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

Aman Negassi, Data Science, aman_negassi@my.uri.edu
Anusha Singamaneni, Computer Science, anushasa10@my.uri.edu
Archana Chittoor, Computer Science, achittoor@my.uri.edu

Contents

Analysis of Big City Smoking and Drinking Data among Students and Adults

4/20/2020

Team:

Aman Negassi, Data Science, aman_negassi@my.uri.edu

Anusha Singamaneni, Computer Science, anushasa10@my.uri.edu

Archana Chittoor, Computer Science, achittoor@my.uri.edu

1. Executive Summary

This project aims to utilize statistical methods such as Linear Regression and Decision Trees, 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 [1] 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 we are considering the demographics of the subjects that are disproportionately scattered among the cities.

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. The intention is to conduct statistical analysis and inference using appropriate methods to isolate any underlying factors and help regulate these societal problems from the core.

2. Descriptive Statistics and Graphical Summaries

Loading Data

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Load the dataset and run summary()
health.data = read.csv("Big_Cities_Health_Data_Inventory.csv")
summary (health.data)
```

```
##                                Indicator.Category
##  HIV/AIDS                                  :2177
##  Injury and Violence                       :1916
##  Nutrition, Physical Activity, & Obesity:1841
##  Infectious Disease                        :1486
##  Cancer                                    :1432
##  Maternal and Child Health                 :1323
##  (Other)                                   :3337
##                                                                                  Indicator
##  Persons Living with HIV/AIDS Rate (Per 100,000 people)                               :  6
23
##  HIV Diagnoses Rate (Per 100,000 people)                                              :  6
00
##  All Types of Cancer Mortality Rate (Age-Adjusted; Per 100,000 people):  5
72
##  Tuberculosis Incidence Rate (Per 100,000 people)                                     :  5
60
##  Heart Disease Mortality Rate (Age-Adjusted; Per 100,000 people)                      :  5
54
##  Diabetes Mortality Rate (Age-Adjusted; Per 100,000 people)                           :  5
38
##  (Other)                                                                              :100
65
##        Year         Gender     Race..Ethnicity     Value
##  2012     :3950   Both  :9409   All     :5757   Min.   :    0.0
##  2013     :3657   Female:2423   White   :1914   1st Qu.:    7.0
##  2011     :3501   Male  :1680   Black   :1869   Median :   16.1
##  2010     :1357                 Hispanic:1688   Mean   :  285.7
```

```
##  2014     :1020                 Asian/PI:1015   3rd Qu.:   45.2
##  2008-2012:   7                 Other   : 570   Max.   :80977.0
##  (Other)  :  20                 (Other) : 699   NA's   :13
##                              Place
##  Phoenix, AZ                    : 772
##  Miami (Miami-Dade County), FL  : 707
##  Long Beach, CA                 : 667
##  New York, NY                   : 651
##  Denver, CO                     : 649
##  Portland (Multnomah County), OR: 623
##  (Other)                        :9443
##
BCHC.Requested.Methodology
##  2012, 2013, 2014: per 100,000 population using 2010 US Census figures. Pl
ease specify population/source/ methodology - this is an area where data sour
ces vary greatly. If 2012-2014 not available, provide three most recent years
of data       :  734
##  HIV cases diagnosed in 2012, 2013, 20014 (as available); report crude rat
e per 100,000 pop using 2010 US Census figures. If 2012-2014 not available, p
rovide three most recent years of data
:  540
##  2012, 2013, 2014; per 100,000 population using 2010 US Census figures, ag
e adjusted to the year 2000 standard population. If 2012-2014 not available,
provide three most recent years of data. Suggested ICD-10 codes: C00-C97
:  516
##
:  508
##  2012, 2013, 2014; per 100,000 population using 2010 US Census figures, ag
e adjusted to the year 2000 standard population. If 2012-2014 not available,
provide three most recent years of data. Suggested ICD-10 codes: I00-I109, I1
1, I13, I20-I51:  502
##  2012, 2013, 2014; per 100,000 population using 2010 US Census figures, ag
e adjusted to the year 2000 standard population. If 2012-2014 not available,
provide three most recent years of data. Suggested ICD-10 Codes: E10-E14
:  482
##  (Other)
:10230
##
Source
##
:2290
##  Source: California Electronic Death Registration System (CA-EDRS), as of
April 1, 2016.  Includes immediate cause of death and contributing causes of
death.   Includes records for which Long Beach was recorded on the death cert
ificate as the decedentâ\200\231s city of residence.  Duplicate records due t
o revisions to the death certificate are not included.  ICD-10 codes were not
included in the dataset; algorithms were written to match key words from caus
e of death to the corresponding ICD-10 diagnosis.: 386
##  Minnesota Department of Health, Vital Records
: 315
```

```
##  NYC DOHMH Bureau of Vital Statistics
: 289
##  BRFSS
: 285
##  Oregon Death Certificates, National Center for Health Statistics Populati
on Estimates, Census Bureau Population Estimates (Vintage 2012)
: 246
##  (Other)
:9701
##
Methods
##
:9280
##  Population denominators based on extrapolation after year 2010
: 233
##  Age-specific rates calculated using annual Los Angeles County population
estimate created by LAC Internal Services Division, adjusted to year 2000 sta
ndard population. Includes records for which Los Angeles was recorded on the
death certificate as the decedentÃ•s city of residence.: 219
##  2011-2013 years are the most recently available data at this time.
: 213
##  2010 US Census; age-adjusted to the year 2000 standard population
: 195
##  age-adjusted total population only, used 2000 standard US population; Pop
ulation denominators: Source: U.S. Census Bureau, 2009-2013 5-Year American C
ommunity Survey
: 193
##  (Other)
:3179
##
Notes
##
:9971
##  Deaths for which cause was listed as â\200œdeferredâ\200\235 for review b
y a coroner or for which cause was missing are not included.  Counts from whi
ch rates are derived are subject to change due to late reporting or revisions
to the cause of death.  : 381
##  Tarrant County (not just Fort Worth)
: 317
##  Bexar County (Not just San Antonio)
: 240
##  2014 data is preliminary
: 219
##  *All races,except for white, contain Hispanic/Latino populations
: 188
##  (Other)
:2196
```

Data Extraction

We first extract 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 us two datasets drink.st and smoke.st. Once the data is extracted, we will retain only the variables of interest and eliminate the others, for further analysis. Additionally, we also rename the column Race..Ethnicity to Ethnicity because it is not recommended to have special symbols in column names. Also, we would need to remove rows with missing values as they generate errors during analysis.

Extracting and Refining Students data

```
require(dplyr)

# Drinking data for High School Students
drink.st <- health.data %>%
  filter(Indicator == "Percent of High School Students Who Binge Drank")

# Smoking data for Students
smoke.st <- health.data %>%
 filter(Indicator == "Percent of High School Students Who Currently Smoke")

# Remove unwanted variables and rename some columns

drink.st <- drink.st[c(3:7)]
drink.st$Ethnicity <- drink.st$Race..Ethnicity
drink.st <- drink.st[-c(3)]

smoke.st <- smoke.st[c(3:7)]
smoke.st$Ethnicity <- smoke.st$Race..Ethnicity
smoke.st <- smoke.st[-c(3)]
```

Extracting and Refining Adults' data

Similarly, we extract 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 us the two datasets drink.ad and smoke.ad. Just like we did for Students data, we eliminate the unwanted columns, which are not required as part of our analysis. Similarly, columns are renamed and missing values are eliminated.

```
# Drinking data for Adults
drink.ad <- health.data %>%
  filter(Indicator == "Percent of Adults Who Binge Drank")

# Smoking data for Adults
smoke.ad <- health.data %>%
  filter(Indicator == "Percent of Adults Who Currently Smoke")
```

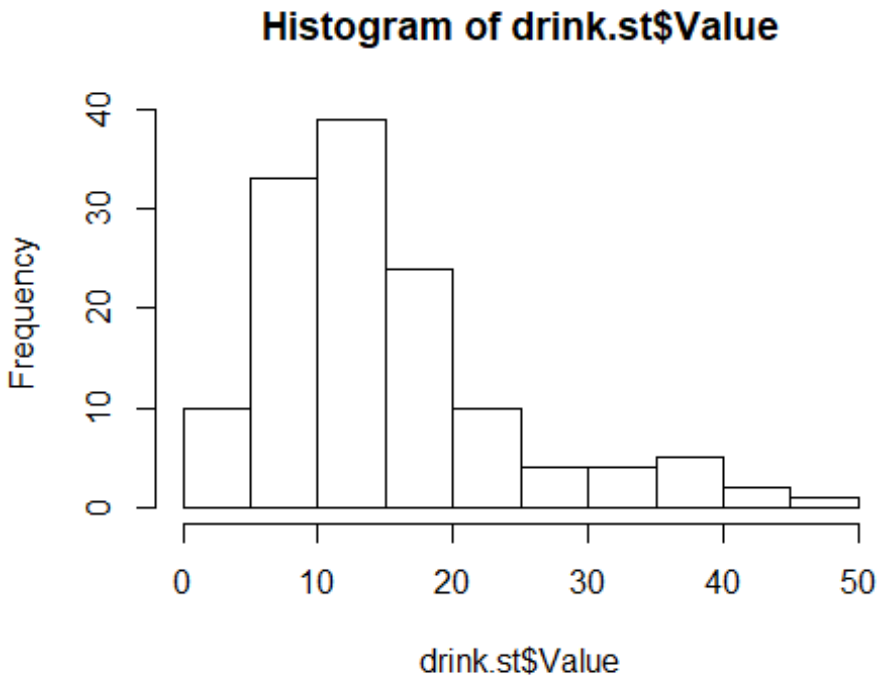
```
# Remove unwanted variables and rename some columns
drink.ad <- drink.ad[c(3:7)]
drink.ad$Ethnicity <- drink.ad$Race..Ethnicity
drink.ad <- drink.ad[-c(3)]

smoke.ad <- smoke.ad[c(3:7)]
smoke.ad$Ethnicity <- smoke.ad$Race..Ethnicity
smoke.ad <- smoke.ad[-c(3)]
```

We found some missing values and these need to eliminted before proceeding with the analysis.

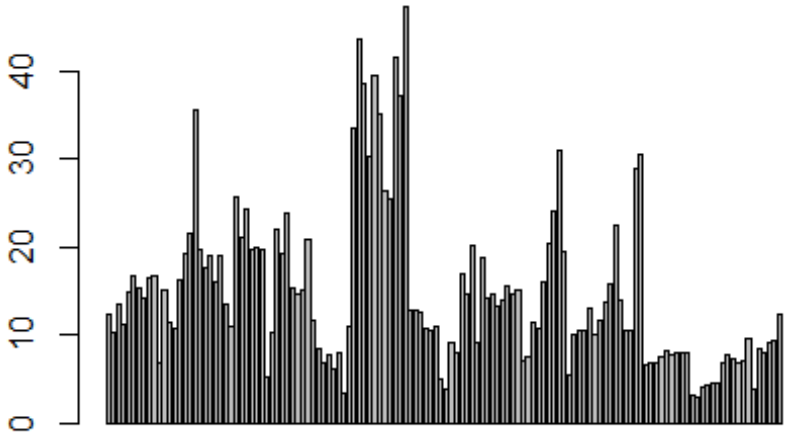
```
# Remove missing data which has been found only in smoke.ad dataset
drink.st <- drink.st %>%
  filter(Value != "NA")
smoke.st <- smoke.st %>%
  filter(Value != "NA")
drink.ad <- drink.ad %>%
  filter(Value != "NA")
smoke.ad <- smoke.ad %>%
  filter(Value != "NA")
```

The above four datasets which have been generated based on our Questions of Interest will be used to perform the analysis.

