ANALYSIS OF BIG CITY SMOKING AND DRINKING DATA AMONG STUDENTS AND ADULTS

Final Report

Anusha Singamaneni

Table of Contents

1. **Executive Summary**

This project aims to analyze the major factors that influence smoking and drinking issues among high school students as well as adults in the most populated urban cities of the United States. The dataset being used is Big Cities Health data which contains health status of twenty-eight of the nation's largest and most urban cities, as captured by 34 health (and six demographics-related) indicators. Each city is rich with its own culture and history and I am considering the demographics of the subjects that are disproportionately scattered among the cities.

2. Questions of Interest

The main objective of the project is to address the following questions:

i) What are the major factors causing smoking and drinking problems among High School students in the most urban cities of the United States? How much are these conditions influenced by the place, ethnicity, and gender of the students?

ii) Similarly, how much effect do predictors like place, gender and ethnicity have on smoking and drinking problems among adults in US's biggest cities?

These questions have been chosen as the basis of research because smoking and binge drinking are major issues of concern especially among young students, and lead to dropouts, termination and such outcomes. Therefore, if analysis can be performed to identify influencing factors, the result can be utilized to curb such problems to an extent.

3. Data Extraction

(i) **Extracting and Refining Students' data**

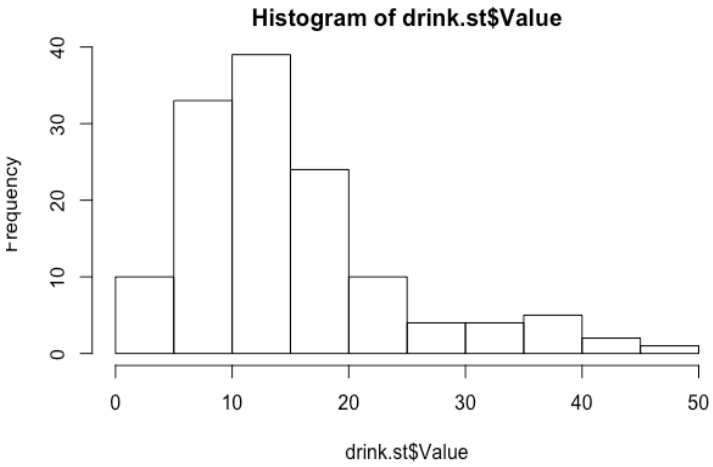
I first extracted the "Drinking and Smoking data" for high school students by applying the filters with Indicator as "Percent of High School Students Who Binge Drank" and "Percent of High School Students Who Currently Smoke" respectively. This gives two datasets drink.st and smoke.st. Once the data is extracted, I will retain only the variables of interest and eliminate the others, for further analysis. Additionally, I also renamed the column Race..Ethnicity to Ethnicity because it is not recommended to have special symbols in column names. Also, I would need to remove rows with missing values as they generate errors during analysis.

(ii)Extracting and Refining Adults' data

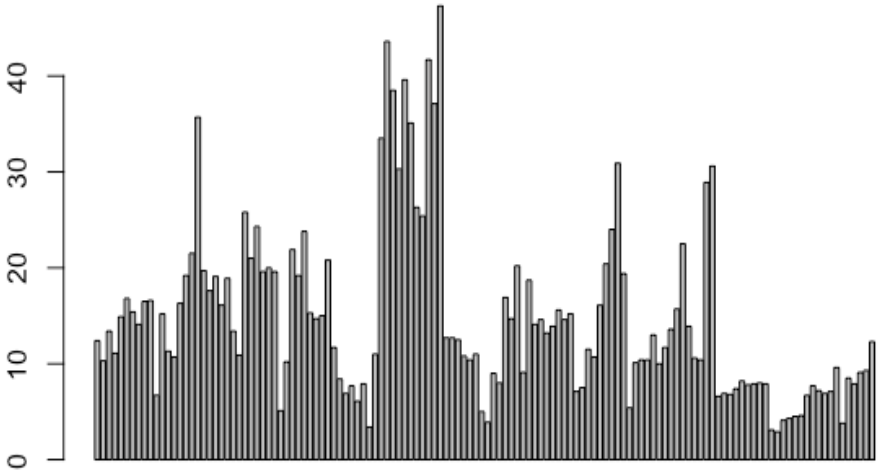
Similarly, I extracted the “Drinking and Smoking data” for Adults by applying the filters with Indicator as "Percent of Adults Who Binge Drank" and "Percent of Adults Who Currently Smoke" respectively. This gives me two datasets drink.ad and smoke.ad. Just like I did for Students data, I eliminate the unwanted columns, which are not required as part of our analysis. Similarly, columns are renamed, and missing values are eliminated.

4. Observation on Drinking and smoking for students

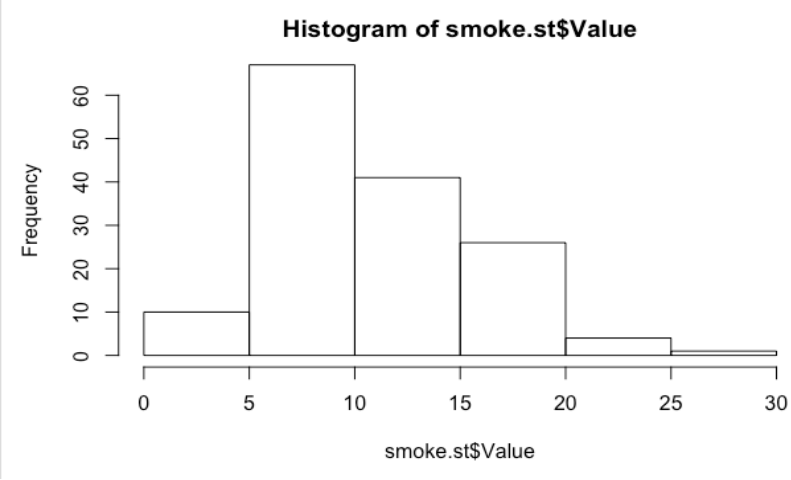
Histogram and bar plot of Drinking Data for students:

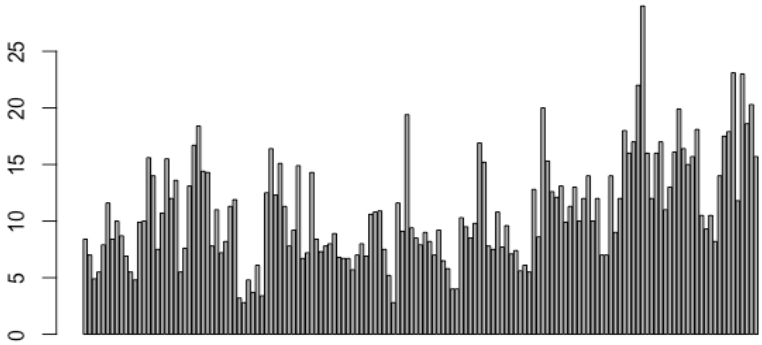


Bar plot of drink.st$Value



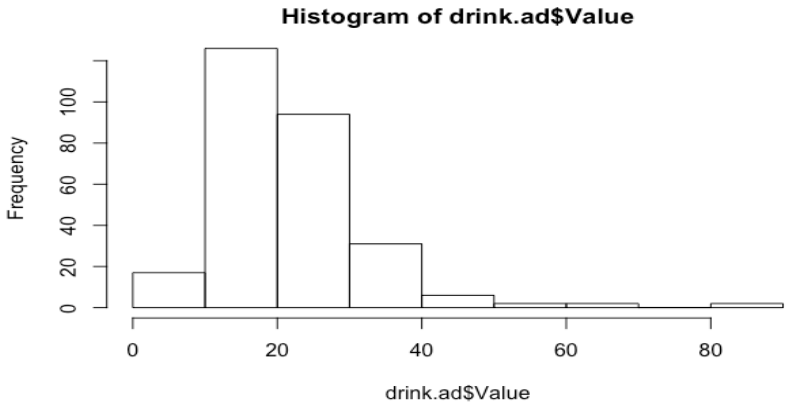
Histogram and bar plot of Smoking data for students:

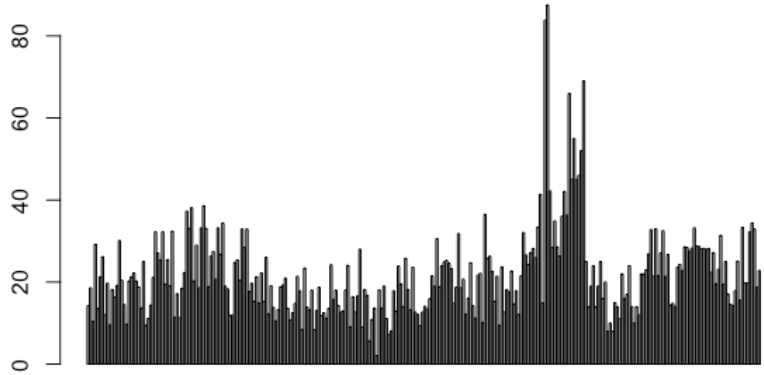




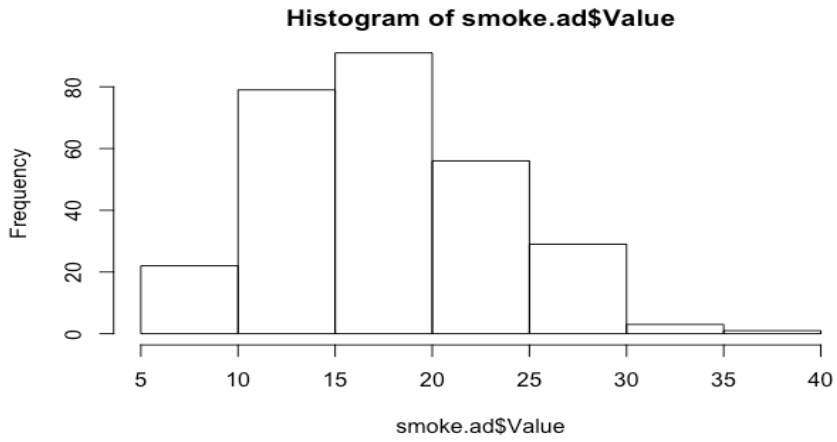
5. Observation on Drinking and smoking for Adults:

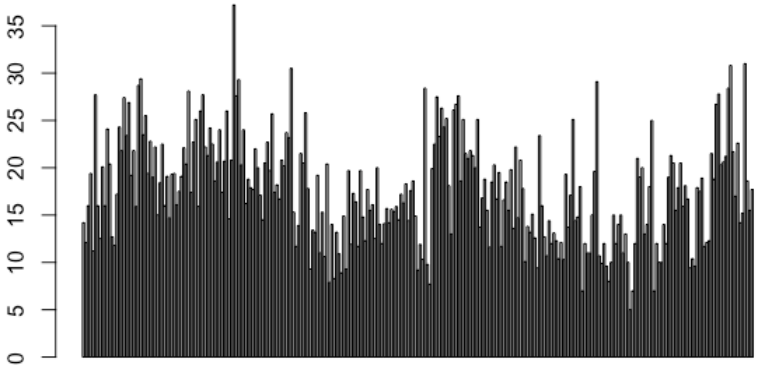
Histogram and bar plot of drinking data for adults:





Histogram and bar plot of Smoking data for Adults:





Observations:

Looking at the bar plots, pie charts and histograms of four datasets, we can make the following important and relevant observations:

