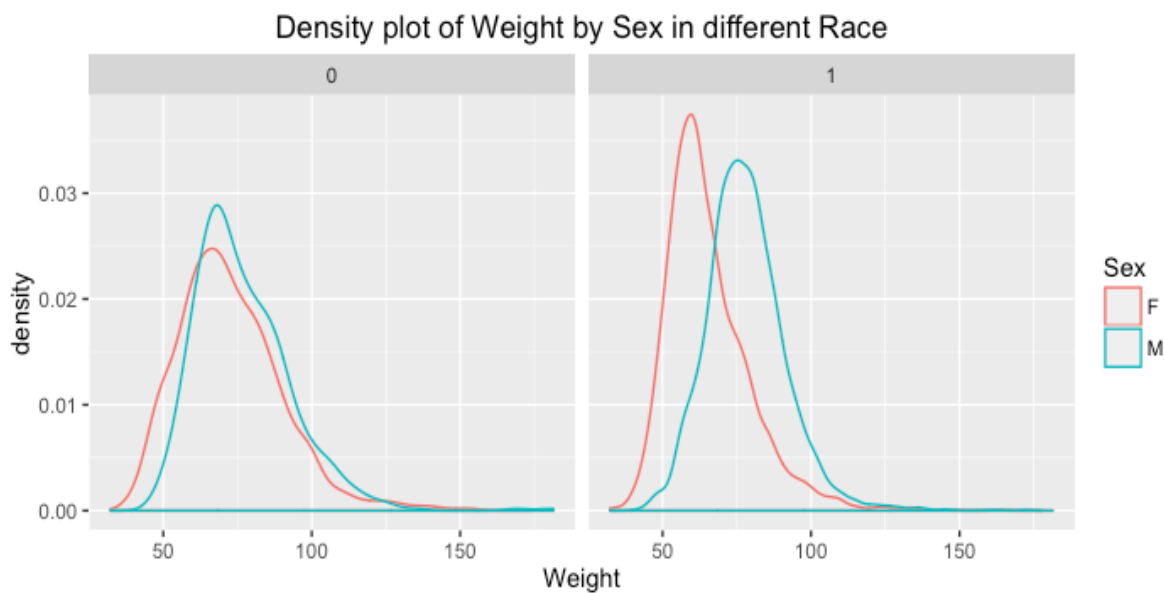
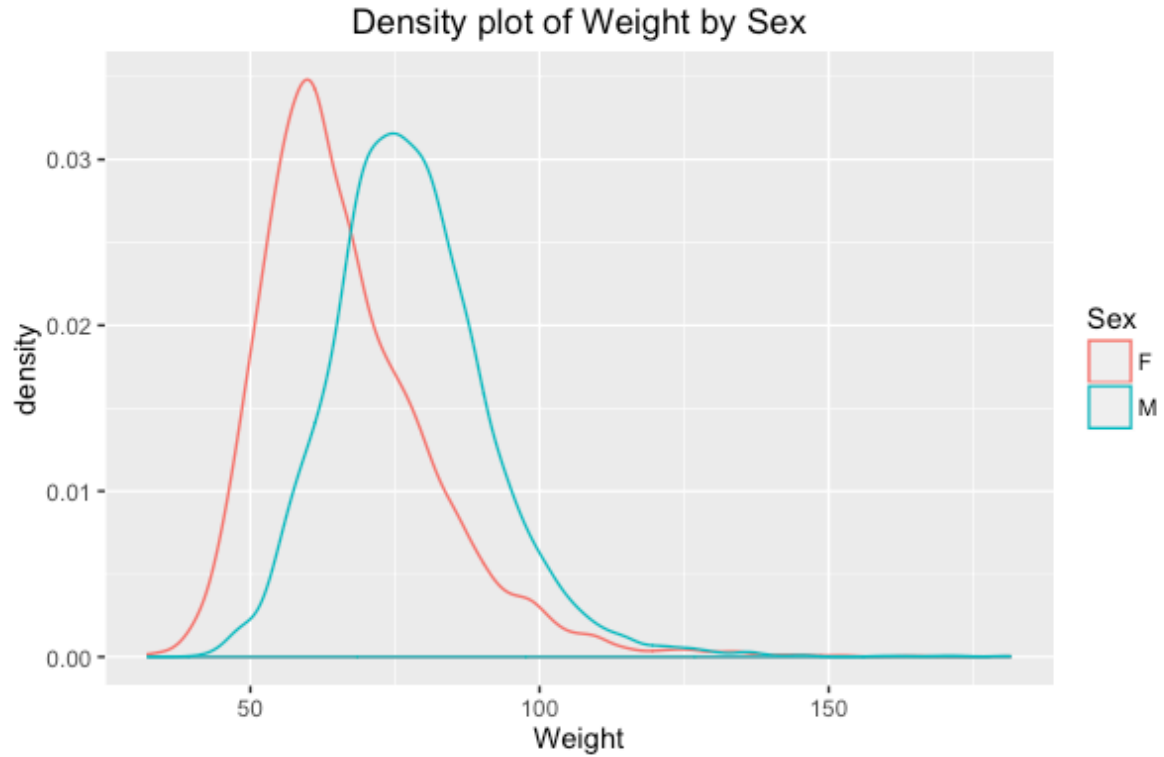
Correlogram of NHANES data