```
# Drinking data for Students
mean(drink.st$Value) # The average is 14.85%

## [1] 14.85076

median(drink.st$Value) # The median value is 12.7

## [1] 12.7

sd(drink.st$Value) # The standard deviation is 9.09

## [1] 9.094126

hist(drink.st$Value) #Based on the graph, it seems to be often 40 drinks high
school students binge on.
```




```r
barplot(drink.st$Value)
```



```r
#pie(drink.st$Value, cex = 0.5)

# Smoking data for Students

mean(smoke.st$Value) # The average is 10.89 %
```

```
## [1] 10.8906
```

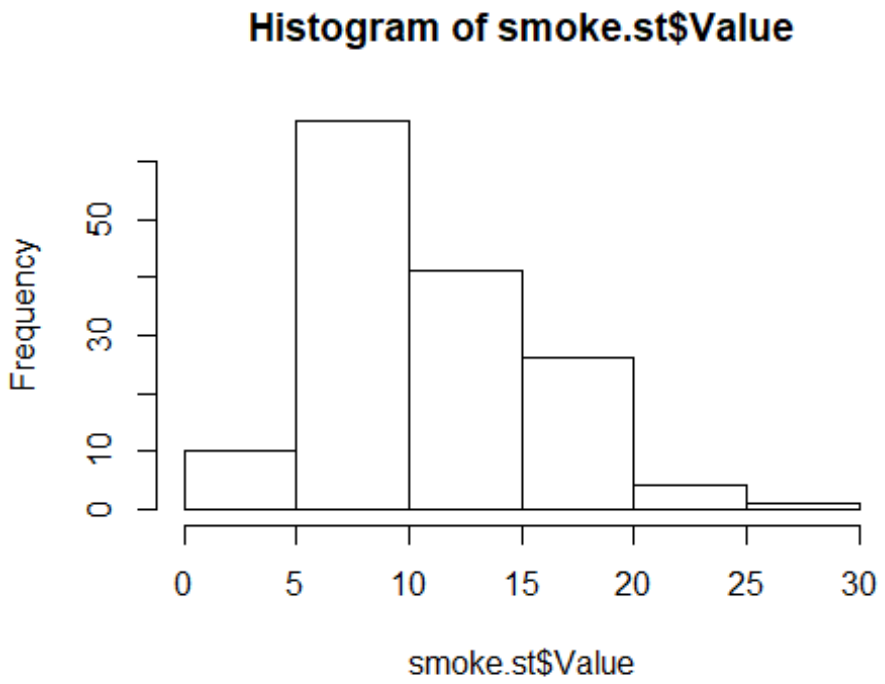
```r
median(smoke.st$Value) # The middle value is 10
```

```
## [1] 10
```

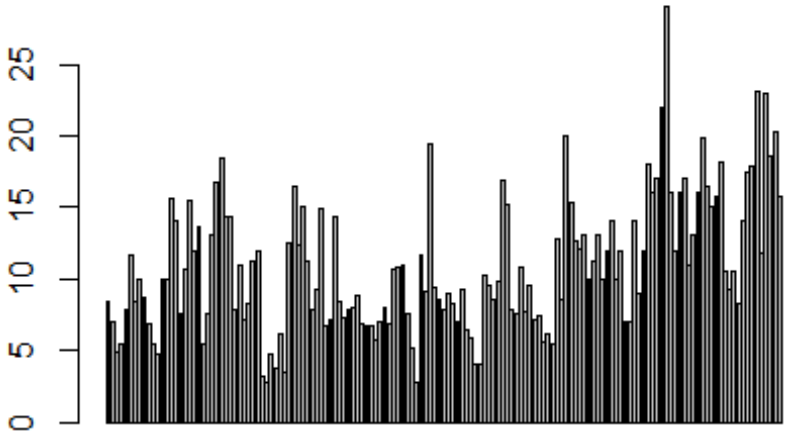
```r
sd(smoke.st$Value) # The standard deviation is 4.69
```

```
## [1] 4.694424
```

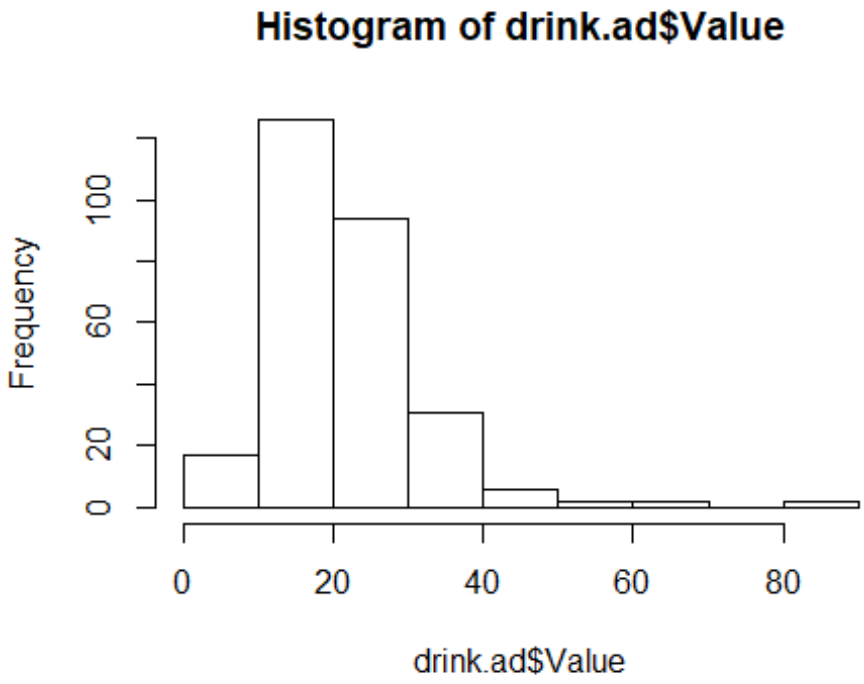
```r
hist(smoke.st$Value) # Based on the histogram, it seems that the students who currently smoke do smoke more than 60 cigarettes often.
```



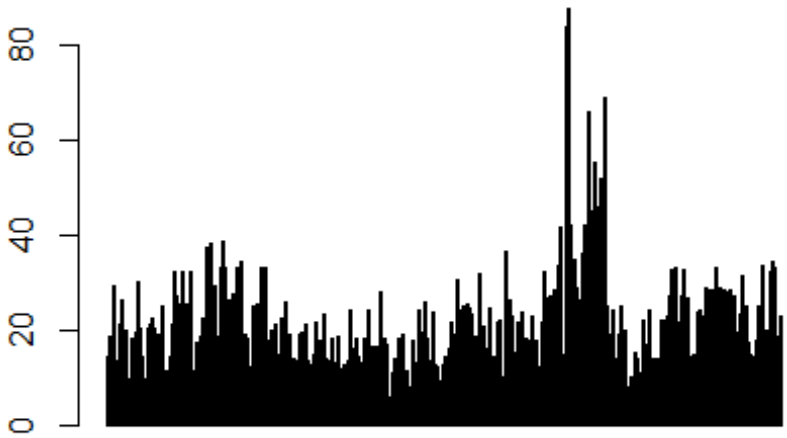

```r
barplot(smoke.st$Value)
```



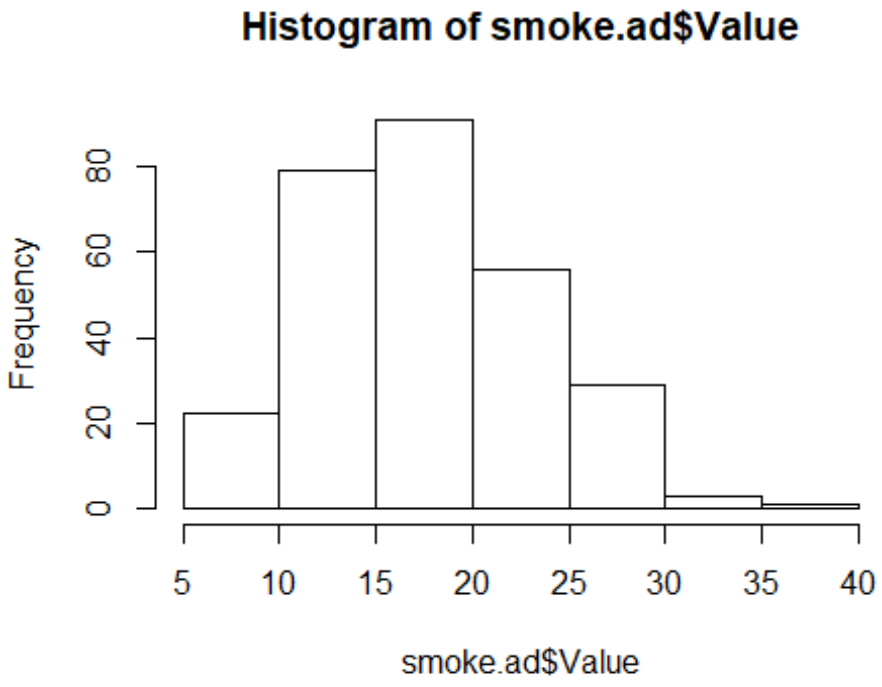

```
#pie(smoke.st$Value, cex = 0.5)

# Drinking data for Adults

mean(drink.ad$Value) #The average was 21.81%

## [1] 21.81393

median(drink.ad$Value) # The middle value is 19.7

## [1] 19.7

sd(drink.ad$Value) # The standard deviation is 10.75

## [1] 10.75459

hist(drink.ad$Value) #Shown from the histogram, the adults frequently consume
more than 120 drinks.
```



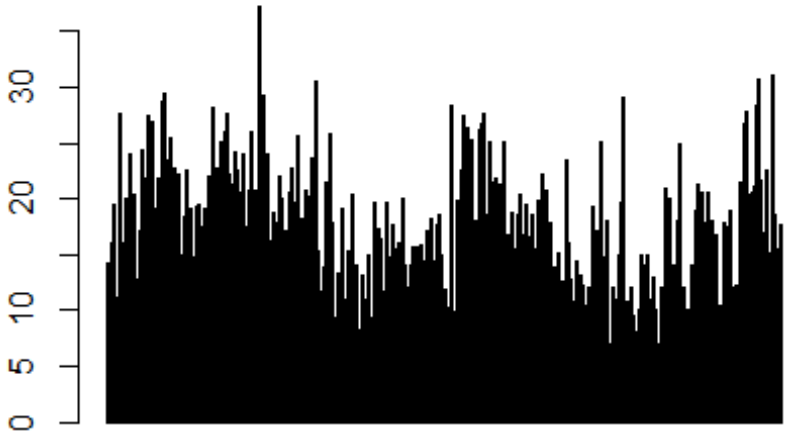

```r
barplot(drink.ad$Value)
```



```r
#pie(drink.ad$Value, cex = 0.5)