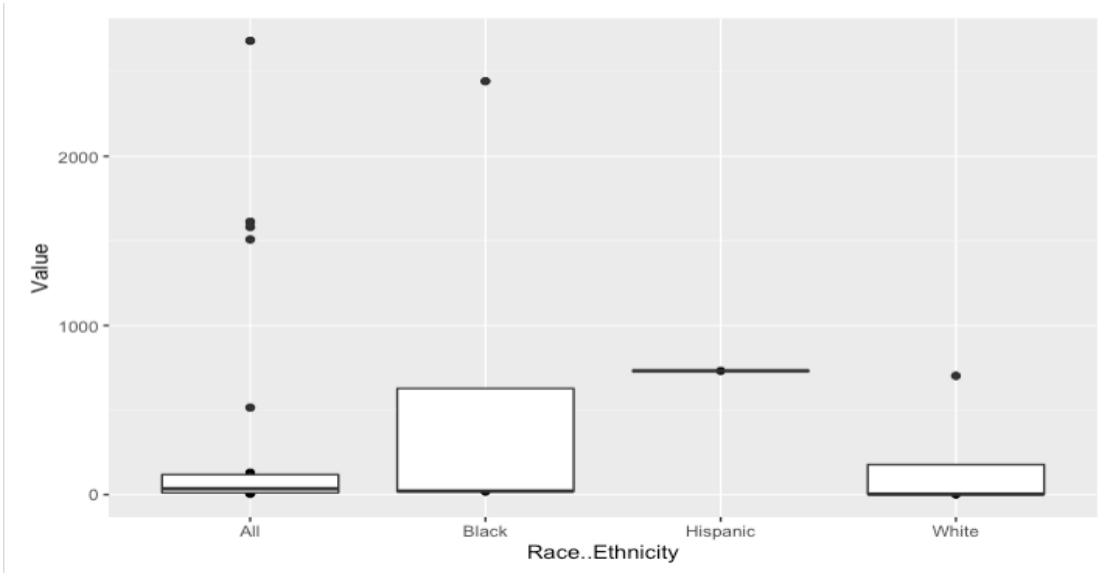
* More often than not, high school students seem to be binging on as many as 40 drinks and smoking more than 60 cigarettes, which is quite high.

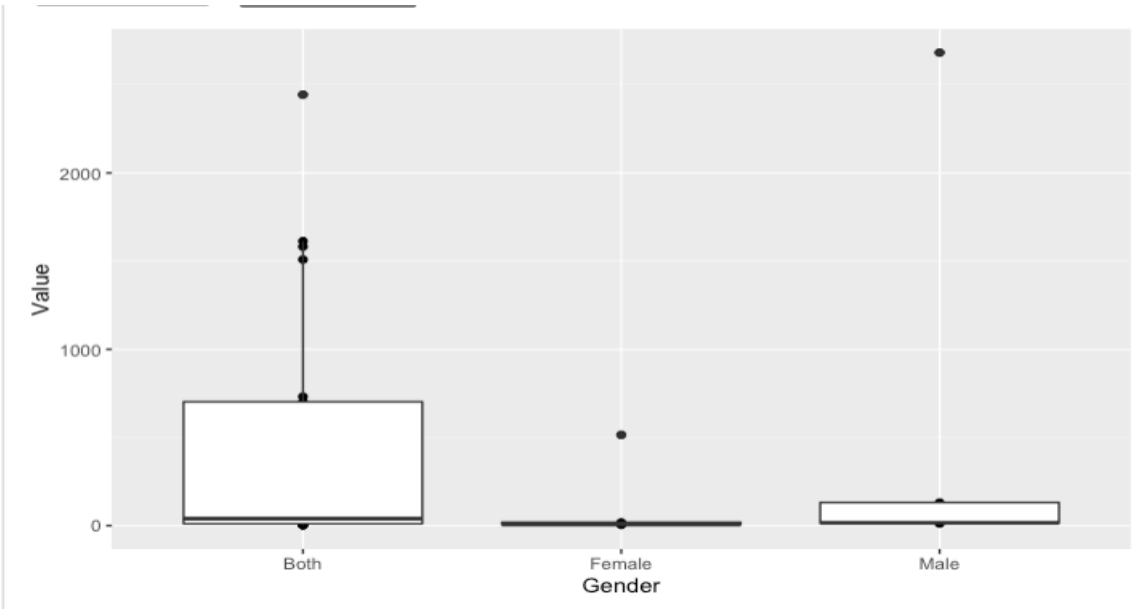
* Coming to the adults with smoking and drinking problems, they seem to be frequently consuming more than 120 drinks and smoking more than 80 cigarettes.

* On average, problems with binge drinking and smoking seem to be more prevalent among adults than high school students.

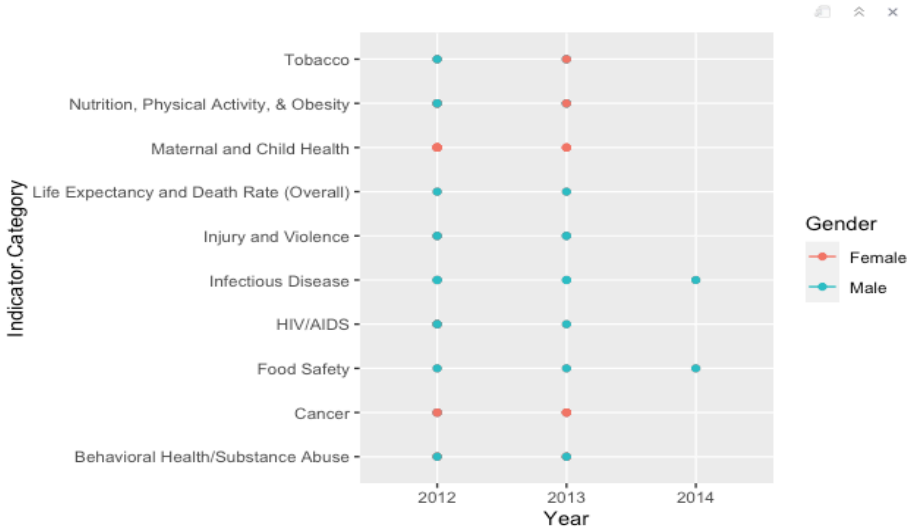
6. Analysis on data with different Plotting Observations:

(i) The plots show that the Ethnicity of Hispanic has the highest value which means that most Hispanic are affected with HIV/Aids and the other plots show that both male and female are influenced with this disease, but percentage of males are more than those of females in Atlanta

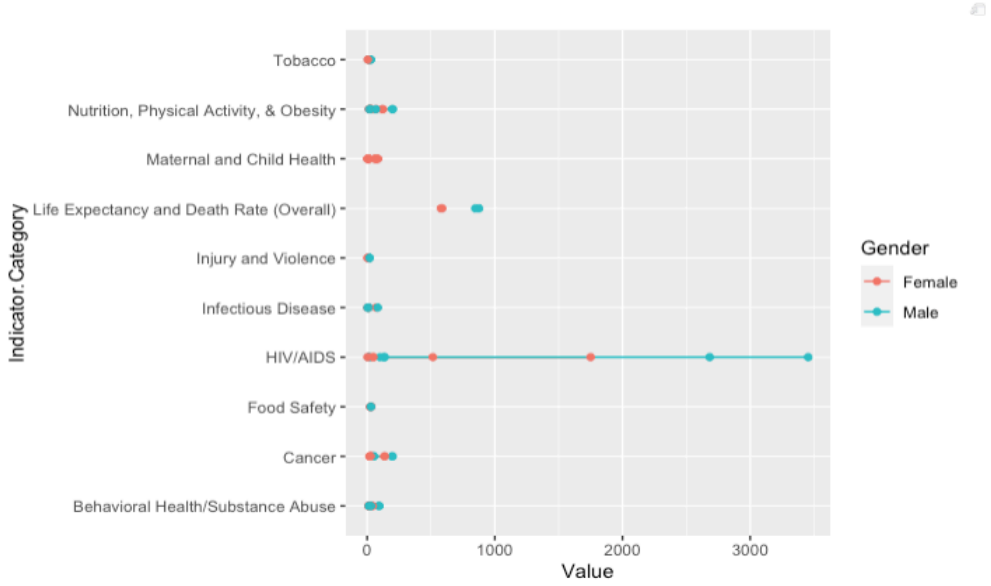




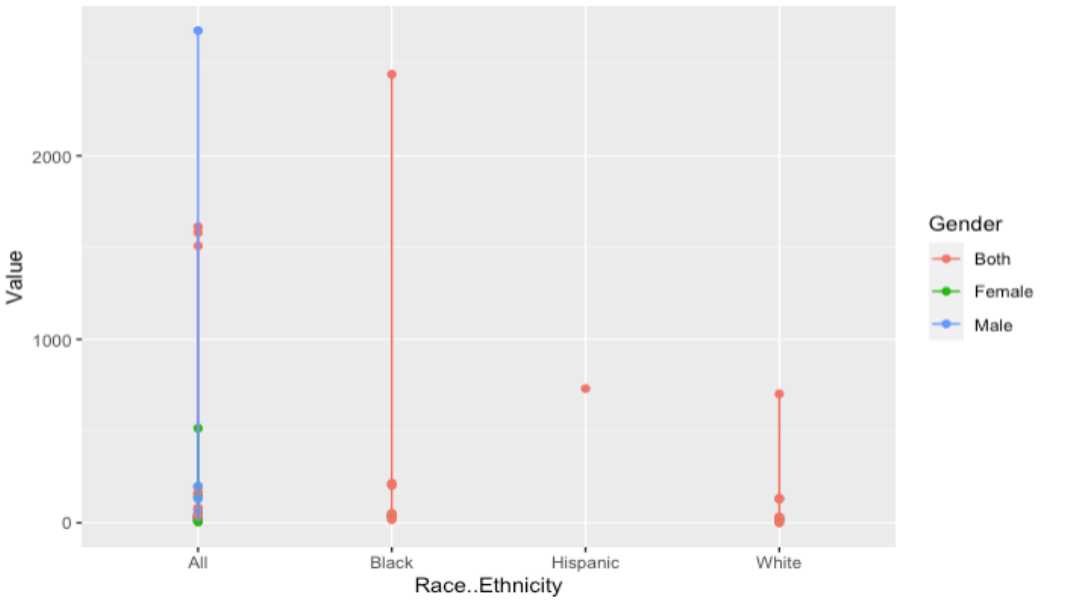
(ii) This plot shows that there are more male than female at Atlanta (Fulton County) , GA and Baltimore for the Years 2012-2014 who are affected with multiple types of disease



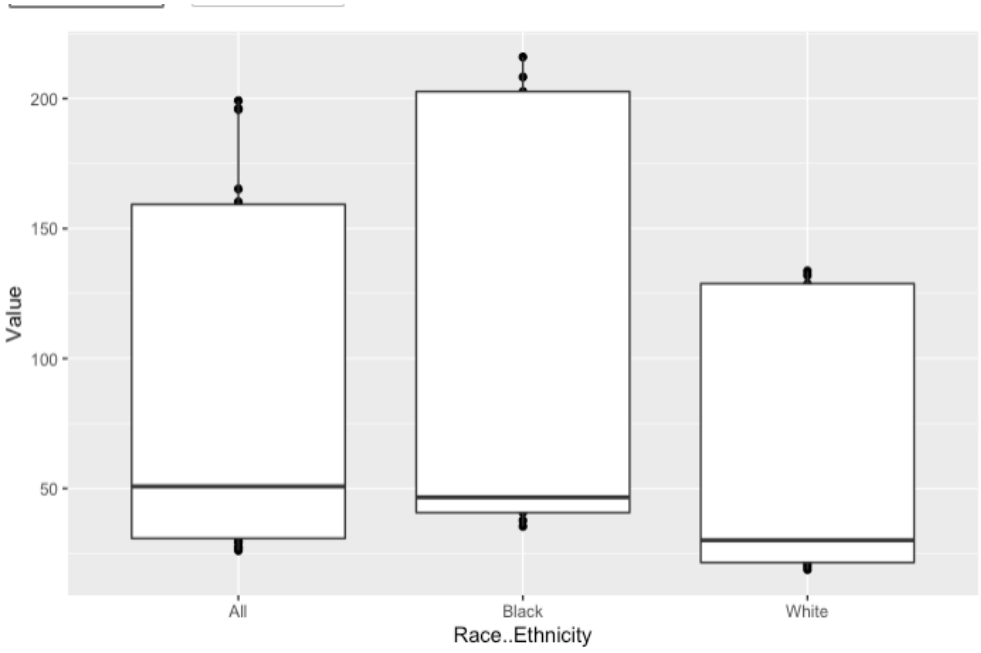
(iii) This plot shows that HIV/AIDS is the most affected disease of all others. And this disease is most seen in male population than in female in Atlanta.