The correlogram shows that both Serum.Iron and Transferin, and BMI and Weight have strong positive correlations while Transferin and TIBC have a relatively strong negative relationship, which confirmed the stated relationship between Weight & BMI, and Transferin

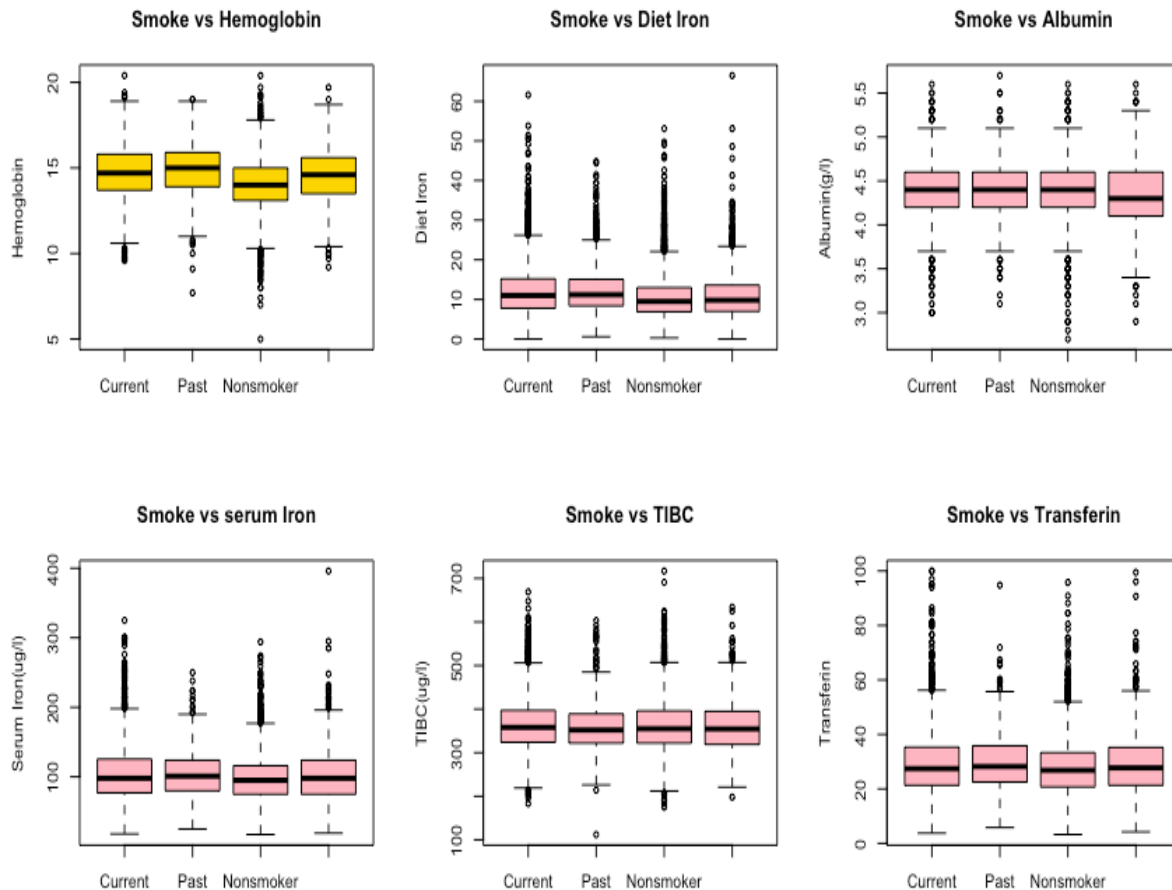
& Serum Iron and TIBC are correct. Hemoglobin also has moderate positive association with Diet Iron, Albumin, Transferin, and Weight.