# Smoking data for Adults

mean(smoke.ad$Value) # The average was 17.61%

## [1] 17.60854

median(smoke.ad$Value) # The median value is 17.4

## [1] 17.4

sd(smoke.ad$Value) # The standard deviation is 5.53

## [1] 5.533269

hist(smoke.ad$Value) # Adults smoke frequently more than 80 cigarettes.
```




```r
barplot(smoke.ad$Value)
```




```
#pie(smoke.ad$Value, cex = 0.8)
```

Observations:

Looking at the bar plots, pie charts and histograms of our 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.

3. Detailed Data Analysis

i) Linear Regression

The first type of analysis we are performing on our data set is linear Regression. We have chosen the linear Regression as our method of data analysis because , it models the relationships between a response variable and one or more predictor variables. It also shows how changes in the predictor values are associated with changes in the response mean. You can also use regression to make predictions based on the values of the

predictors.Here we are using Response as Value and Gender,Year, Ethnicity, Place as predictors and left all the remaining Variables for future analysis.

Response (variable of interest) with the type and units:

For all four data sets that we work on, **Value** is the response or variable of interest. It is a Numerical (or quantitative) variable. It has no units but rather it is a numerical indicator about a particular health condition that we are interested in.

Explanatory/grouping variable(s) with the type and units:

The explanatory variables for the four datasets (drink.st, smoke.st, drink.ad and smoke.ad) are:

- Year (type: Date)
- Gender (Factor with 3 levels - Male, Female, Both)
- Ethnicity (Factor with 9 levels such as Native American, Asian/PI and so on)
- Place (Factor with 29 levels)

We will not use the other columns like Indicator. Category, Indicator, BCBH.Requested.Methodology, Source , Methods and Notes in the data as they do not add any value and we obtain no new relationships or dependencies when these are taken into account.

How linear Regression applies for our analysis:

To address our question of interest , we need to find relationships/ dependencies between our response **Value** and the predictors given by **Place, Gender, Year and Ethnicity**. As the response is numerical, it makes sense to use linear regression to examine our data.

Limitations :

The major limitation on our dataset is that it is prone to outliers and leverage points. We can see them in almost every dataset (4 datasets). So, outliers and leverage points should be analyzed and removed before applying Linear Regression to the dataset.

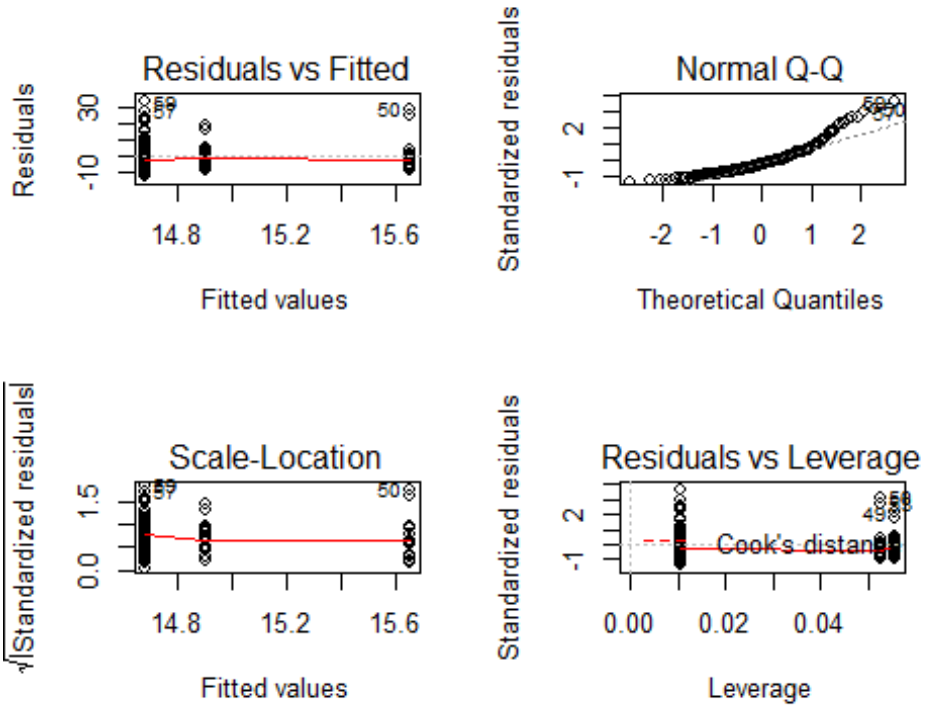
However Linear Regression performs well when the dataset is linearly separable, and is easier to implement, interpret and very efficient to train.

We first fit a simple linear regression model on students data , who currently drink and smoke, between dependent variable “Value” and explanatory variable “Gender”.

```
# Fit linear regression with Gender as predictor and Value as response
# Students data who Drink
lm.fitg=lm(Value~Gender,data=drink.st)
summary(lm.fitg)
```

```
##
## Call:
## lm(formula = Value ~ Gender, data = drink.st)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.781  -6.781  -2.131   4.064  32.619
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.6811     0.9396  15.625   <2e-16 ***
## GenderFemale   0.9716     2.3015   0.422    0.674
## GenderMale     0.2189     2.3542   0.093    0.926
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.158 on 129 degrees of freedom
## Multiple R-squared:  0.001384,   Adjusted R-squared:  -0.0141
## F-statistic: 0.08941 on 2 and 129 DF,  p-value: 0.9145
```

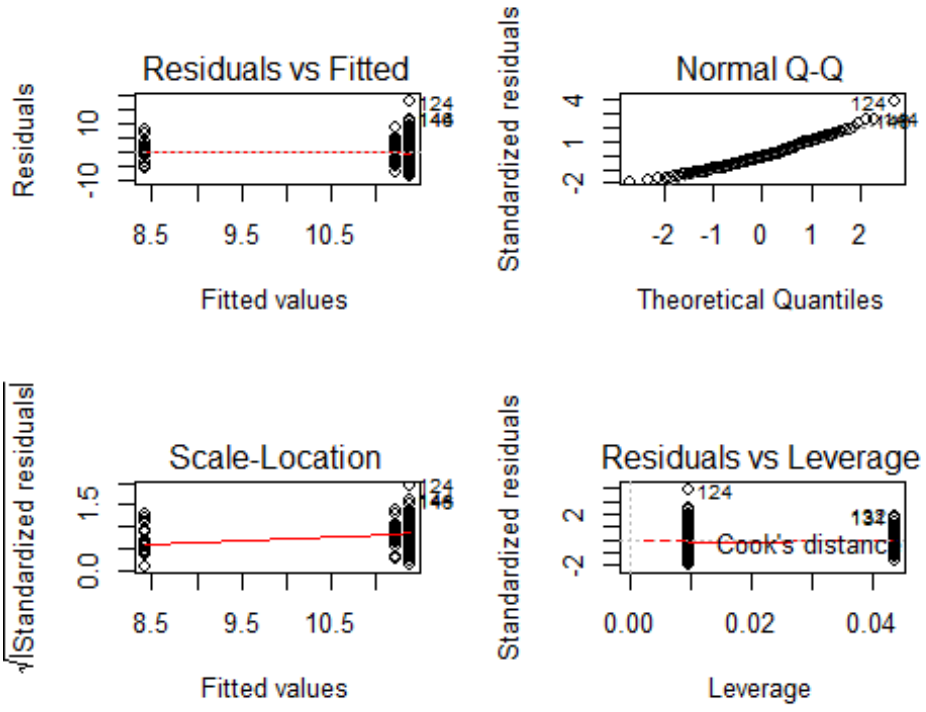
```
par(mfrow = c(2, 2))
plot(lm.fitg)
```




```
# High School students who smoke
lm.fitg=lm(Value~Gender,data=smoke.st)
summary(lm.fitg)
```

```
##
## Call:
## lm(formula = Value ~ Gender, data = smoke.st)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.568 -3.417 -0.568  2.783 17.632
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.3680     0.4537  25.054  < 2e-16 ***
## GenderFemale  -2.9419     1.0620  -2.770  0.00633 **
## GenderMale    -0.1506     1.0620  -0.142  0.88745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.605 on 146 degrees of freedom
## Multiple R-squared:  0.05078,    Adjusted R-squared:  0.03778
## F-statistic: 3.905 on 2 and 146 DF,  p-value: 0.02227
```

```
par(mfrow = c(2, 2))
plot(lm.fitg)
```



Interpretation:

In Simple linear Regression model for students, drink data and smoke data , it shows Gender is not statistically significant , as it results in very high p-Value, and a negative intercept (beta1)

So, there is not much effect of Gender in causing the high school students to smoke and Drink.

Multiple Linear Regression

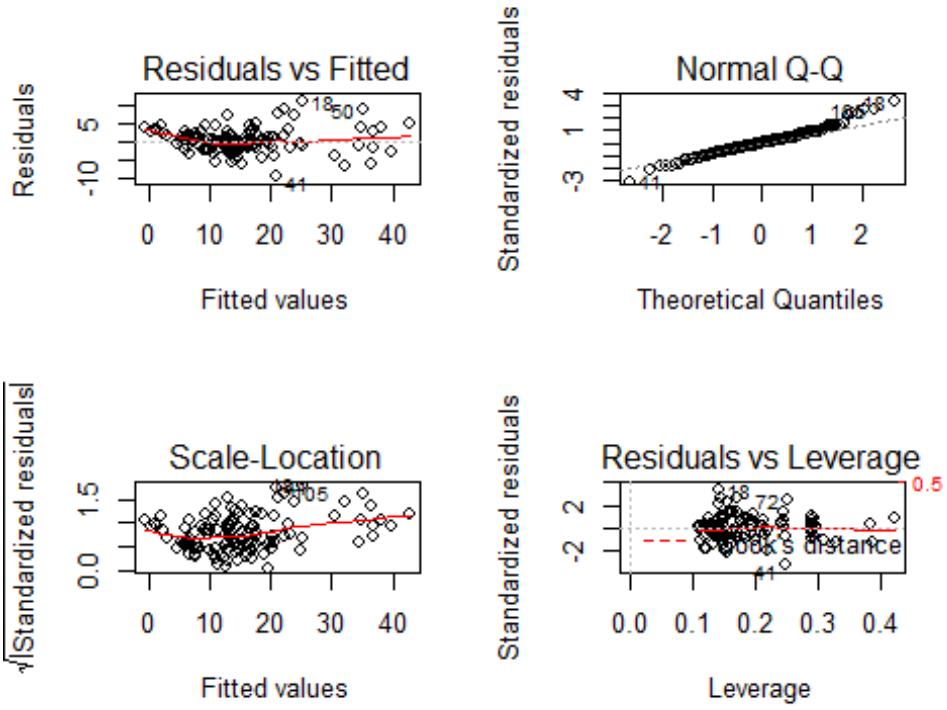
Using Ethnicity , Place and Gender as Independent Variables on High school students drink data.