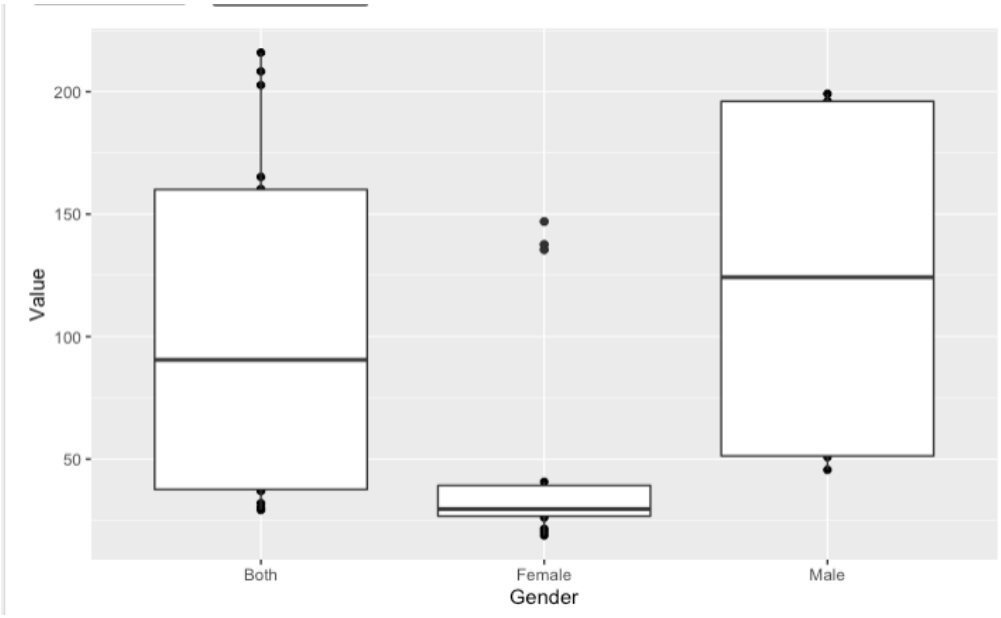


(iv) This plot shows that both male and female in Atlanta (Fulton County) , GA, with ALL ethnicity are the most affected population with Cancer and HIV/AIDS.



(v) This plot shows the Black (Ethnicity) are the people who are most affected with the Cancer in Atlanta. And most are the male population that are affected





7. Detailed analysis on four datasets (Performed linear Regression for more accurate results)

(i) Analysis on Drinking data for high school students:

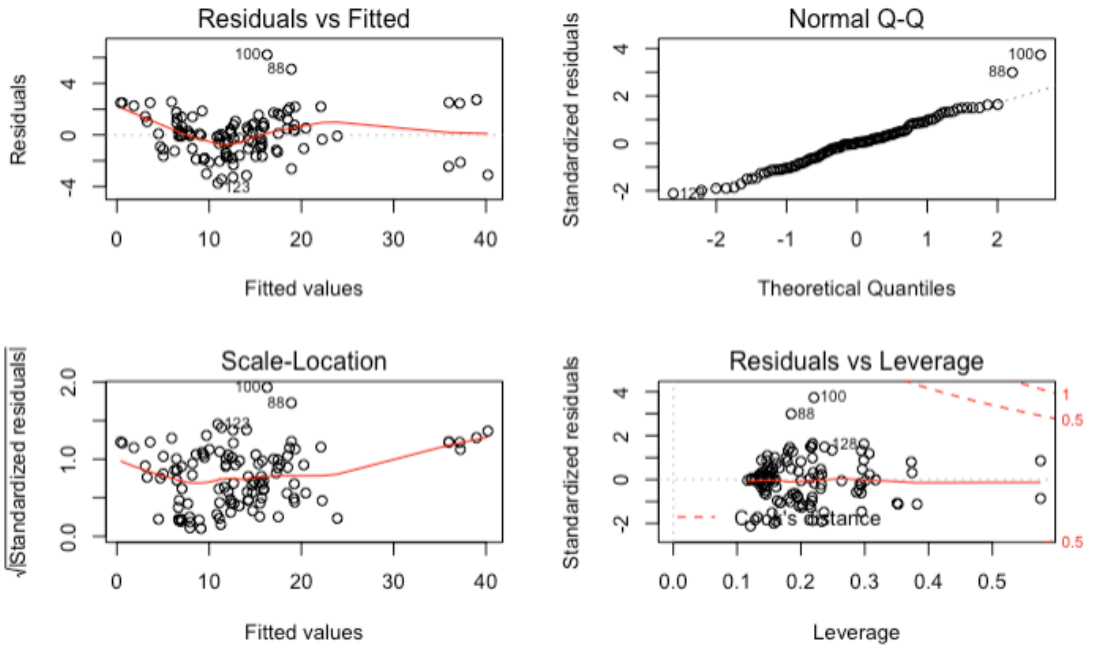
The below plot shows the students who are affected with drinking problems according to their place, Ethnicity and year. It shows Miami ,FL has the highest value which means the students of this state are highly affected with drinking problem.

To know what the major factors are causing this drinking problem performed linear regression on the students drinking data.

The results show a strong evidence that Place and Ethnicity are the major influencing factors causing high school students to binge drink

Below shown residual plots also adds strength for the same.

There are few outliers found while performing the linear fit, so removed them and run the linear regression again for better fit.



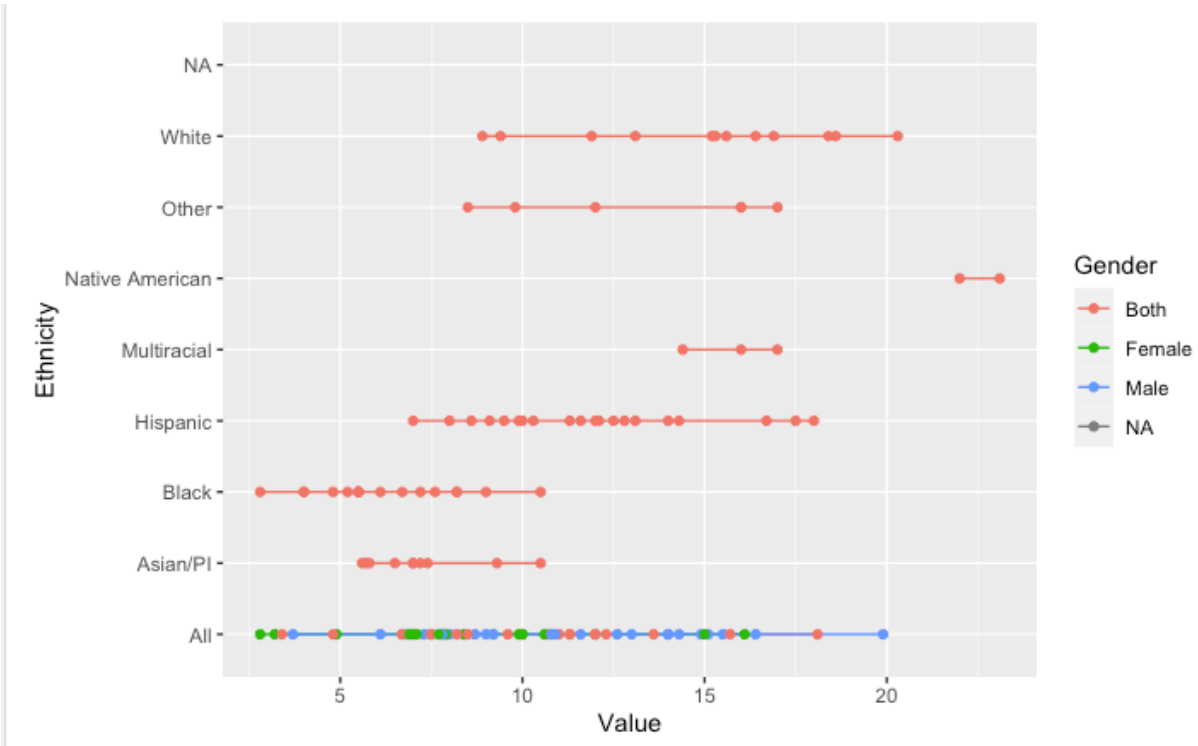
Approximately 93 % of variation in value variable can be explained by this model with these two independent variables (Place and Ethnicity). Very low P-value also strenghtens this assumption.

The residual standard error also shows there is not much distance between our observed value(Y) from the predicted Value(Yhat).

We can see from the plots that there exist leverage points , so removed all those as they are not much influential, and run the model again to show best fit.

(ii) Analysis on High school students who smoke

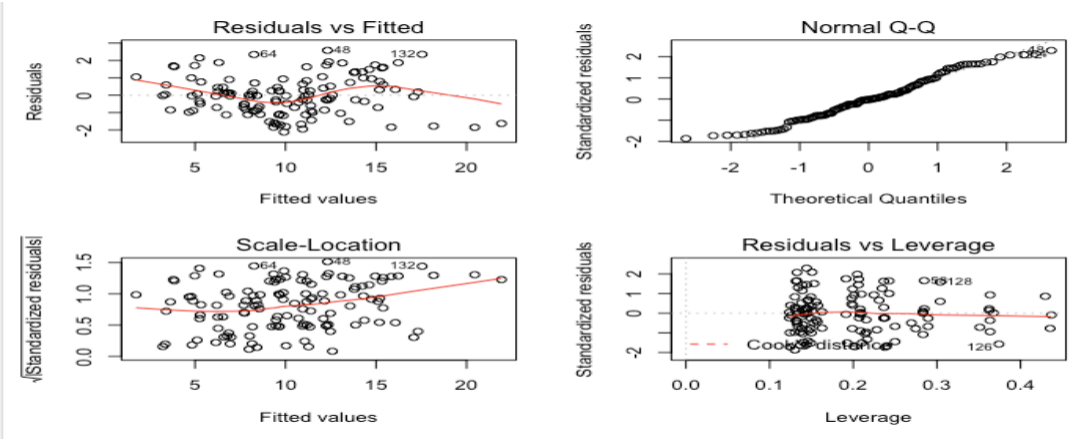
The below plot shows the population of students who are affected with smoking problem according to their place, and year. It shows in the year 2013 the US Total both male and female with of White Ethnicity are mostly affected with the smoking problems.



To show what are the major influencing factors causing this problem performed linear regression analysis on student's smoke data.

This fit also shows that place and Ethnicity are the major influencing factors

And below are the residual plots for the regression analysis.



After preforming Linear regression on both smoke and drinking data of High school students, the observations are,

Approximately 93 % of variation in value variable can be explained by this model with these two independent variables (Place and Ethnicity). Very low P-value also strengthens this assumption.