III. Interpretation



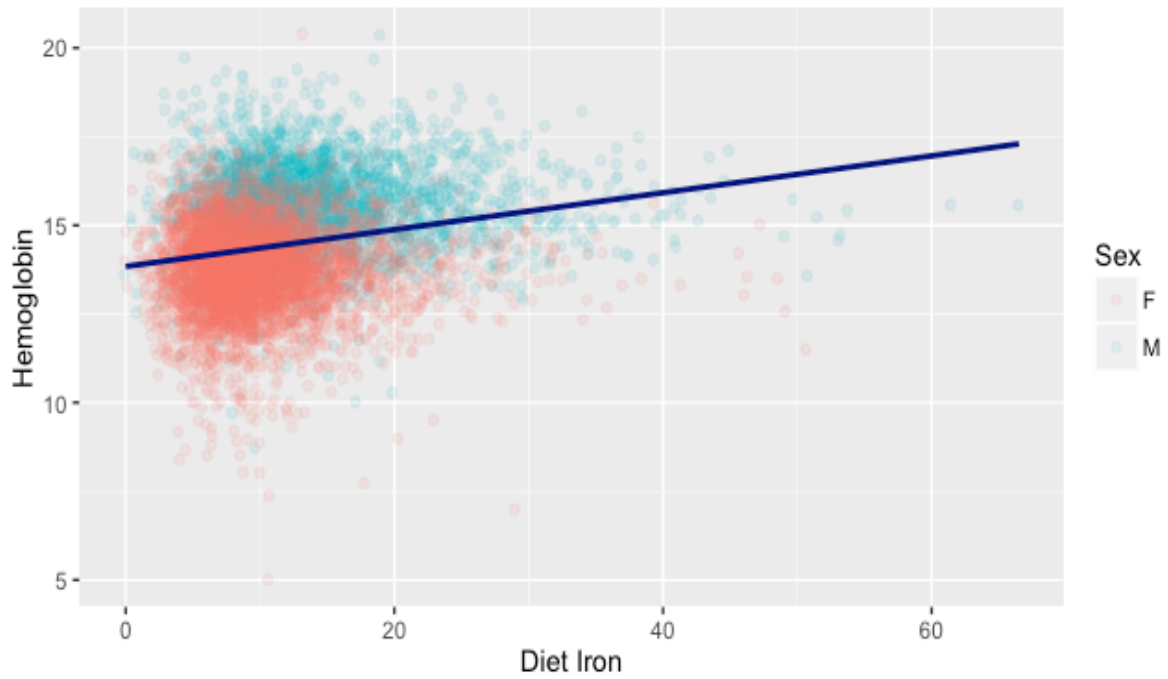
While exploring patterns, I found interesting that overall men have higher weight than women, but by race, non-Caucasian male and female have the same weight.

Since I am interested in whether smoking affects any chemical levels, I used boxplot to analyze and found that smoking would only affect Hemoglobin. Non-smoker has lower Hemoglobin, which means Hemoglobin would increase for people who smoke currently or in the past.