```
# Drink data

lm.fit=lm(Value~.,data=drink.st)
summary(lm.fit)

##
## Call:
## lm(formula = Value ~ ., data = drink.st)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.383 -1.800  0.000  1.433 10.608
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        15.54680    2.90814   5.346 5.20e-07 **
*
## Year2011                           -0.89720    2.38555  -0.376 0.707596
## Year2012                           -1.78140    2.59098  -0.688 0.493243
## Year2013                           -2.59779    2.24871  -1.155 0.250592
## GenderFemale                        0.16316    1.12364   0.145 0.884825
## GenderMale                         -0.35549    1.14299  -0.311 0.756397
## PlaceBoston, MA                     2.70262    2.02634   1.334 0.185146
## PlaceChicago, IL                    4.15433    2.24871   1.847 0.067474 .
## PlaceDenver, CO                     4.56376    2.15747   2.115 0.036745 *
## PlaceLas Vegas (Clark County), NV   0.39425    2.34649   0.168 0.866890
## PlaceLos Angeles, CA               -8.26230    2.34069  -3.530 0.000617 **
*
## PlaceMiami (Miami-Dade County), FL 22.03981    2.06494  10.673  < 2e-16 **
*
## PlaceNew York, NY                  -1.97374    2.01059  -0.982 0.328498
## PlacePhiladelphia, PA               1.58833    2.02634   0.784 0.434880
## PlaceSan Antonio, TX                6.45099    3.93376   1.640 0.103989
## PlaceSan Diego County, CA          -9.24960    4.01394  -2.304 0.023151 *
## PlaceSan Francisco, CA             -0.15584    2.22111  -0.070 0.944195
## PlaceU.S. Total                    -7.09086    1.98589  -3.571 0.000537 **
```

```
*
## PlaceWashington, DC                -0.64901    3.93376  -0.165 0.869270
## EthnicityAsian/PI                  -6.44344    1.38866  -4.640 1.00e-05 **
*
## EthnicityBlack                     -4.40370    1.16124  -3.792 0.000249 **
*
## EthnicityHispanic                   3.23606    1.14299   2.831 0.005550 **
## EthnicityMultiracial               -0.63869    2.27471  -0.281 0.779426
## EthnicityNative American            0.07476    1.98482   0.038 0.970026
## EthnicityOther                      0.30894    1.81045   0.171 0.864831
## EthnicityWhite                      7.73997    1.18484   6.532 2.31e-09 **
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.463 on 106 degrees of freedom
## Multiple R-squared:  0.8826, Adjusted R-squared:  0.855
## F-statistic: 31.89 on 25 and 106 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))
plot(lm.fit)
```

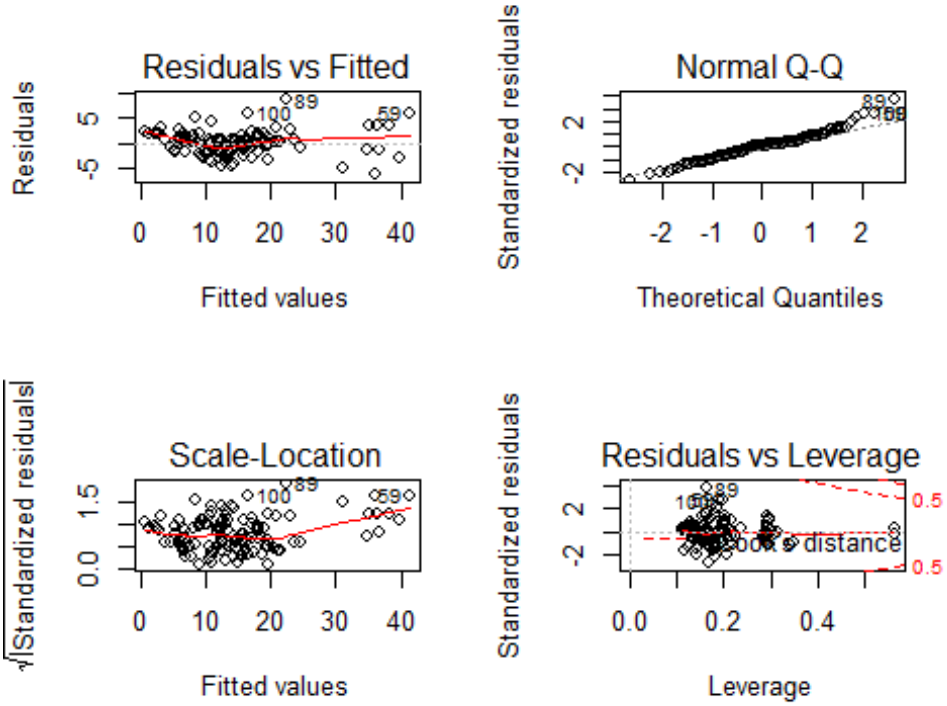
```
## Warning: not plotting observations with leverage one:
##   90, 91, 132

## Warning: not plotting observations with leverage one:
##   90, 91, 132
```



```r
HighLeverage <- cooks.distance(lm.fit) > (4/nrow(drink.st))

drink.st <- drink.st[!HighLeverage,]
lm.fit=lm(Value~.,data=drink.st)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Value ~ ., data = drink.st)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9043 -1.4836  0.1896  1.0421  8.4631
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        15.2801     2.1113   7.237 1.08e-10 **
*
## Year2011                           -0.8581     1.7370  -0.494 0.622401
## Year2012                           -1.8875     1.9026  -0.992 0.323647
## Year2013                           -2.3400     1.6323  -1.434 0.154913
## GenderFemale                       -0.3012     0.8275  -0.364 0.716672
## GenderMale                         -0.3560     0.8290  -0.429 0.668582
## PlaceBoston, MA                     1.9829     1.4784   1.341 0.182982
## PlaceChicago, IL                    4.4121     1.6323   2.703 0.008115 **
## PlaceDenver, CO                     5.3661     1.6081   3.337 0.001202 **
## PlaceLas Vegas (Clark County), NV   2.8653     1.7896   1.601 0.112605
## PlaceLos Angeles, CA               -7.5303     1.7146  -4.392 2.87e-05 **
*
## PlaceMiami (Miami-Dade County), FL 22.1383     1.5467  14.313  < 2e-16 **
*
## PlaceNew York, NY                  -2.1178     1.4641  -1.446 0.151265
## PlacePhiladelphia, PA               1.8031     1.4747   1.223 0.224405
## PlaceSan Francisco, CA             -1.1380     1.6581  -0.686 0.494151
## PlaceU.S. Total                    -6.6808     1.4498  -4.608 1.24e-05 **
*
## EthnicityAsian/PI                  -5.6261     1.0496  -5.360 5.62e-07 **
*
## EthnicityBlack                     -3.9031     0.8584  -4.547 1.57e-05 **
*
## EthnicityHispanic                   3.2356     0.8290   3.903 0.000176 **
*
## EthnicityMultiracial                1.2696     1.9714   0.644 0.521110
## EthnicityNative American            0.2582     1.4422   0.179 0.858272
## EthnicityOther                     -1.7286     1.4381  -1.202 0.232288
## EthnicityWhite                      6.2118     0.9443   6.578 2.42e-09 **
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.512 on 97 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.9297, Adjusted R-squared:  0.9138
## F-statistic: 58.31 on 22 and 97 DF,  p-value: < 2.2e-16

par(mfrow = c(2, 2))
plot(lm.fit)
```



Interpretation

After performing a multiple fit , we can assume that **Ethnicity and Place** are statistically significant predictors for Value. These are the most influencing fators that causes drinking among the High school students

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.

Performing Multiple regression on High school students Smoke data.

```
#Smoke Data
```

```r
lm.fitall=lm(Value~.,data=smoke.st)
summary(lm.fitall)
```

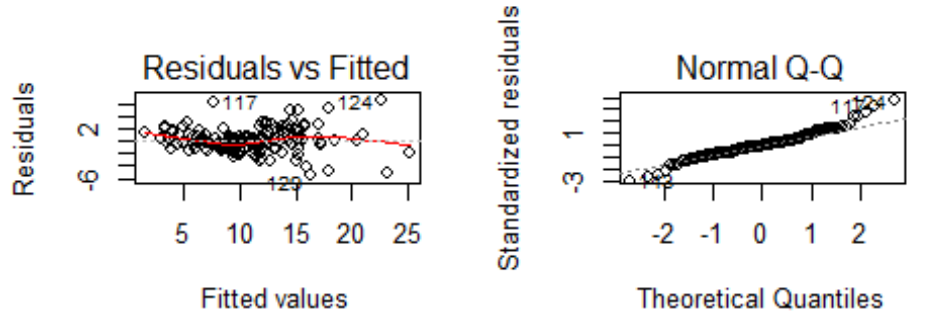
```
##
## Call:
## lm(formula = Value ~ ., data = smoke.st)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3710 -1.2003 -0.1258  0.8056  6.3699
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          9.9320     1.7008   5.840 4.43e-08 **
*
## Year2011                            -0.4933     1.3553  -0.364 0.716498
## Year2012                            -1.5455     0.8978  -1.721 0.087727 .
## Year2013                            -2.4068     1.2786  -1.882 0.062171 .
## GenderFemale                        -1.4739     0.6209  -2.374 0.019167 *
## GenderMale                           1.3174     0.6209   2.122 0.035884 *
## PlaceBoston, MA                      0.7610     1.2519   0.608 0.544373
## PlaceChicago, IL                     3.7459     1.3667   2.741 0.007049 **
## PlaceDenver, CO                      3.0231     1.3113   2.305 0.022835 *
## PlaceDetroit, MI                    -4.4298     1.3849  -3.199 0.001760 **
## PlaceHouston, TX                     3.1689     1.2592   2.517 0.013145 *
## PlaceLas Vegas (Clark County), NV   -0.5888     1.4259  -0.413 0.680386
## PlaceLos Angeles, CA                -0.4250     1.4286  -0.297 0.766618
## PlaceMiami (Miami-Dade County), FL   0.6058     1.2445   0.487 0.627275
## PlaceNew York, NY                    0.3023     1.2155   0.249 0.804019
## PlacePhiladelphia, PA                1.2442     1.2274   1.014 0.312739
## PlaceSan Antonio, TX                 3.1312     1.4259   2.196 0.029986 *
## PlaceSeattle, WA                     3.4218     1.8029   1.898 0.060060 .
## PlaceU.S. Total                      6.3844     1.2094   5.279 5.74e-07 **
*
## PlaceWashington, DC                  8.1747     2.3855   3.427 0.000833 **
*
## EthnicityAsian/PI                   -3.7565     0.8273  -4.540 1.33e-05 **
*
## EthnicityBlack                      -4.1446     0.6725  -6.163 9.52e-09 **
*
## EthnicityHispanic                    1.3115     0.6505   2.016 0.045998 *
## EthnicityMultiracial                 3.8965     1.3578   2.870 0.004845 **
## EthnicityNative American             9.2762     1.1975   7.746 3.13e-12 **
*
## EthnicityOther                       2.0593     0.9190   2.241 0.026847 *
## EthnicityWhite                       4.5626     0.6970   6.546 1.47e-09 **
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.106 on 122 degrees of freedom
## Multiple R-squared:  0.8342, Adjusted R-squared:  0.7988
## F-statistic:  23.6 on 26 and 122 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(lm.fitall)
```

```
## Warning: not plotting observations with leverage one:
##   149
```

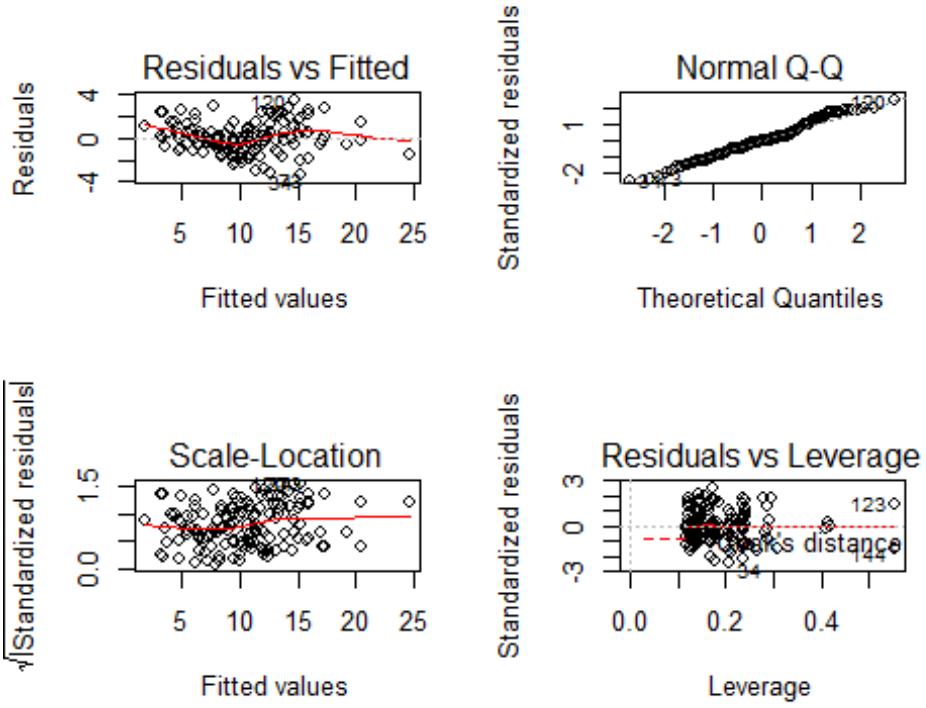
```
## Warning: not plotting observations with leverage one:
##   149
```