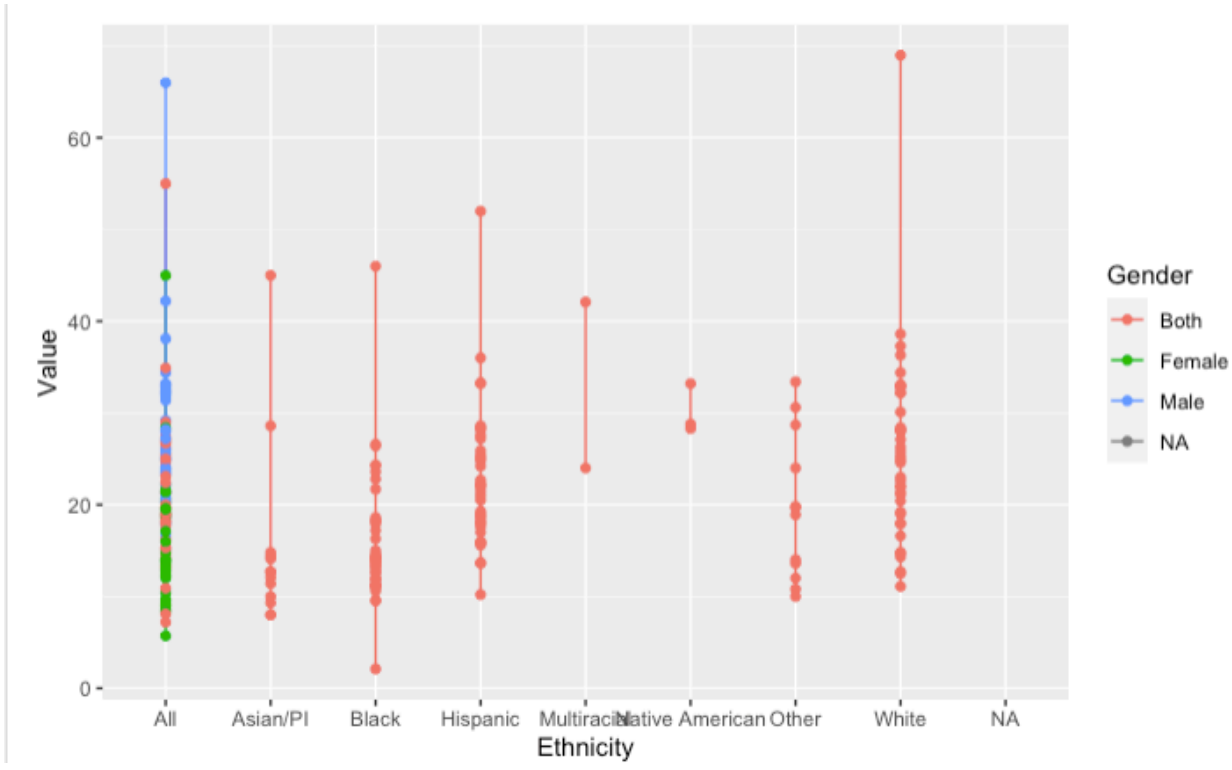
The residual standard error also shows there is not much distance between our observed value(Y) from the predicted Value (Yhat).

We can see from the plots that there exist leverage points, so removed all those as they are not much influential and run the model again to show best fit.

(iii) Analysis on Adults drinking data

The plot shows the value that the people are affected with drinking problem according to their respective place and ethnicity, both male and female adults are affected with drinking problem, but males are most addicted than females

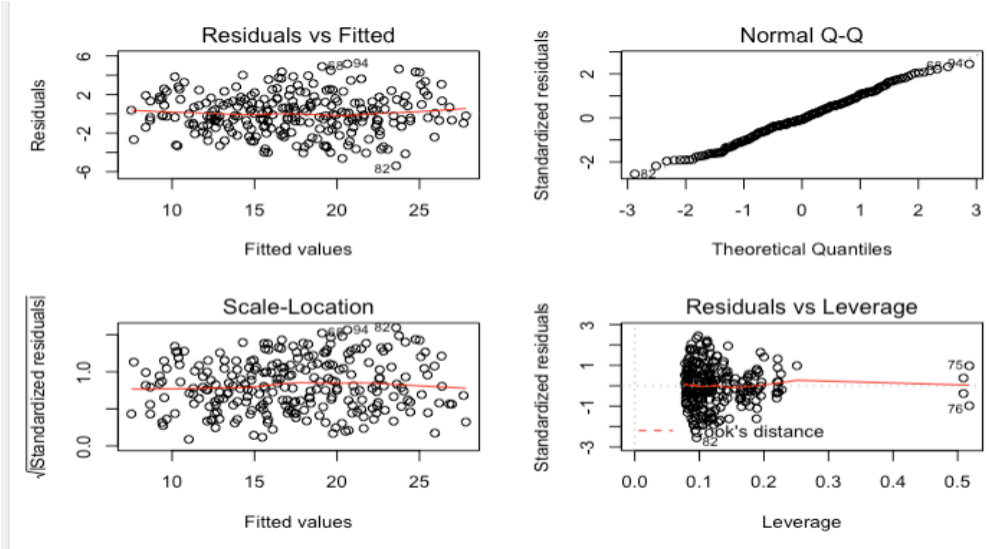

NA
Washington, DC
U.S. Total
Seattle, WA
San Jose, CA
San Diego County, CA
San Antonio, TX
Phoenix, AZ
Philadelphia, PA
New York, NY
Minneapolis, MN
Miami (Miami-Dade County), FL
Los Angeles, CA
Las Vegas (Clark County), NV
Houston, TX
Fort Worth (Tarrant County), TX
Denver, CO
Chicago, IL
Boston, MA
Baltimore, MD
Atlanta (Fulton County), GA
Place
All
Asian/PI
Black
Hispanic
Multiracial
Native American
Other
White
NA
Ethnicity
Value
60
40
20


60
40
20
0
Value
All
Asian/PI
Black
Hispanic
Multiracial
Native American
Other
White
NA
Ethnicity
Gender
Both
Female
Male
NA

The linear regression analysis on this adult drinking data shows that Drinking is more among male adults than in female.

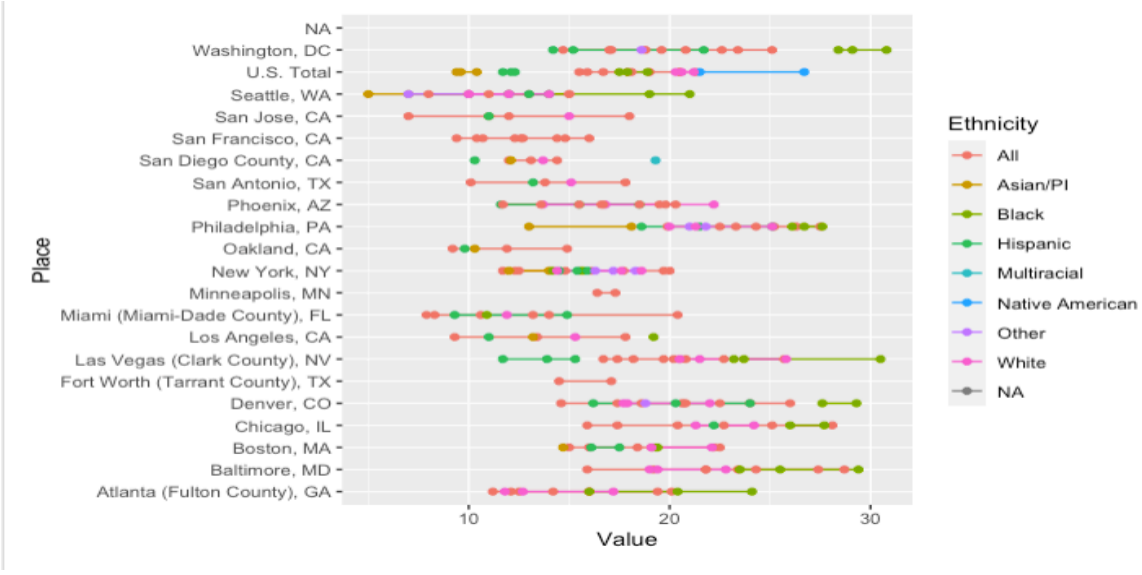
This fit shows us that all the predictors are significant for the response Value. But Year and Gender are best predictors of all four. Residual standard error show there is not much difference between observed and predicted value. Approximately 88% of variance can be explained using this model, and the p-value is very low, stating that this model is statistically significant.

Leverage points are removed and re-run the fit again, to get a best fit.



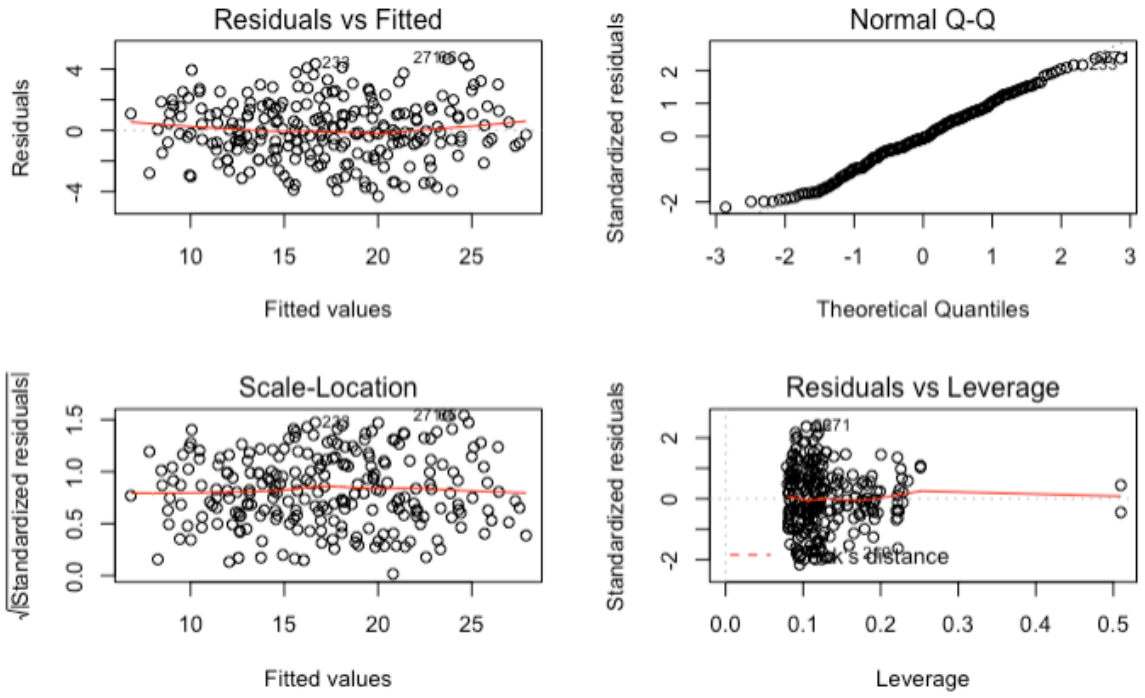
(iv) Analysis on Adults smoking Data

The plots show the adults smoking data according to their place and Ethnicity. It can be seen that Las Vegas NV, and Washington DC are the most affected states with smoking problems among adults.



This fit shows that Place and Ethnicity are most influential factors in causing smoking among Adults. The residual error with 2.801 on 193 degrees of freedom show there is no significant difference between the observed and predicted value. This model explains approximately shows 76% of variance and it has a very low P-value. So, this model is statistically significant.

Leverage points removed and re-run the fit again as there is no influence of these points in our data.