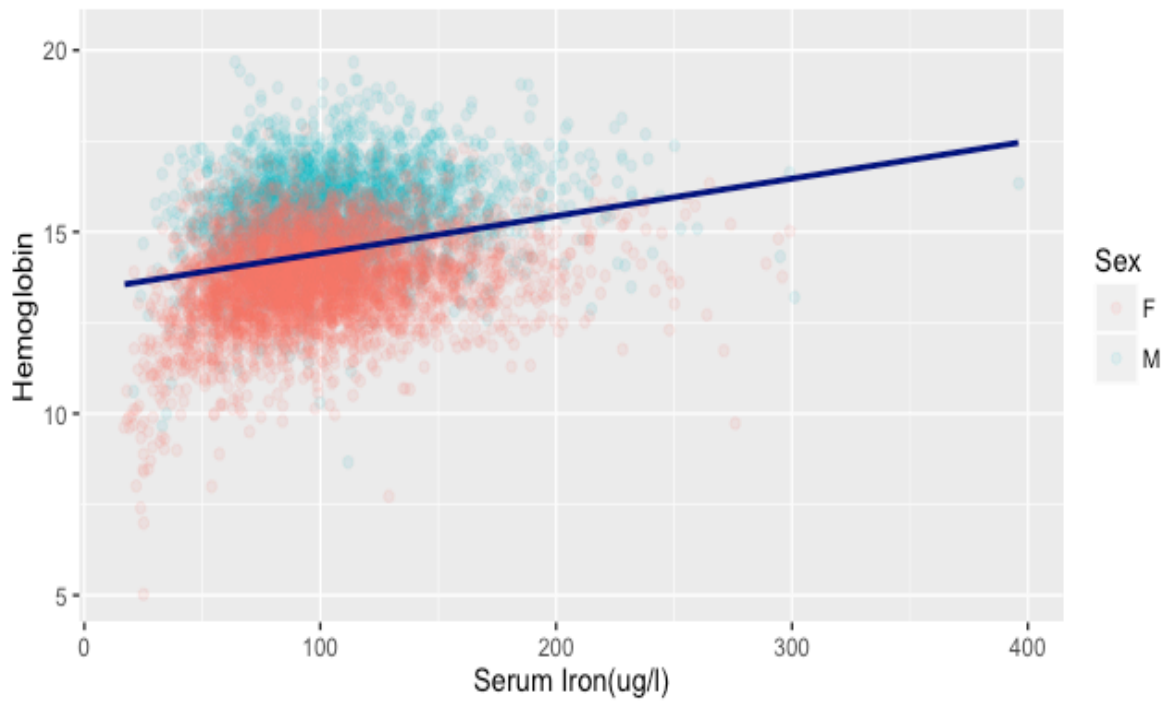


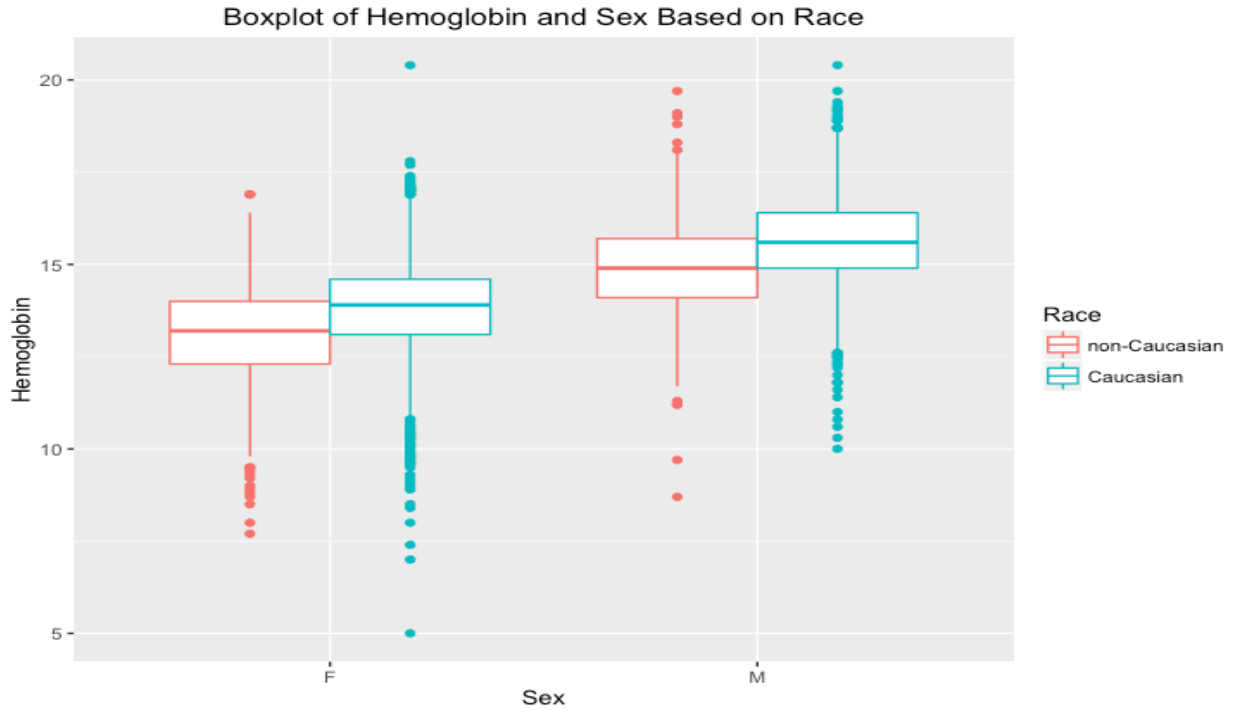
With further analysis, I observed that there are natural clusters for both Serum.Iron and Diet.Iron vs Hemoglobin when based on Sex on the plots below. The more Diet.Iron would lead to a slightly more Hemoglobin while Male and Caucasian would have higher level of Hemoglobin than female and non-Caucasian.