```
HighLeverage <- cooks.distance(lm.fitall) > (4/nrow(smoke.st))

smoke.st <- smoke.st[!HighLeverage,]
lm.fitall=lm(Value~.,data=smoke.st)
summary(lm.fitall)
```

```
##
## Call:
## lm(formula = Value ~ ., data = smoke.st)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2960 -0.8716 -0.0477  0.8083  3.3721
##
```

```
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         10.0944     1.2393   8.145 5.90e-13 **
*
## Year2011                            -1.1064     0.9896  -1.118 0.265976
## Year2012                            -1.4667     0.7348  -1.996 0.048360 *
## Year2013                            -2.4265     0.9316  -2.605 0.010446 *
## GenderFemale                        -1.6410     0.4584  -3.580 0.000510 **
*
## GenderMale                           1.3174     0.4520   2.915 0.004302 **
## PlaceBoston, MA                      0.9889     0.9130   1.083 0.281078
## PlaceChicago, IL                     3.6951     0.9958   3.711 0.000323 **
*
## PlaceDenver, CO                      2.9589     0.9553   3.097 0.002467 **
## PlaceDetroit, MI                    -4.2202     1.0095  -4.181 5.79e-05 **
*
## PlaceHouston, TX                     3.4507     0.9188   3.756 0.000276 **
*
## PlaceLas Vegas (Clark County), NV   -1.4707     1.1007  -1.336 0.184224
## PlaceLos Angeles, CA                -0.5254     1.0404  -0.505 0.614577
## PlaceMiami (Miami-Dade County), FL   0.3980     0.9130   0.436 0.663734
## PlaceNew York, NY                    0.5553     0.8884   0.625 0.533171
## PlacePhiladelphia, PA                1.1038     0.8990   1.228 0.222080
## PlaceSan Antonio, TX                 3.0181     1.0394   2.904 0.004444 **
## PlaceSeattle, WA                     3.2093     1.3318   2.410 0.017595 *
## PlaceU.S. Total                      6.9701     0.8942   7.795 3.59e-12 **
*
## EthnicityAsian/PI                   -3.8137     0.6033  -6.322 5.43e-09 **
*
## EthnicityBlack                      -4.5483     0.4986  -9.122 3.50e-15 **
*
## EthnicityHispanic                    1.3242     0.4736   2.796 0.006090 **
## EthnicityMultiracial                 3.8775     0.9967   3.890 0.000170 **
*
## EthnicityNative American             8.6524     1.1830   7.314 4.15e-11 **
*
## EthnicityOther                       1.8753     0.7648   2.452 0.015745 *
## EthnicityWhite                       4.5692     0.5579   8.190 4.67e-13 **
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.533 on 112 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8951, Adjusted R-squared:  0.8717
## F-statistic: 38.23 on 25 and 112 DF,  p-value: < 2.2e-16

par(mfrow = c(2, 2))
plot(lm.fitall)
```



Interpretation

Multiple fit for students smoke data shows that Place and Ethnicity are most influencing factors that caused smoking among High school students, Approximately 89% of variation in value can be explained by this model, and has a very low P-value, And the residual error is 3.34 on 238 degrees of freedom tell us that there is no bigger distance between observed value(Y) and predicted value(Yhat), which shows this is statistically significant. It can be seen from the plots that there are few leverage points , which doesnt seem to have any effect on our model , so removed them and ran the fit again.

Multiple Regression analysis on Adults drinking data

```
# Similarly run the fit on Adults data
lm.fitAd=lm(Value~Gender,data=drink.ad)
summary(lm.fitAd)

##
## Call:
## lm(formula = Value ~ Gender, data = drink.ad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.648  -6.861  -2.125   4.459  65.752
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.7475     0.7303  29.779  < 2e-16 ***
```

```
## GenderFemale  -5.7449     1.8079  -3.178 0.001653 **
## GenderMale     5.9184     1.7706   3.343 0.000944 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.33 on 277 degrees of freedom
## Multiple R-squared:  0.08435,    Adjusted R-squared:  0.07774
## F-statistic: 12.76 on 2 and 277 DF,  p-value: 5.004e-06
```

```
lm.fitAd=lm(Value~.,data=drink.ad)
summary(lm.fitAd)
```

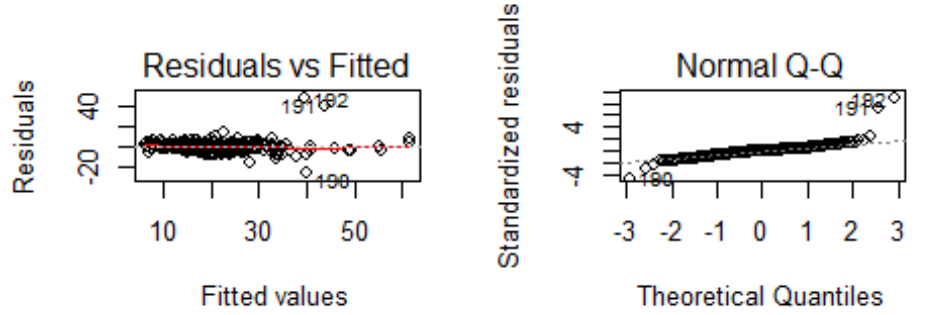
```
##
## Call:
## lm(formula = Value ~ ., data = drink.ad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.189  -2.395   0.044   1.926  48.000
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        16.15084    2.24636   7.190 7.76e-12
***
## Year2011                            4.11968    1.55203   2.654  0.00846
**
## Year2012                            2.31195    1.57621   1.467  0.14371
## Year2013                            0.58867    1.50493   0.391  0.69602
## Year2015                            0.01968    8.45297   0.002  0.99814
## GenderFemale                       -5.76996    1.31610  -4.384 1.73e-05
***
## GenderMale                          6.21951    1.29769   4.793 2.85e-06
***
## PlaceBaltimore, MD                  0.34289    2.20331   0.156  0.87646
## PlaceBoston, MA                     7.57666    2.36183   3.208  0.00151
**
## PlaceChicago, IL                    9.19518    2.90157   3.169  0.00172
**
## PlaceDenver, CO                     6.16921    2.06177   2.992  0.00305
**
## PlaceFort Worth (Tarrant County), TX  -0.72805    3.84009  -0.190  0.84978
## PlaceHouston, TX                    -2.51540    2.07516  -1.212  0.22662
## PlaceLas Vegas (Clark County), NV   -3.69746    2.06177  -1.793  0.07415
.
## PlaceLong Beach, CA                 2.52948    6.15161   0.411  0.68129
## PlaceLos Angeles, CA                -3.37281    2.76788  -1.219  0.22418
## PlaceMiami (Miami-Dade County), FL  -2.75447    2.41636  -1.140  0.25543
## PlaceMinneapolis, MN               -11.71666    4.50880  -2.599  0.00993
**
## PlaceNew York, NY                   1.30812    1.96851   0.665  0.50698
```

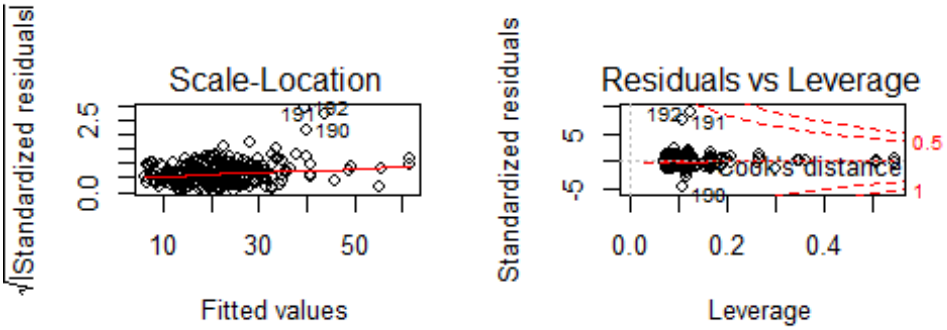
```
## PlacePhiladelphia, PA                  2.18210    2.12160   1.029  0.30472
## PlacePhoenix, AZ                      -1.66037    2.43091  -0.683  0.49523
## PlaceSan Antonio, TX                  17.11350    2.31504   7.392 2.25e-12
***
## PlaceSan Diego County, CA             14.37708    2.70137   5.322 2.31e-07
***
## PlaceSan Jose, CA                     38.45820    2.76937  13.887  < 2e-16
***
## PlaceSeattle, WA                      -1.02443    1.95056  -0.525  0.59992
## PlaceU.S. Total                        8.27923    1.97998   4.181 4.03e-05
***
## PlaceWashington, DC                    5.08264    2.02258   2.513  0.01261
*
## EthnicityAsian/PI                     -9.38972    1.94279  -4.833 2.37e-06
***
## EthnicityBlack                        -6.14034    1.31015  -4.687 4.60e-06
***
## EthnicityHispanic                      0.46047    1.34593   0.342  0.73255
## EthnicityMultiracial                   6.10315    4.43521   1.376  0.17005
## EthnicityNative American               1.18585    3.22429   0.368  0.71335
## EthnicityOther                        -0.60851    1.93686  -0.314  0.75366
## EthnicityWhite                         6.23608    1.28495   4.853 2.16e-06
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.876 on 246 degrees of freedom
## Multiple R-squared:  0.7368, Adjusted R-squared:  0.7015
## F-statistic: 20.87 on 33 and 246 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))
plot(lm.fitAd)
```

```
## Warning: not plotting observations with leverage one:
##   279, 280
```

```
## Warning: not plotting observations with leverage one:
##   279, 280
```



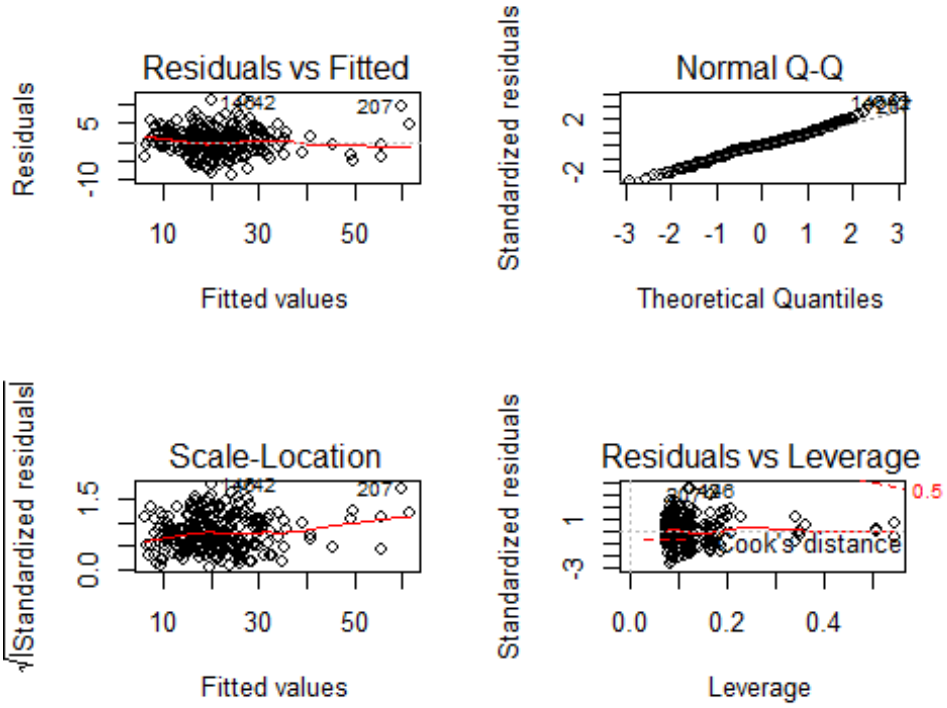