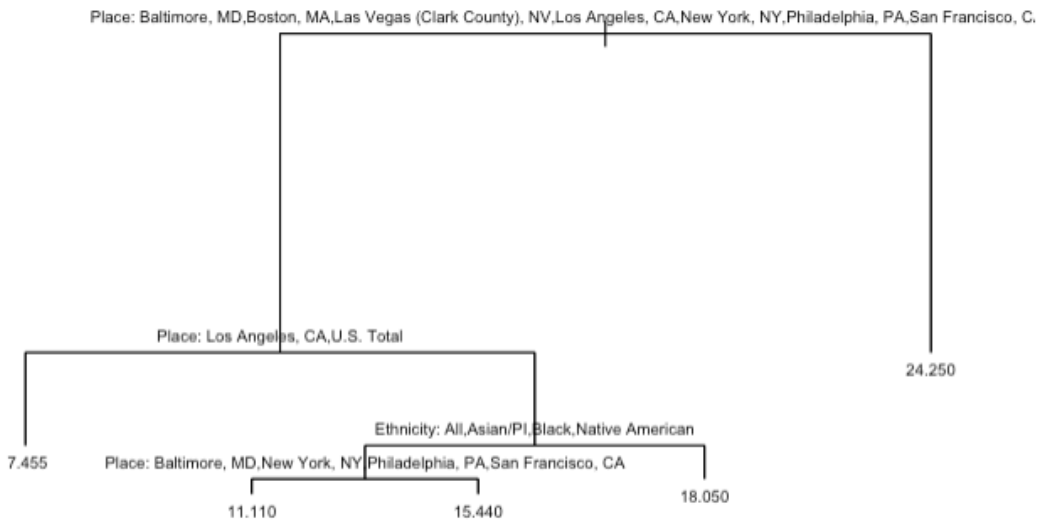
Conclusion:

After analysis on all four data sets, we can say that Place is the most significant factor followed by Ethnicity. These are the major influencing factors causing drinking and smoking among both High school students and Adults.

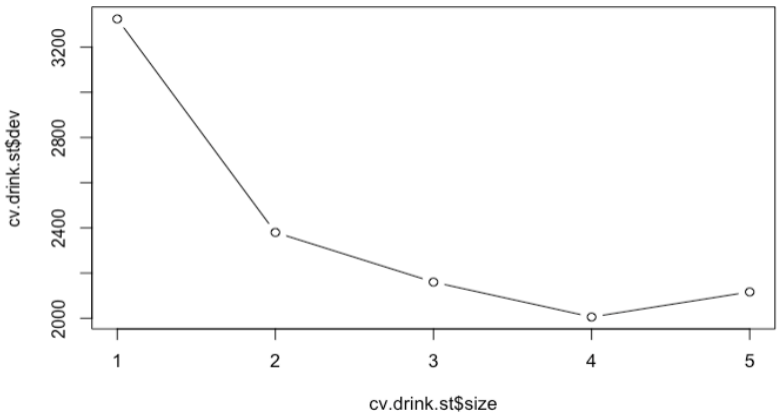
8. Regression Trees analysis

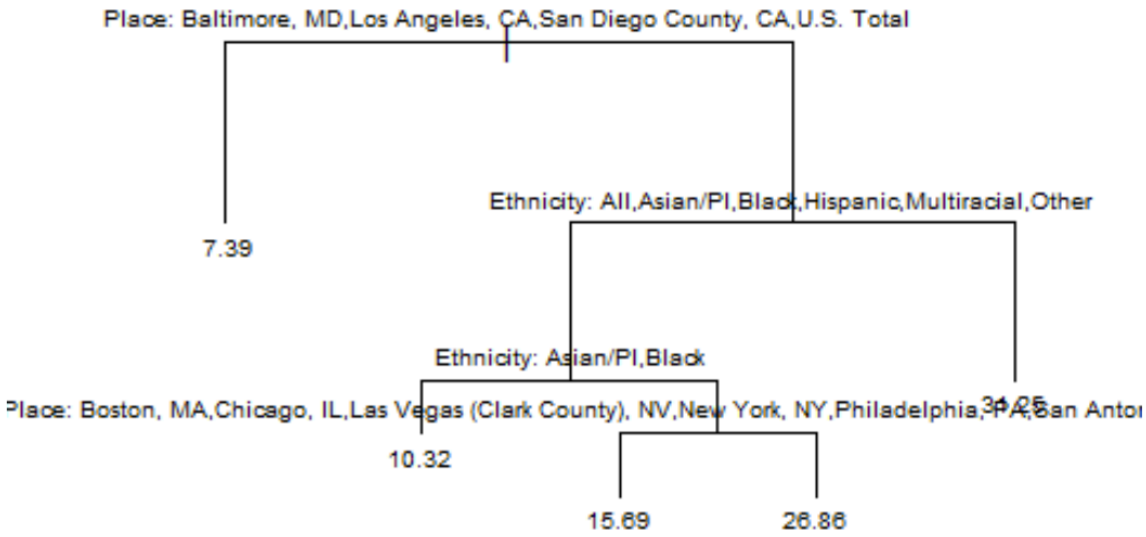
Performed another type of statistical model called Regression Trees, because it has a good graphical representation

(i) Basic Regression trees for Students drinking data

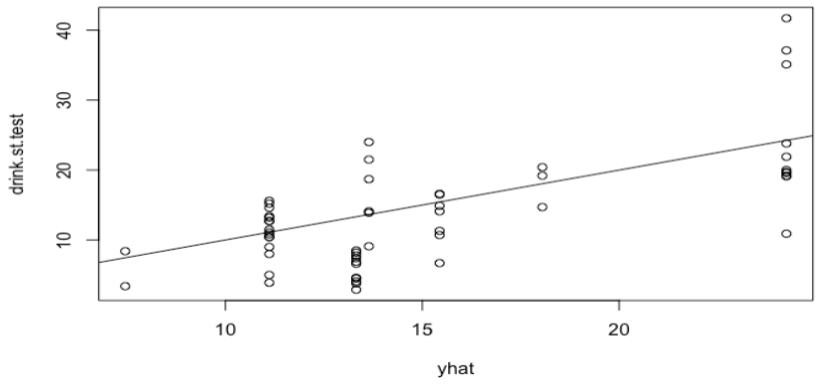


After pruning the tree



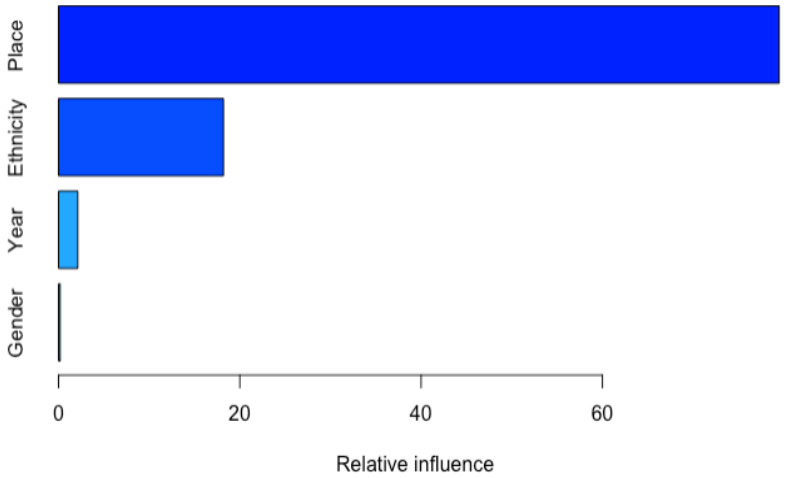


Plot on test data:

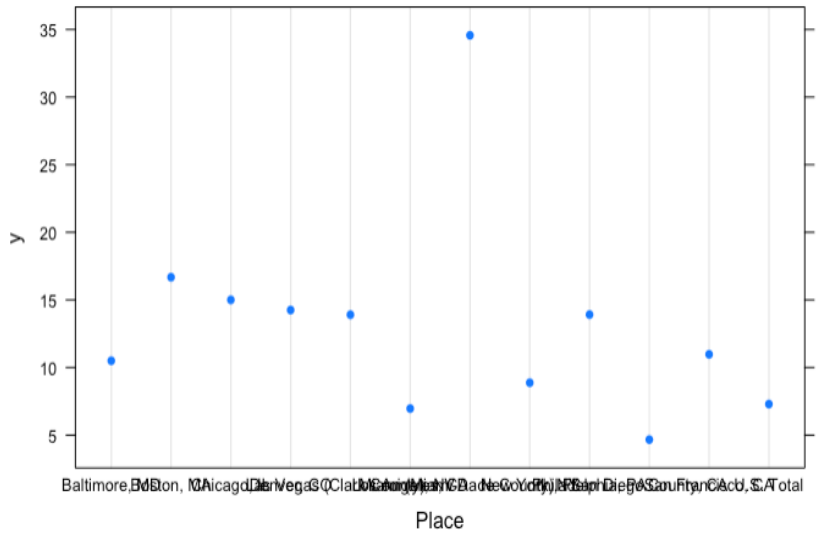


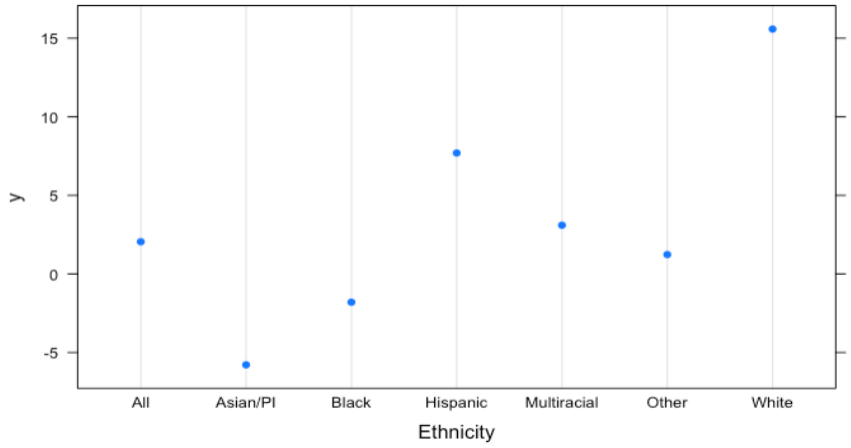
The tree after pruning to 5 terminal nodes seems to be easier to interpret and has a better graphical representation. The best predictor seems to be Place because it is used for the initial split, where the Place is Baltimore, Los Angeles, San Diego on one side and all other cities (25 cities) are on the other side. Ethnicity is the next best predictor used and the tree is split depending on it being Asian/PI, Black, Hispanic, Multiracial on one side and all other ethnicities on the other side. The tree() function has used only Place and Ethnicity for building the Regression tree. In addition, we can see that the predictions have higher values on the left sub-tree as compared to the other side, which is as expected.

Performed Boosting analysis to get better fit and clear plots:



Place and Ethnicity are the most important variables as seen above. We can also produce partial dependence plots for these two variables. The plots below show marginal effect of selected variables on the response.





Conclusion:

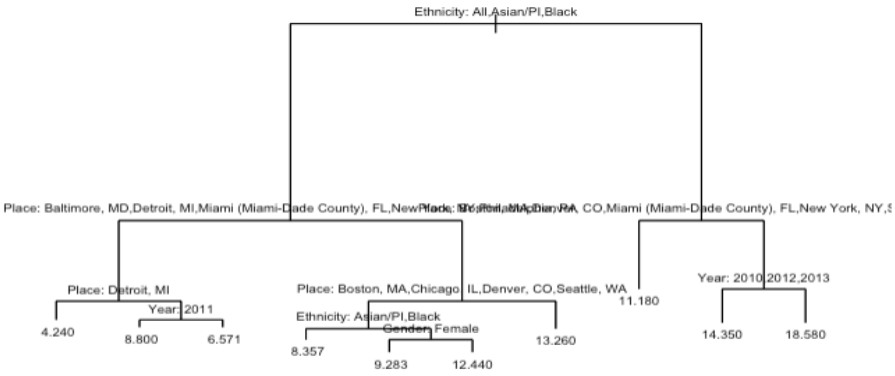
As seen above the most important predictor is **Place** and the next best predictor is **Ethnicity**.