Diet Iron and Hemoglobin based on Sex

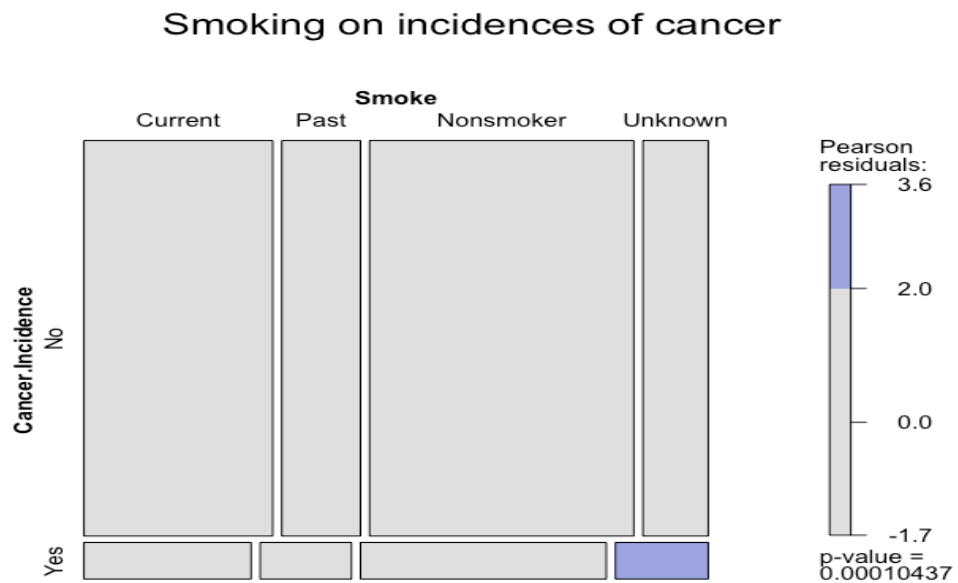


Serum Iron and Hemoglobin based on Sex

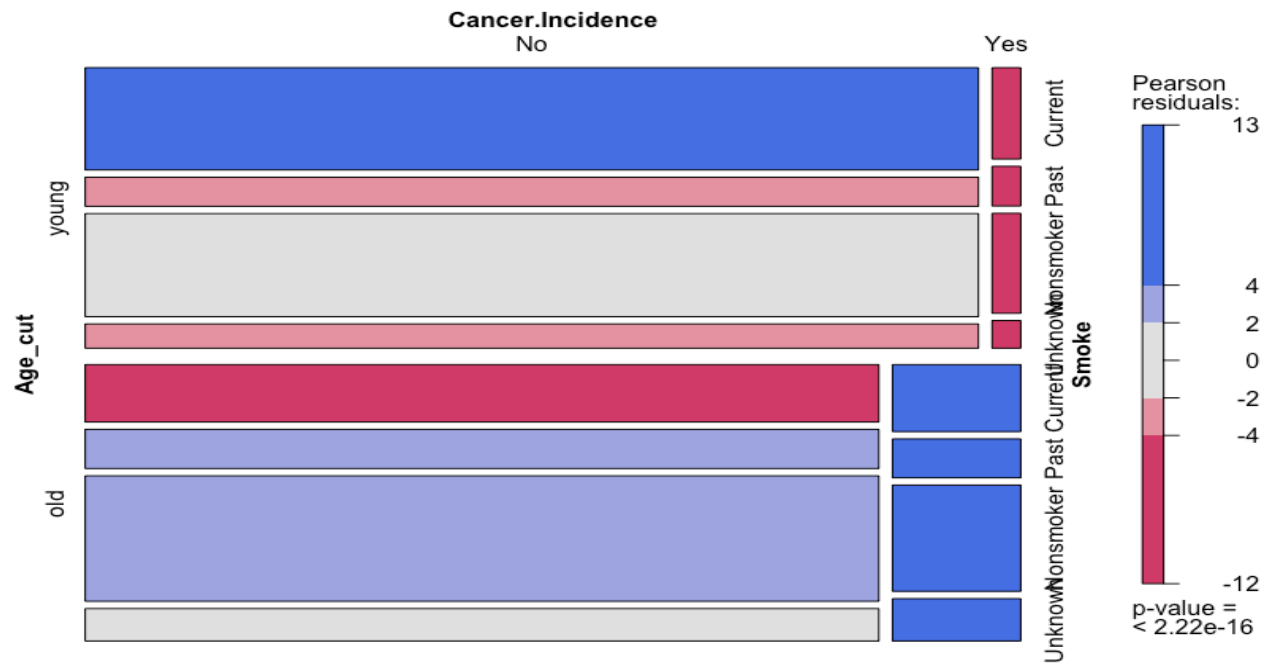




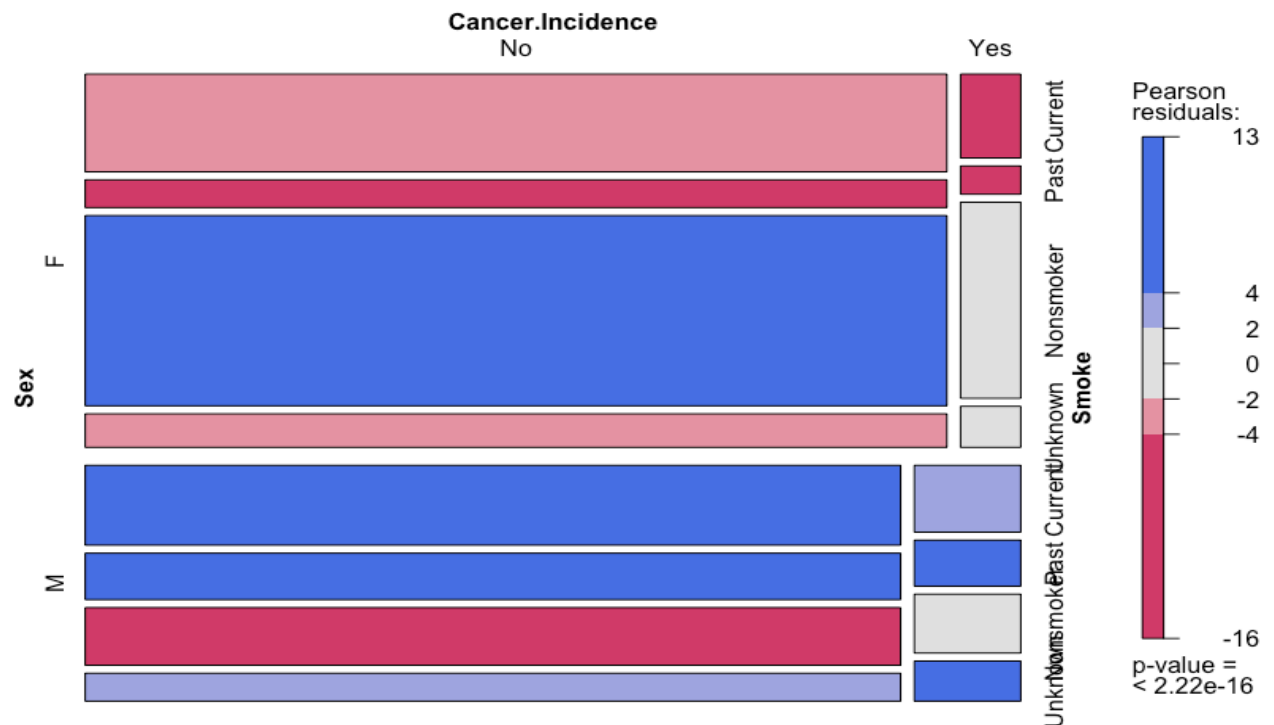
To examine the smoking effect on cancer incidences across age group, gender and race, I divided age into 2 groups, young (25-48) and old (48-75) and the mosaic plots below visualized the relationships:

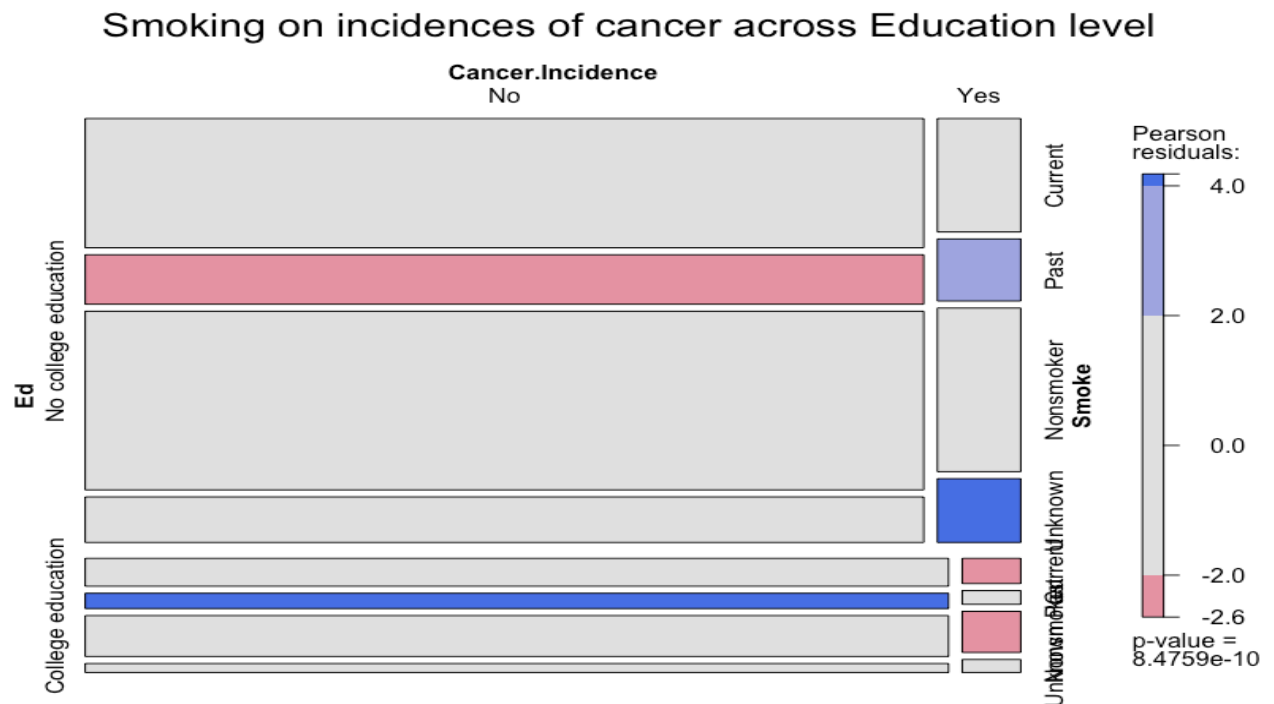
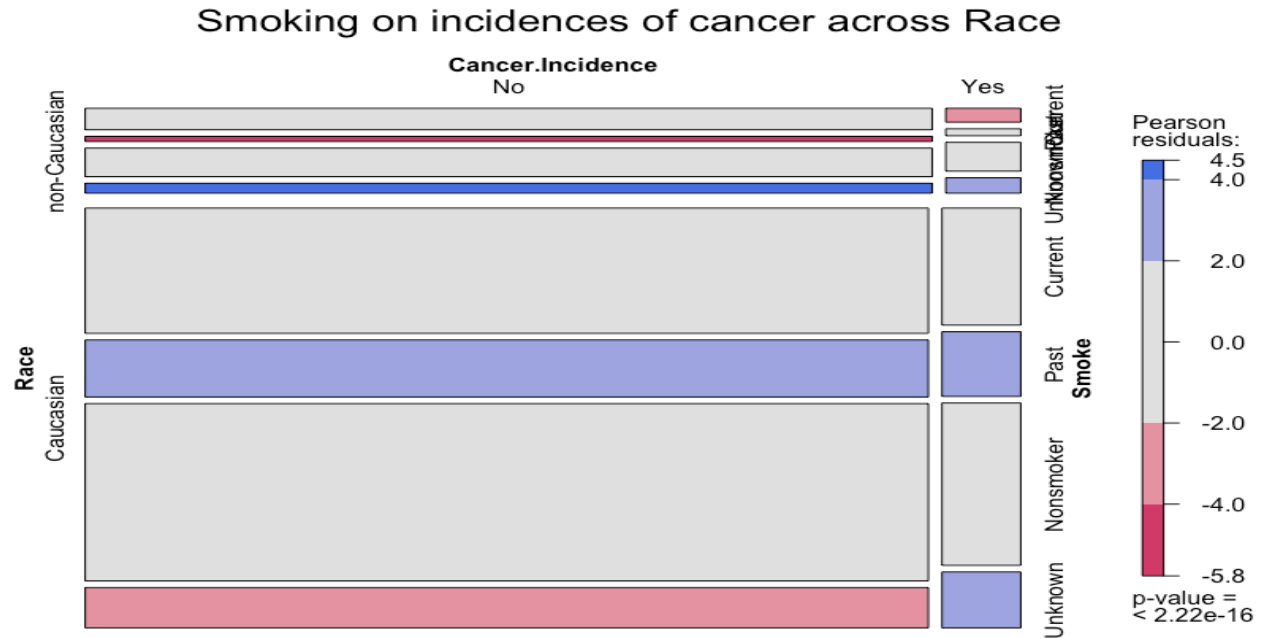