```
HighLeverage <- cooks.distance(lm.fitAd) > (4/nrow(drink.ad))

drink.ad <- drink.ad[!HighLeverage,]
lm.fitAd=lm(Value~.,data=drink.ad)
summary(lm.fitAd)

##
## Call:
## lm(formula = Value ~ ., data = drink.ad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5009 -1.8323 -0.0566  1.8601 10.5922
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         14.0635     1.3004  10.814  < 2e-16
***
## Year2011                             5.9869     0.9201   6.507 4.49e-10
***
## Year2012                             4.6477     0.9239   5.030 9.64e-07
***
## Year2013                             3.7807     0.8954   4.223 3.44e-05
***
## GenderFemale                        -5.6355     0.7577  -7.437 1.85e-12
***
## GenderMale                           6.0287     0.7430   8.114 2.62e-14
```

```
***
## PlaceBaltimore, MD                      1.4069     1.2556   1.121  0.26362
## PlaceBoston, MA                         8.2969     1.3468   6.160 3.07e-09
***
## PlaceChicago, IL                        9.7556     1.6512   5.908 1.19e-08
***
## PlaceDenver, CO                         6.1318     1.1722   5.231 3.70e-07
***
## PlaceFort Worth (Tarrant County), TX   -0.0417     2.1853  -0.019  0.98479
## PlaceHouston, TX                       -2.4514     1.1804  -2.077  0.03891
*
## PlaceLas Vegas (Clark County), NV      -3.7348     1.1722  -3.186  0.00163
**
## PlaceLos Angeles, CA                   -2.7911     1.5754  -1.772  0.07771
.
## PlaceMiami (Miami-Dade County), FL     -1.9229     1.3783  -1.395  0.16430
## PlaceMinneapolis, MN                  -11.7308     2.5640  -4.575 7.66e-06
***
## PlaceNew York, NY                       1.3441     1.1200   1.200  0.23130
## PlacePhiladelphia, PA                   2.4479     1.2258   1.997  0.04697
*
## PlacePhoenix, AZ                       -1.9183     1.3821  -1.388  0.16645
## PlaceSan Antonio, TX                   11.4782     1.5921   7.210 7.41e-12
***
## PlaceSan Diego County, CA              14.9406     1.5373   9.719  < 2e-16
***
## PlaceSan Jose, CA                      37.7150     1.5750  23.946  < 2e-16
***
## PlaceSeattle, WA                       -0.7152     1.1158  -0.641  0.52216
## PlaceU.S. Total                         7.9566     1.1329   7.023 2.26e-11
***
## PlaceWashington, DC                     5.0826     1.1504   4.418 1.51e-05
***
## EthnicityAsian/PI                      -9.8788     1.1079  -8.917  < 2e-16
***
## EthnicityBlack                         -6.0888     0.7516  -8.101 2.84e-14
***
## EthnicityHispanic                       0.3071     0.7814   0.393  0.69462
## EthnicityMultiracial                    5.8869     2.5235   2.333  0.02049
*
## EthnicityNative American                3.2748     2.1172   1.547  0.12326
## EthnicityOther                         -1.3867     1.1398  -1.217  0.22495
## EthnicityWhite                          4.3525     0.7468   5.828 1.81e-08
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.34 on 238 degrees of freedom
##   (2 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.8857, Adjusted R-squared:  0.8708
## F-statistic:  59.5 on 31 and 238 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(lm.fitAd)
```



Interpretation

From the simple regression fit, we can understand 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 standared 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-vaue 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.

Multiple Regression analysis on Adults smoking data

```
#Adults who smoke
lm.fitall=lm(Value~.,data=smoke.ad)
summary(lm.fitall)
```

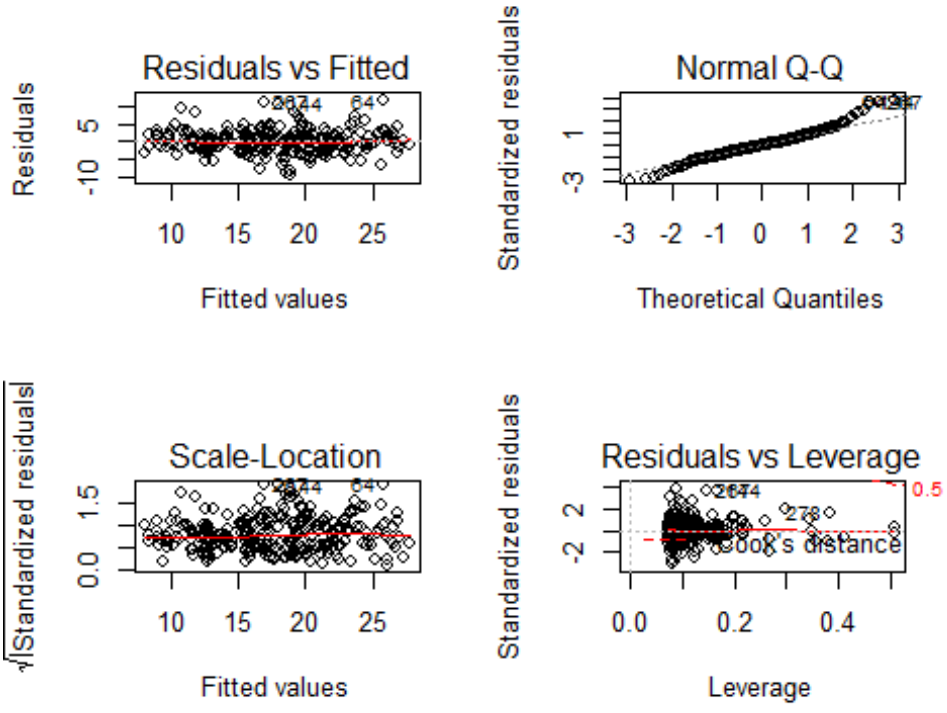
```
##
## Call:
## lm(formula = Value ~ ., data = smoke.ad)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -9.215 -1.690  0.000  1.555 11.422
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       15.70619    1.20666  13.016  < 2e-16
***
## Year2011                           0.30346    0.82360   0.368 0.712851
## Year2012                           0.35060    0.80645   0.435 0.664131
## Year2013                           0.07100    0.81618   0.087 0.930749
## Year2014                          -1.29344    1.40266  -0.922 0.357367
## Year2015                          -1.89654    4.67939  -0.405 0.685613
## GenderFemale                      -3.01087    0.69906  -4.307 2.40e-05
***
## GenderMale                         3.15890    0.69906   4.519 9.68e-06
***
## PlaceBaltimore, MD                 6.53033    1.22007   5.352 1.99e-07
***
## PlaceBoston, MA                    2.85110    1.30155   2.191 0.029427
*
## PlaceChicago, IL                   6.38353    1.34923   4.731 3.77e-06
***
## PlaceDenver, CO                    5.62516    1.13110   4.973 1.24e-06
***
## PlaceFort Worth (Tarrant County), TX  1.27579    2.12517   0.600 0.548845
## PlaceLas Vegas (Clark County), NV     4.14129    1.14331   3.622 0.000355
***
## PlaceLong Beach, CA                1.69035    3.41043   0.496 0.620592
## PlaceLos Angeles, CA              -1.48487    1.53793  -0.965 0.335248
## PlaceMiami (Miami-Dade County), FL   -3.59697    1.33122  -2.702 0.007374
**
## PlaceMinneapolis, MN               0.81678    2.49803   0.327 0.743970
## PlaceNew York, NY                 -0.03313    1.09162  -0.030 0.975810
## PlaceOakland, CA                  -2.48487    1.53793  -1.616 0.107441
## PlacePhiladelphia, PA              7.31987    1.21767   6.011 6.63e-09
***
## PlacePhoenix, AZ                   1.73320    1.28524   1.349 0.178731
## PlaceSan Antonio, TX              -1.32841    1.72518  -0.770 0.442035
## PlaceSan Diego County, CA         -2.55402    1.56555  -1.631 0.104093
## PlaceSan Francisco, CA            -3.38466    1.41295  -2.395 0.017351
*
## PlaceSan Jose, CA                 -2.30235    1.62479  -1.417 0.157749
## PlaceSeattle, WA                  -3.16890    1.08002  -2.934 0.003662
**
## PlaceU.S. Total                    1.21547    1.09686   1.108 0.268890
## PlaceWashington, DC                3.26808    1.12300   2.910 0.003945
**
## EthnicityAsian/PI                 -4.60514    0.98653  -4.668 5.01e-06
***
## EthnicityBlack                     4.37555    0.74730   5.855 1.52e-08
```

```
***
## EthnicityHispanic                    -1.92895    0.75434  -2.557 0.011157
*
## EthnicityMultiracial                  7.55958    2.01507   3.752 0.000219
***
## EthnicityNative American              6.91730    1.78396   3.877 0.000136
***
## EthnicityOther                       -0.14446    1.13658  -0.127 0.898964
## EthnicityWhite                       -0.46296    0.71117  -0.651 0.515669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.257 on 245 degrees of freedom
## Multiple R-squared:  0.6968, Adjusted R-squared:  0.6535
## F-statistic: 16.09 on 35 and 245 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(lm.fitall)
```

```
## Warning: not plotting observations with leverage one:
##   280, 281

## Warning: not plotting observations with leverage one:
##   280, 281
```

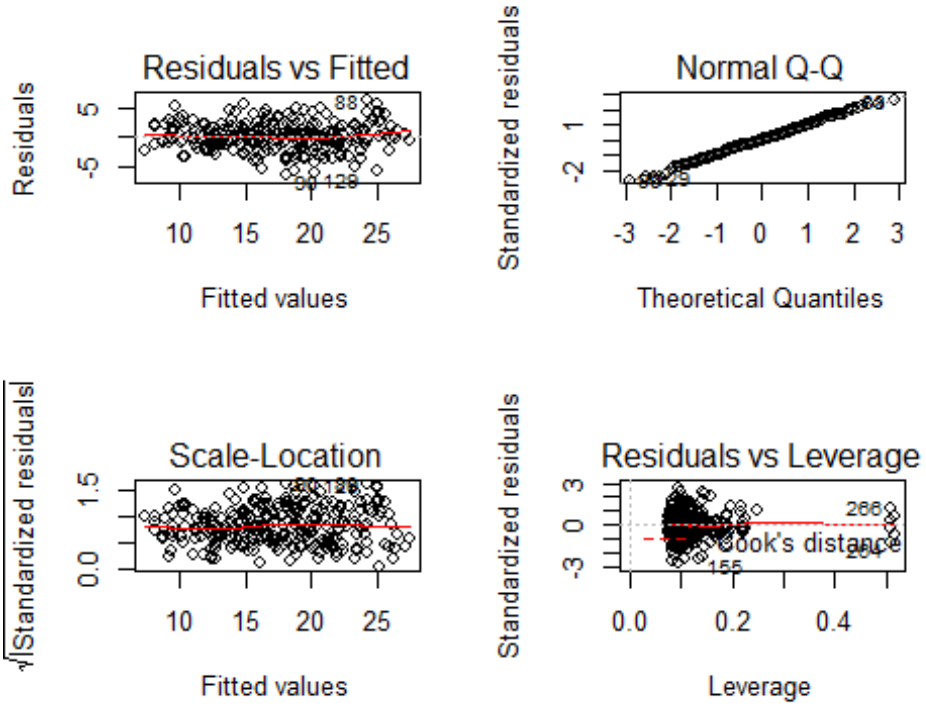



```
HighLeverage <- cooks.distance(lm.fitall) > (4/nrow(smoke.ad))
```

```r
smoke.ad <- smoke.ad[!HighLeverage,]
lm.fitall=lm(Value~.,data=smoke.ad)
summary(lm.fitall)
```