On average, when we examine the plots generated after Boosting, we find that the cities **Miami, Florida and San Antonio, TX** have the highest problem of Binge Drinking among High School students. Cities such as **Los Angeles, CA and San Diego County, CA** have the least indicator values leading to the inference that these cities seem to have least binge drinking problems among students.

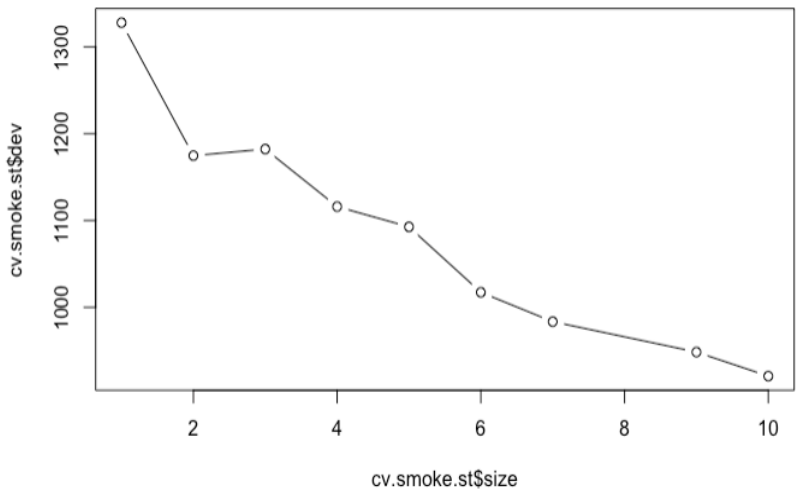
When it comes to ethnicities, we find that White community has the highest drinking rate among students and Black, Asian/PI have the lowest rates.

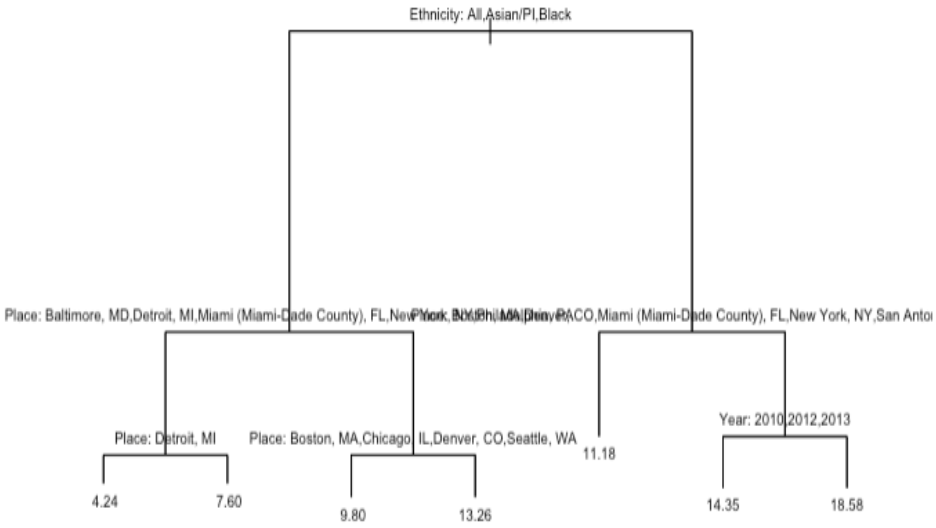
(ii) Smoking data for Students:

Basic Regression tree for smoke data



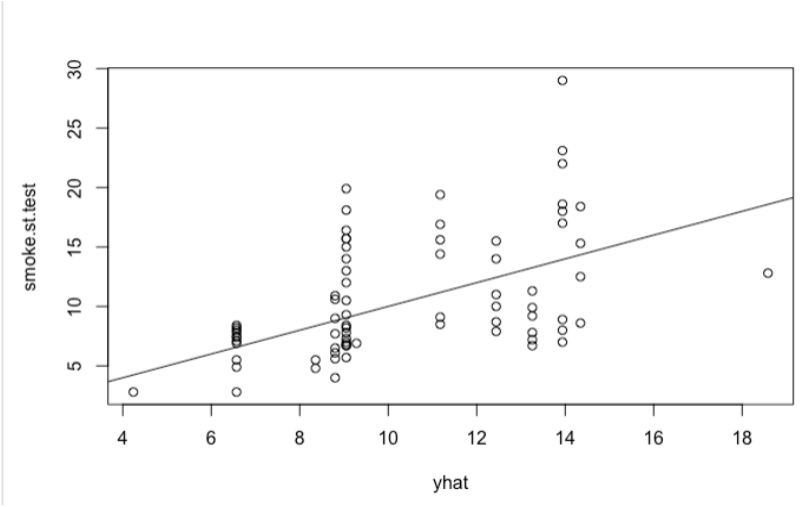
After pruning the tree



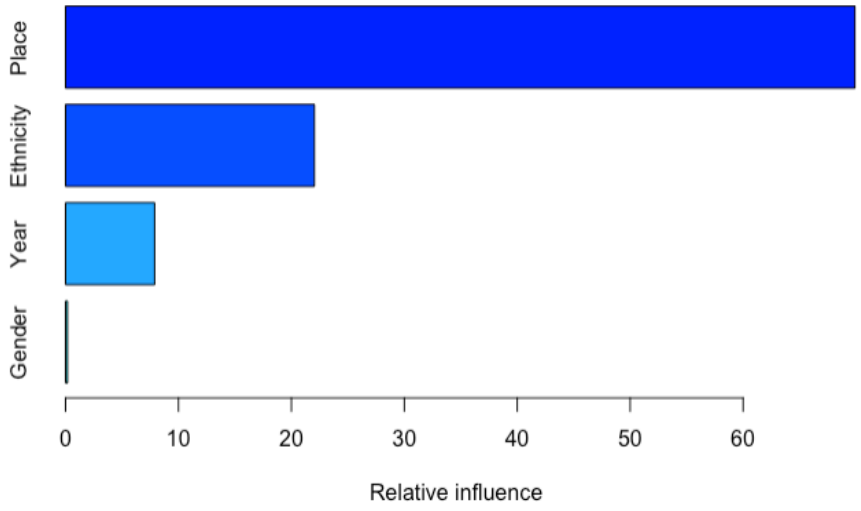


The tree after pruning to 7 terminal nodes seems to be easier to interpret and has a better graphical representation. The best predictor seems to be Ethnicity in this case because it is used for the initial split, where the Place is Asian/PI, lack on one side on one side and all other ethnicities are on the other side. Place seems to be the next best predictor used and the tree is split with Miami, New York, San Francisco on the higher side and all others on the other lower side. The Year is also used to split the data on the left sub-tree.

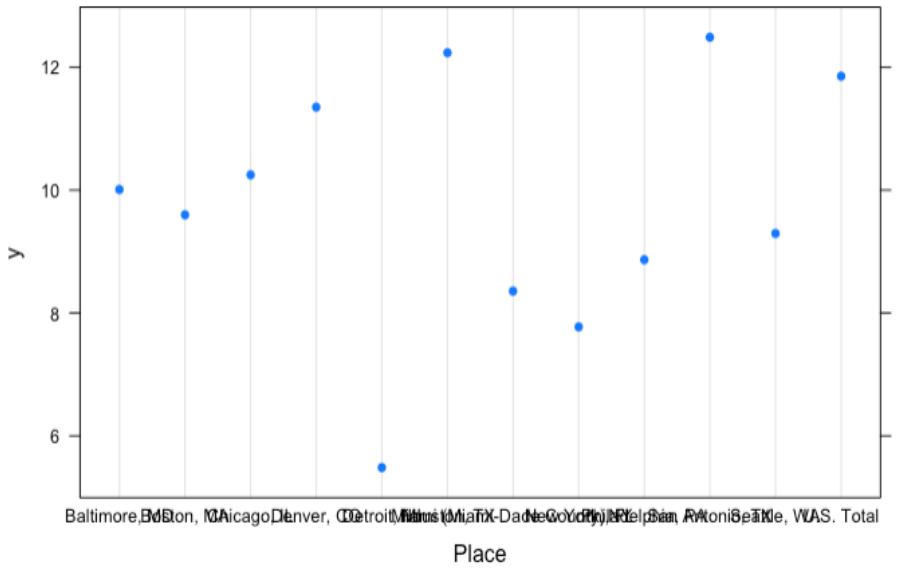
Plot on predicted test data:

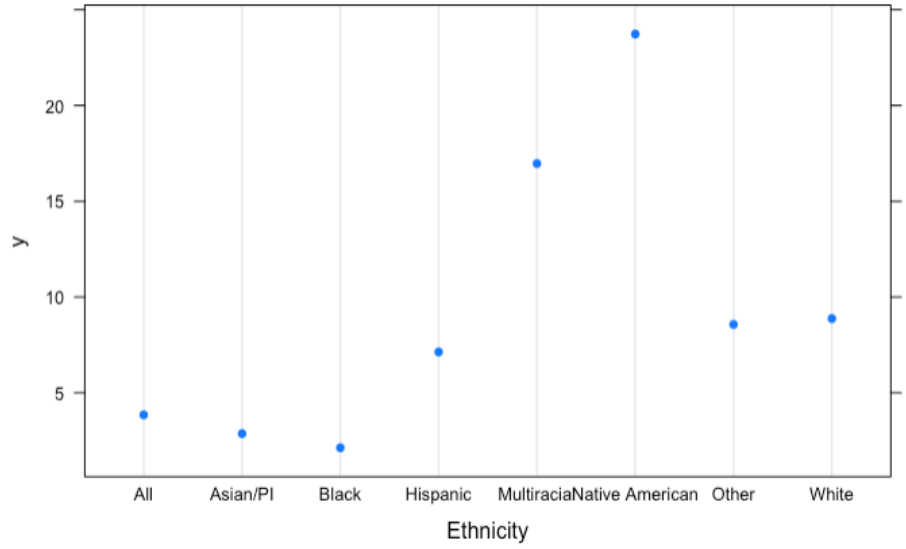


Boosting :



Place and Ethnicity are the most important variables as seen above. We can also produce partial dependence plots for these two variables. The plots below show marginal effect of selected variables on the response.



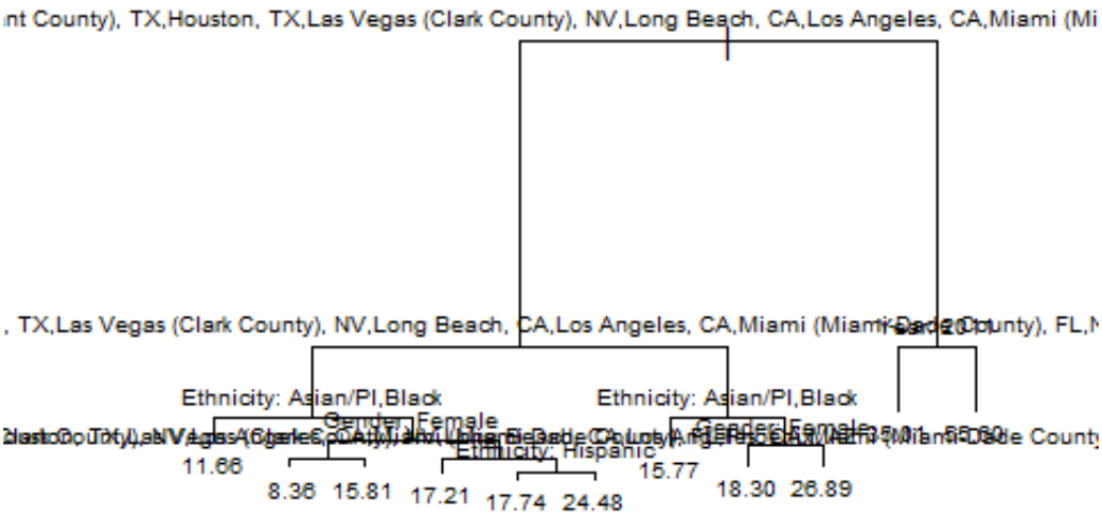


As seen above the most important predictor is **Place** and the next best predictor is **Ethnicity**.