Smoking on incidences of cancer across age groups



Smoking on incidences of cancer across Sex

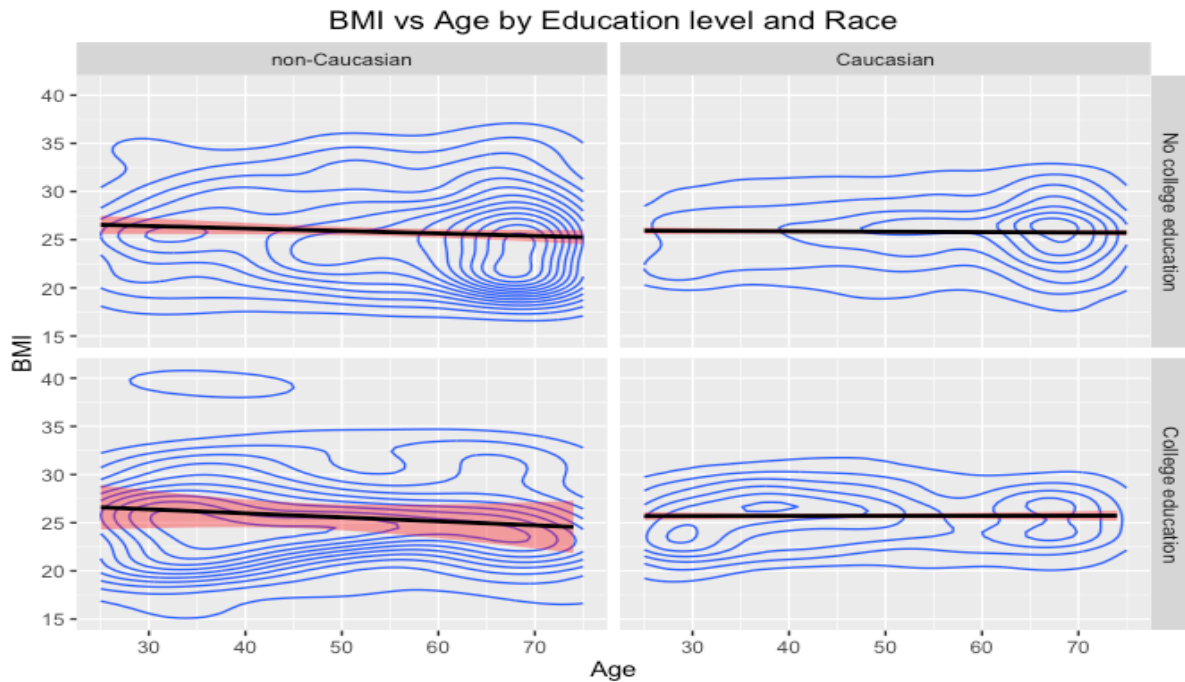




Smoking doesn't have significant effect on cancer incidences. However, more young people are current smoker and more females are non-smoker. Also, more Caucasian smoked in the past but their Cancer.incidence rate is not significantly different than non-Caucasian. Educated people have lower chance to get cancer because they might have better life-styles and health knowledge. Older people and male also have higher probability of getting cancer.

IV. Prediction

Although all the chemicals seem to have no relationship with BMI, there is a linear relationship with Age. Race and Education level could also help to predict BMI since non-Caucasian tends to have a wider range of BMI in different age and for educated non-Caucasian, there is a higher uncertainty on the true regression line.



V. Conclusion

To conclude, smoking doesn't contribute to Cancer. Incidence rate, but male, older and non-educated people have higher probability of getting cancer because age is a great risk factor for developing cancer while female and educated people usually concern their health more, therefore lead to lower Cancer. incidence. On the other hand, smoking would affect Hemoglobin in the body while Male and Caucasian would have higher level. Since the data have a lot missing values, it is helpful to predict the BMI with Age, Race and Education.