```
##
## Call:
## lm(formula = Value ~ ., data = smoke.ad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3962 -1.3926 -0.0606  1.4914  6.1759
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         15.3375     0.9578  16.012  < 2e-16
***
## Year2011                             0.2854     0.6501   0.439 0.661082
## Year2012                             0.2640     0.6375   0.414 0.679164
## Year2013                            -0.4789     0.6491  -0.738 0.461412
## Year2014                            -1.5667     1.0862  -1.442 0.150567
## GenderFemale                        -3.1985     0.5425  -5.896 1.33e-08
***
## GenderMale                           2.9458     0.5424   5.431 1.43e-07
***
## PlaceBaltimore, MD                   6.4202     0.9786   6.561 3.55e-10
***
## PlaceBoston, MA                      3.5872     1.0200   3.517 0.000527
***
## PlaceChicago, IL                     6.8186     1.0566   6.453 6.49e-10
***
## PlaceDenver, CO                      5.6747     0.9016   6.294 1.57e-09
***
## PlaceFort Worth (Tarrant County), TX   0.3198     1.9558   0.164 0.870258
## PlaceLas Vegas (Clark County), NV    4.8124     0.8991   5.353 2.11e-07
***
## PlaceLos Angeles, CA                -1.0379     1.1990  -0.866 0.387604
## PlaceMiami (Miami-Dade County), FL  -2.9116     1.1097  -2.624 0.009282
**
## PlaceMinneapolis, MN                 1.2378     1.9348   0.640 0.522982
## PlaceNew York, NY                    0.7631     0.8608   0.886 0.376293
## PlaceOakland, CA                    -2.9845     1.3513  -2.209 0.028195
*
## PlacePhiladelphia, PA                7.9506     0.9613   8.271 1.11e-14
***
## PlacePhoenix, AZ                     2.4089     1.0101   2.385 0.017912
*
## PlaceSan Antonio, TX                -0.4061     1.3423  -0.303 0.762488
## PlaceSan Diego County, CA           -1.8922     1.2649  -1.496 0.136060
## PlaceSan Francisco, CA              -2.8475     1.1092  -2.567 0.010893
*
```

```
## PlaceSan Jose, CA                       -1.3946     1.2654  -1.102 0.271581
## PlaceSeattle, WA                        -2.9237     0.8672  -3.372 0.000878
***
## PlaceU.S. Total                          1.5907     0.8762   1.815 0.070785
.
## PlaceWashington, DC                      5.5315     0.9293   5.952 9.89e-09
***
## EthnicityAsian/PI                       -4.5215     0.7620  -5.934 1.09e-08
***
## EthnicityBlack                           3.8888     0.5926   6.562 3.53e-10
***
## EthnicityHispanic                       -2.3177     0.5870  -3.949 0.000105
***
## EthnicityMultiracial                     5.5693     2.7365   2.035 0.042988
*
## EthnicityNative American                 7.2686     1.8985   3.829 0.000167
***
## EthnicityOther                          -1.3976     0.9165  -1.525 0.128649
## EthnicityWhite                           0.3078     0.5708   0.539 0.590163
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.511 on 228 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.7995, Adjusted R-squared:  0.7705
## F-statistic: 27.55 on 33 and 228 DF,  p-value: < 2.2e-16

par(mfrow = c(2, 2))
plot(lm.fitall)

## Warning: not plotting observations with leverage one:
##   195

## Warning: not plotting observations with leverage one:
##   195
```



Interpretation

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 High school students.

ii) Regression Tree

The second type of analysis we perform on our data is Decision Tree analysis. We have chosen the Regression tree as our method of data analysis because our response Value is numerical (quantitative).

The Regression tree has advantages over other Regression models. It is easier to interpret and has a good graphical representation. Our intention is to investigate the relationship between the response Value and the predictors Gender, Ethnicity, Value and Place by using

Decision Trees. We will also implement Bagging, Boosting and Random Forests, selecting the best method that produces the minimum Mean Squared Error (MSE). In addition, we will compare and analyse the importance of each of our predictor variables in relationship to the Value indicator.

Response (variable of interest) with the type and units:

For all four data sets that we work on, **Value** is the response or variable of interest. It is a Numerical (or quantitative) variable. It has no units but rather a numerical indicator about a particular health condition that we are interested in.

Explanatory/grouping variable(s) with the type and units:

The explanatory variables for the four datasets (drink.st, smoke.st, drink.ad and smoke.ad) are:
* Year (type: Date) * Gender (Factor with 3 levels - Male, Female, Both) * Ethnicity (Factor with 9 levels such as Native American, Asian/PI and so on) * Place (Factor with 29 levels)

We will not use the other columns like Indicator. Category, Indicator, BCBH.Requested.Methodology, Source , Methods and Notes in the data as they do not add any value and we obtain no new relationships or dependencies when these are taken into account.

How Regression Trees apply to our analysis:

We use Regression trees to investigate the relationships in our data as per the below questions of interest:

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?

To address the above questions, we need to find relationships/ dependencies between our response **Value** and the predictors given by **Place, Gender, Year and Ethnicity**. As the response is numerical, it makes sense to use Regression trees to generate trees that examine our data. We will implement Bagging, Random Forest and Boosting to reduce the Mean Squared Error. Furthermore, we can determine which of the variables are the most important, and list them down according to their significance.

Limitations of using Regression Trees on our data: We are limited by the number of variables which is four. Due to this, Random Forest would be same as Bagging because we need to use all four variables in both, and anything less than that will not give us relevant outputs.

However, Regression trees prove useful for analyzing the data when the response is numerical as mentioned above. They provide an easy interpretation and a good graphical representation as well.

Regression Tree Analysis

```
require(dplyr)

# Drinking data for High School Students
drink.st <- health.data %>%
  filter(Indicator == "Percent of High School Students Who Binge Drank")

# Smoking data for Students
smoke.st <- health.data %>%
 filter(Indicator == "Percent of High School Students Who Currently Smoke")

# Remove unwanted variables and rename some columns

drink.st <- drink.st[c(3:7)]
drink.st$Ethnicity <- drink.st$Race..Ethnicity
drink.st <- drink.st[-c(3)]

smoke.st <- smoke.st[c(3:7)]
smoke.st$Ethnicity <- smoke.st$Race..Ethnicity
smoke.st <- smoke.st[-c(3)]
```

Extracting and Refining Adults' data

```
# Drinking data for Adults
drink.ad <- health.data %>%
  filter(Indicator == "Percent of Adults Who Binge Drank")

# Smoking data for Adults
smoke.ad <- health.data %>%
  filter(Indicator == "Percent of Adults Who Currently Smoke")

# Remove unwanted variables and rename some columns
drink.ad <- drink.ad[c(3:7)]
drink.ad$Ethnicity <- drink.ad$Race..Ethnicity
drink.ad <- drink.ad[-c(3)]

smoke.ad <- smoke.ad[c(3:7)]
smoke.ad$Ethnicity <- smoke.ad$Race..Ethnicity
smoke.ad <- smoke.ad[-c(3)]

# Remove missing data which has been found only in smoke.ad dataset
smoke.ad <- smoke.ad %>%
  filter(Value != "NA")
```

Basic Regression Trees

We first create basic Regression trees for each of our four datasets **drink.st, smoke.st, drink.ad and smoke.ad**.

a) Drinking data for Students

```
require(tree)
```

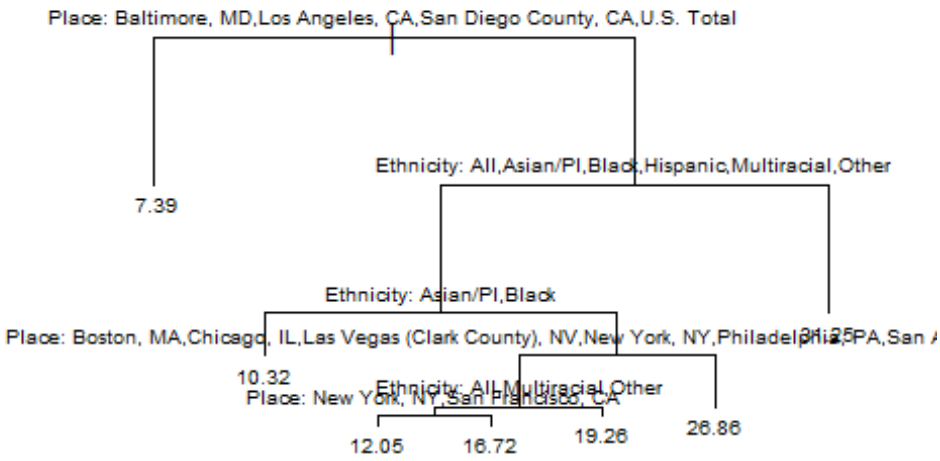
```
## Loading required package: tree
```

```
## Warning: package 'tree' was built under R version 3.6.3
```

```
set.seed(1)

## Create the Training dataset

train = sample(1:nrow(drink.st), nrow(drink.st)/2)
tree.drink.st=tree(Value~.,drink.st,subset=train)
summary(tree.drink.st)
```

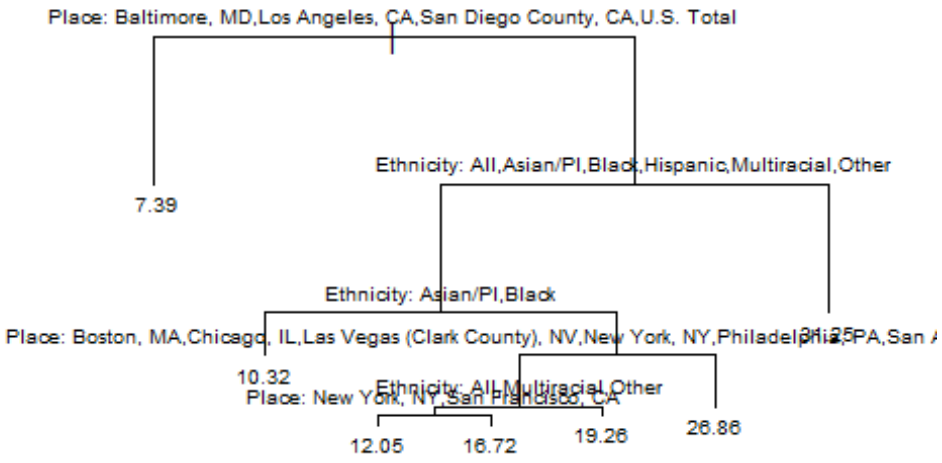
```
##
## Regression tree:
## tree(formula = Value ~ ., data = drink.st, subset = train)
## Variables actually used in tree construction:
## [1] "Place"     "Ethnicity"
## Number of terminal nodes:  7
## Residual mean deviance:  18.91 = 1116 / 59
## Distribution of residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -12.55000  -1.89200  -0.01818   0.00000   1.45000  16.05000
```