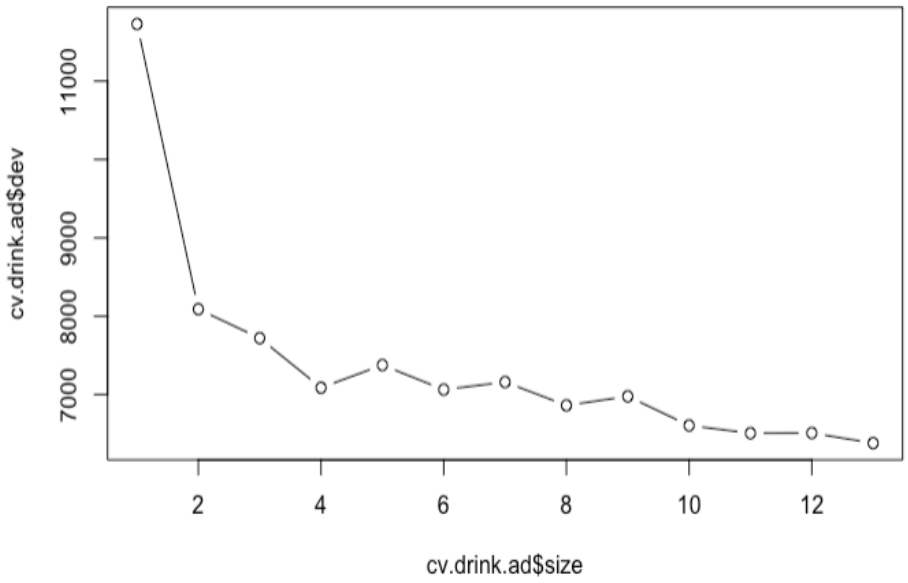
Also, as seen in the plots generated after Boosting, we find that **Miami and Seattle** have higher smoking rates among high school students whereas the rates are lowest in **Detroit**. Similarly, when it comes to ethnicities, smoking rates are highest in Multiracial section of society and lowest in Black, Asian/PI communities.

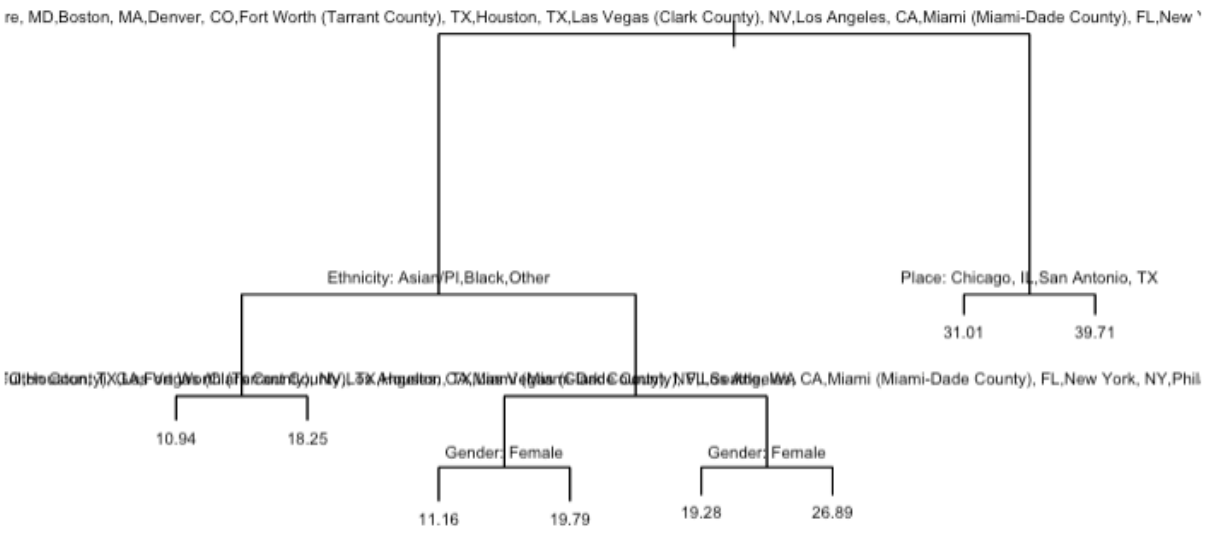
(iii) Drinking data for Adults:

Basic Regression Tree:



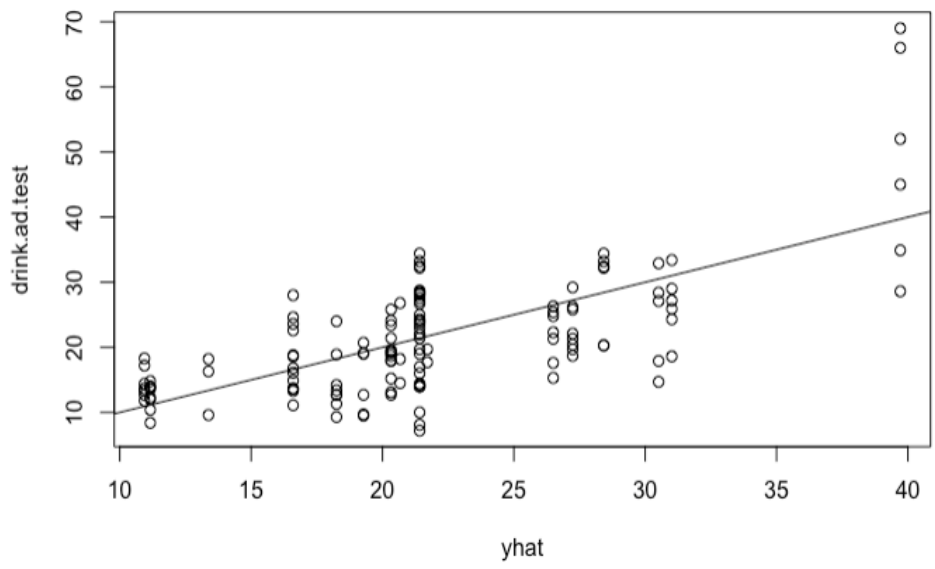
After Pruning the tree:





The tree after pruning to 8 terminal nodes seems to be easier to interpret and has a better graphical representation. The best predictor seems to be Place followed by ethnicity where Asian/PI, Black are on the lower side. Year also seems to be important as the year 2013 seems to have higher rates of drinking issues among adults.

Plots on test data:



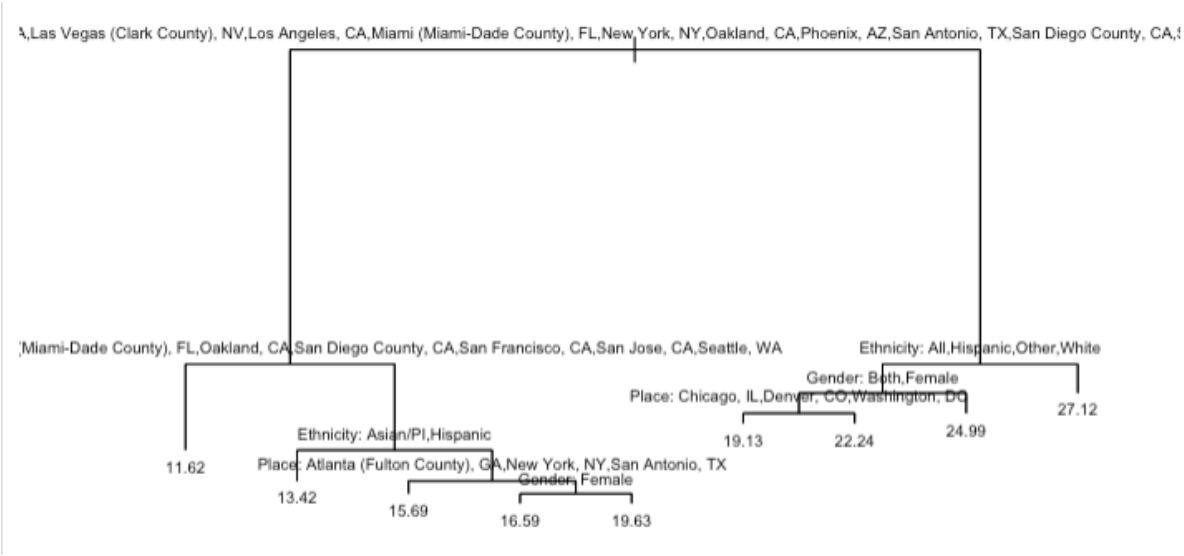
Again, it is observed that the most important predictor is **Place** and the next best predictor is **Ethnicity**.

When we check the plots, it is clear that **San Jose** has the highest rate of binge drinking among adults and Las Vegas and Houston seem to have lower binge drinking rates.

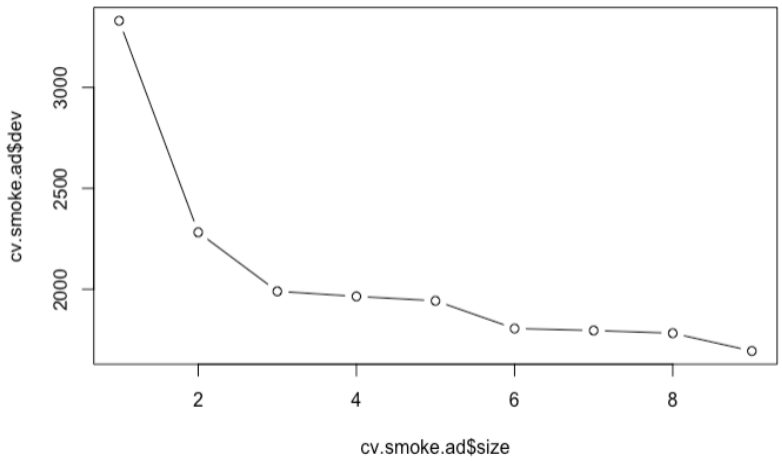
When we check ethnicities, it can be observed that Multiracial has highest rate and Asian/PI and Black seem to have the lowest rates.

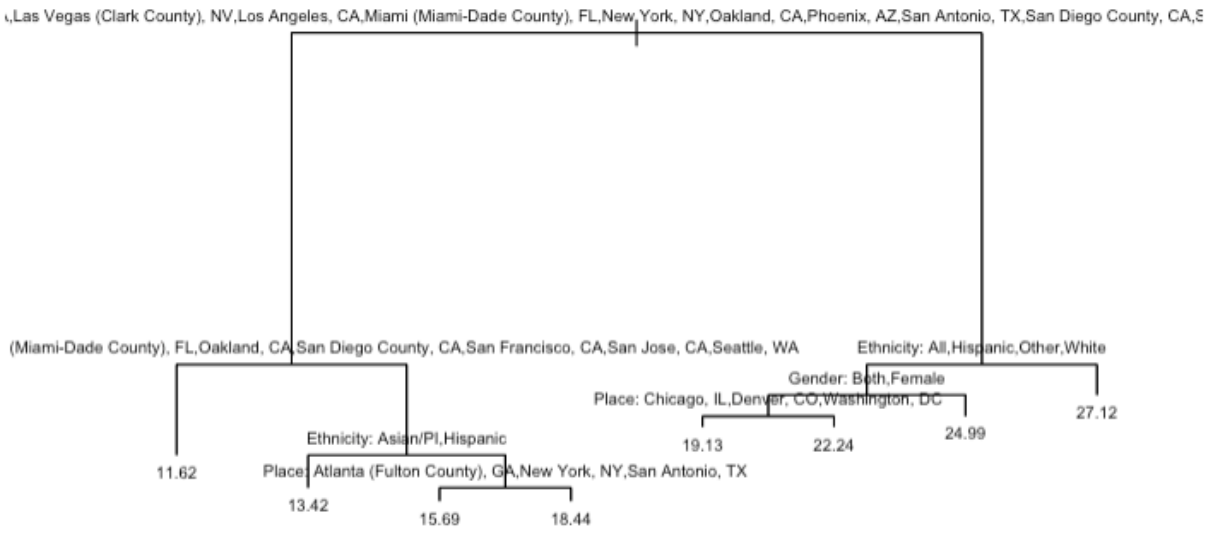
(iv) Smoking data for Adults:

Basic Regression tree :



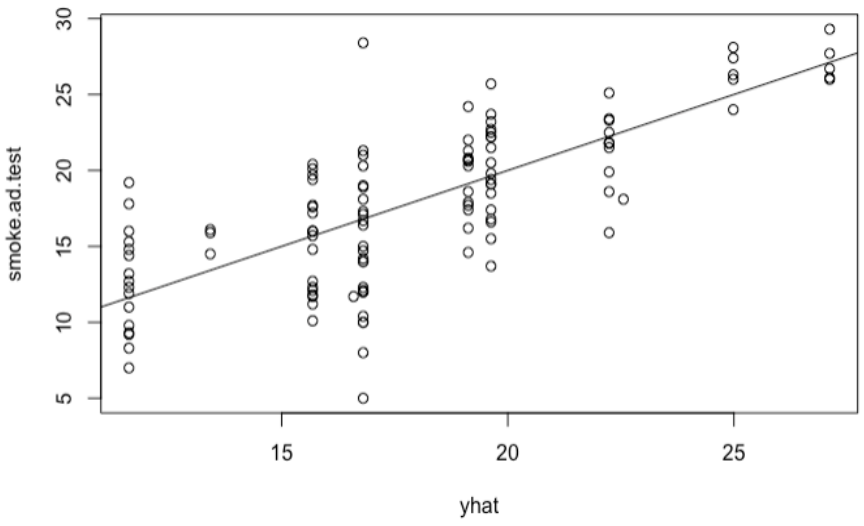
After pruning the tree:



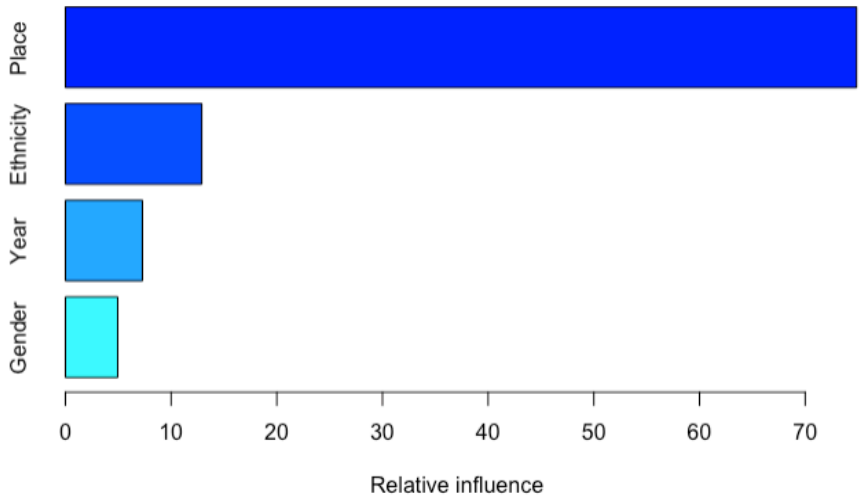


For easier interpretation and better graphical representation, we performed tree pruning to 8 terminal nodes. The best predictor seems to be Place again because it is used for the initial split. Cities Miami, Oakland, San Antonia seem be on the lower side and cities Las Vegas, Chicago and Boston seem to be on the higher end of the smoking rates among adults. The next predictor Ethnicity has Black, Multiracial on the higher side. The tree() function has used Place, Ethnicity and Gender for building the Regression tree.

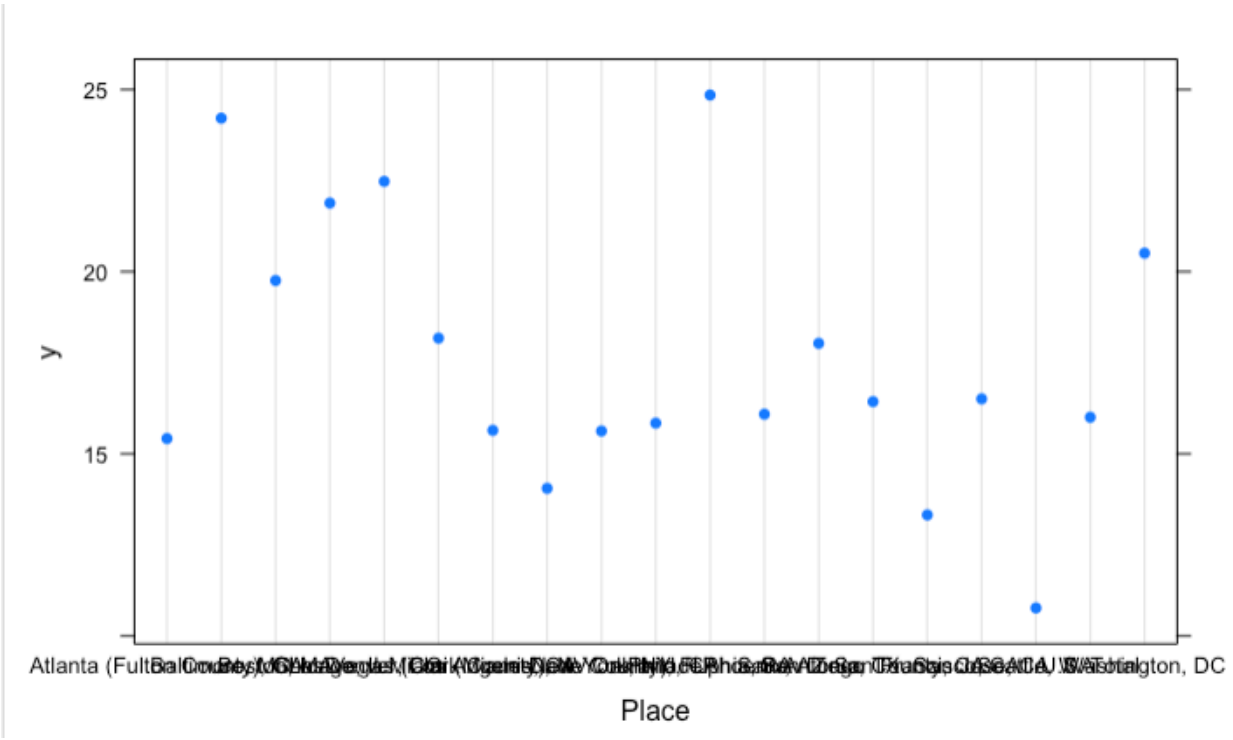
Plot for test data:

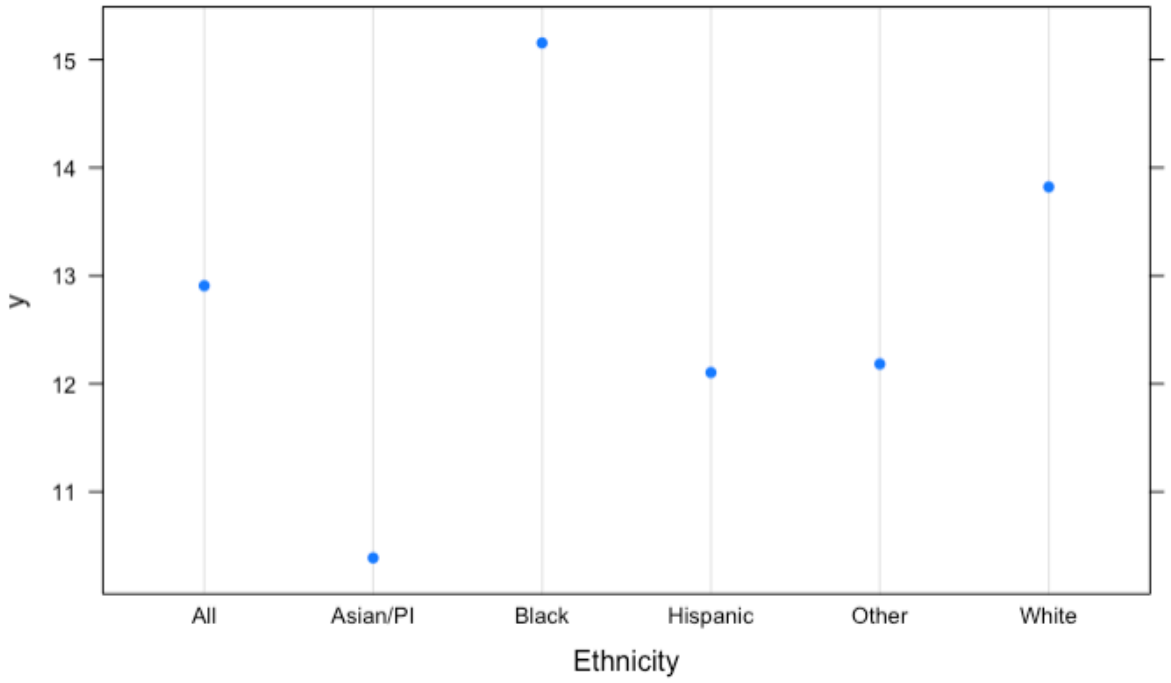


Boosting:



Place and Ethnicity are the most important variables as seen above. We can also produce partial dependence plots for these two variables. The plots below show marginal effect of selected variables on the response





As seen above the most important predictor is Place and the next best predictor is Ethnicity again.
Looking at the plots after Boosting we can see that, the cities **Philadelphia and Denver** seem to have higher rates of smoking among adults, while the cities **Seattle and Miami** have lower smoking rates.
Coming to ethnicities we have Multiracial and Native American communities with higher values; Asian/PI and Hispanic have lower values among the adult population.

9. Overall Conclusion:

Overall, we can conclude that **Place and Ethnicity** are the most important predictors that influence the Indicator values for Smoking and Drinking among High School students as well as Smoking and Drinking among Adults. In other words, certain cities have Values in the higher range, while others exhibit lower indicators.

Also, some communities like Multiracial seem to have more tendency to be in the higher range of indicator values, that is, more adults and high school students in these communities seem to have smoking and binge drinking problems. Whereas, communities such as Asian/PI and Black seem to have lower indicator values.

Therefore, we can place emphasis on the cities and communities which have high tendency for Smoking and Drinking problems and perform further analysis to examine what needs to be done in order to curb these societal problems.

10. Future Research:

Analyze using more predictors:

Performed analysis based on four variables which I felt are the most influential based on my Questions of Interest. All the remaining variables in the Original dataset (Health.Data) have not been considered as they are either not significant logistically or they have not been collected in a format that can be utilized for analysis.
Therefore, future research can be performed by processing these data columns.

Perform analysis on disease indicators:

Original dataset includes diagnosis rate, mortality rate and incidence rate indicators for various diseases such as HIV/AIDS, Cancer,Heart Disease in the most urban cities of the United States. Our intention is to further explore the data and examine the various dependencies and influencing factors that can lead to these conditions. This can be a good case for future research as well.

11. References

[1] https://data.world/health/big-cities-health

[2] http://faculty.marshall.usc.edu/gareth-james/ISL/