```
plot(tree.drink.st)
text(tree.drink.st,pretty=0, cex = 0.6)
```




```
tree.drink.st

## node), split, n, deviance, yval
##       * denotes terminal node
##
##   1) root 66 4981.00 14.54
##     2) Place: Baltimore, MD,Los Angeles, CA,San Diego County, CA,U.S. Tota
l 20   76.08  7.39 *
##     3) Place: Boston, MA,Chicago, IL,Denver, CO,Las Vegas (Clark County),
NV,Miami (Miami-Dade County), FL,New York, NY,Philadelphia, PA,San Antonio, T
X,San Francisco, CA 46 3438.00 17.65
##       6) Ethnicity: All,Asian/PI,Black,Hispanic,Multiracial,Other 40 1678.
00 15.61
##        12) Ethnicity: Asian/PI,Black 11   62.60 10.32 *
##        13) Ethnicity: All,Hispanic,Multiracial,Other 29 1191.00 17.62
##          26) Place: Boston, MA,Chicago, IL,Las Vegas (Clark County), NV,Ne
w York, NY,Philadelphia, PA,San Antonio, TX,San Francisco, CA 24  335.10 15.6
9
##            52) Ethnicity: All,Multiracial,Other 19  222.20 14.75
##             104) Place: New York, NY,San Francisco, CA 8   90.50 12.05 *
##             105) Place: Boston, MA,Chicago, IL,Las Vegas (Clark County), N
V,Philadelphia, PA,San Antonio, TX 11   30.82 16.72 *
##            53) Ethnicity: Hispanic 5   32.41 19.26 *
##          27) Place: Denver, CO,Miami (Miami-Dade County), FL 5  339.40 26.
86 *
##       7) Ethnicity: White 6  483.80 31.25 *
```

```
require(tree)

set.seed(1)

## Create the Training dataset

train = sample(1:nrow(drink.st), nrow(drink.st)/2)
tree.drink.st=tree(Value~.,drink.st,subset=train)
summary(tree.drink.st)

##
## Regression tree:
## tree(formula = Value ~ ., data = drink.st, subset = train)
## Variables actually used in tree construction:
## [1] "Place"     "Ethnicity"
## Number of terminal nodes:  7
## Residual mean deviance:  18.91 = 1116 / 59
## Distribution of residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -12.55000  -1.89200  -0.01818   0.00000   1.45000  16.05000

plot(tree.drink.st)
text(tree.drink.st,pretty=0, cex = 0.6)
```



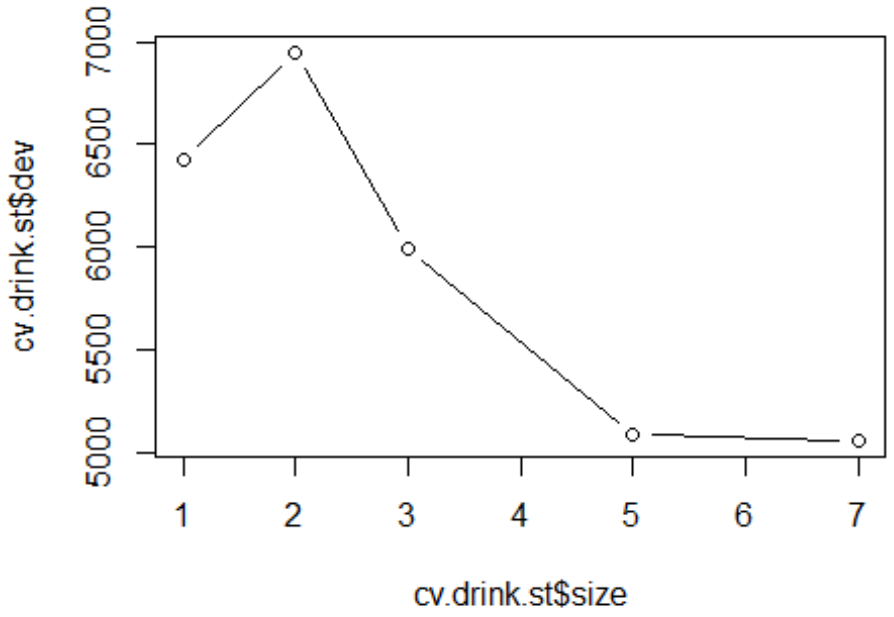

```
tree.drink.st
```

```
## node), split, n, deviance, yval
##       * denotes terminal node
##
##   1) root 66 4981.00 14.54
##     2) Place: Baltimore, MD,Los Angeles, CA,San Diego County, CA,U.S. Tota
l 20   76.08  7.39 *
##     3) Place: Boston, MA,Chicago, IL,Denver, CO,Las Vegas (Clark County),
NV,Miami (Miami-Dade County), FL,New York, NY,Philadelphia, PA,San Antonio, T
X,San Francisco, CA 46 3438.00 17.65
##       6) Ethnicity: All,Asian/PI,Black,Hispanic,Multiracial,Other 40 1678.
00 15.61
##        12) Ethnicity: Asian/PI,Black 11   62.60 10.32 *
##        13) Ethnicity: All,Hispanic,Multiracial,Other 29 1191.00 17.62
##          26) Place: Boston, MA,Chicago, IL,Las Vegas (Clark County), NV,Ne
w York, NY,Philadelphia, PA,San Antonio, TX,San Francisco, CA 24  335.10 15.6
9
##            52) Ethnicity: All,Multiracial,Other 19  222.20 14.75
##             104) Place: New York, NY,San Francisco, CA 8   90.50 12.05 *
##             105) Place: Boston, MA,Chicago, IL,Las Vegas (Clark County), N
V,Philadelphia, PA,San Antonio, TX 11   30.82 16.72 *
##            53) Ethnicity: Hispanic 5   32.41 19.26 *
##          27) Place: Denver, CO,Miami (Miami-Dade County), FL 5  339.40 26.
86 *
##       7) Ethnicity: White 6  483.80 31.25 *
```

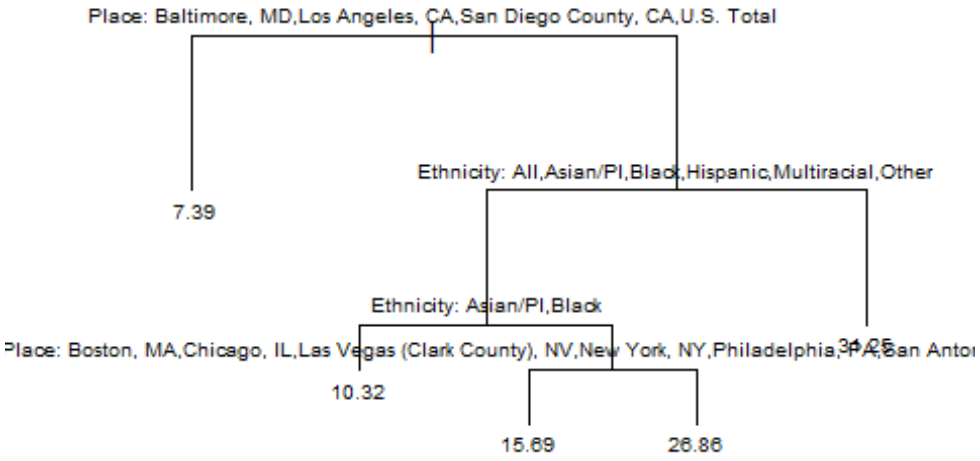
We will prune the tree now.

Pruning:

```
cv.drink.st=cv.tree(tree.drink.st)
plot(cv.drink.st$size,cv.drink.st$dev,type='b')
```




```
prune.drink.st=prune.tree(tree.drink.st,best=5)
plot(prune.drink.st)
text(prune.drink.st,pretty=0, cex = 0.6)
```



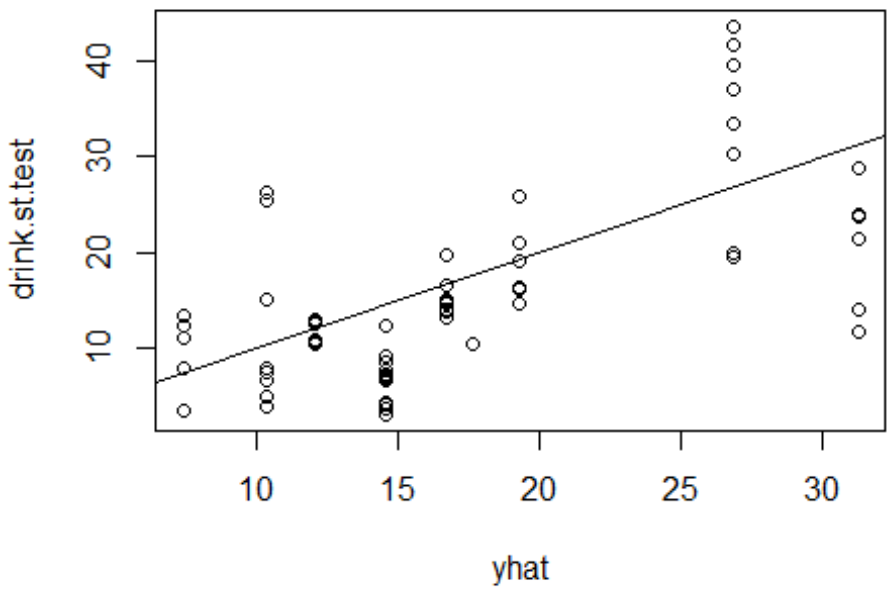
```
prune.drink.st

## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 66 4981.00 14.54
##    2) Place: Baltimore, MD,Los Angeles, CA,San Diego County, CA,U.S. Total
20   76.08  7.39 *
##    3) Place: Boston, MA,Chicago, IL,Denver, CO,Las Vegas (Clark County), N
V,Miami (Miami-Dade County), FL,New York, NY,Philadelphia, PA,San Antonio, TX
,San Francisco, CA 46 3438.00 17.65
##      6) Ethnicity: All,Asian/PI,Black,Hispanic,Multiracial,Other 40 1678.0
0 15.61
##       12) Ethnicity: Asian/PI,Black 11   62.60 10.32 *
##       13) Ethnicity: All,Hispanic,Multiracial,Other 29 1191.00 17.62
##         26) Place: Boston, MA,Chicago, IL,Las Vegas (Clark County), NV,New
York, NY,Philadelphia, PA,San Antonio, TX,San Francisco, CA 24  335.10 15.69
*
##         27) Place: Denver, CO,Miami (Miami-Dade County), FL 5  339.40 26.8
6 *
##      7) Ethnicity: White 6  483.80 31.25 *
```

Interpretation:

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.

Making Predictions on test data:

```
yhat=predict(tree.drink.st,newdata=drink.st[-train,])
drink.st.test = drink.st[-train,"Value"]
plot(yhat,drink.st.test)
abline(0,1)
```



Mean Squared Error

```
mean((yhat-drink.st.test)^2)
```

```
## [1] 54.8252
```

The error rate is quite high and we need to implement Bagging, Random Forest or Boosting to reduce the error and see if we can obtain a better fit.

Bagging

```
require(randomForest)
```

```
## Loading required package: randomForest
```

```
## Warning: package 'randomForest' was built under R version 3.6.3
```

```
## randomForest 4.6-14
```

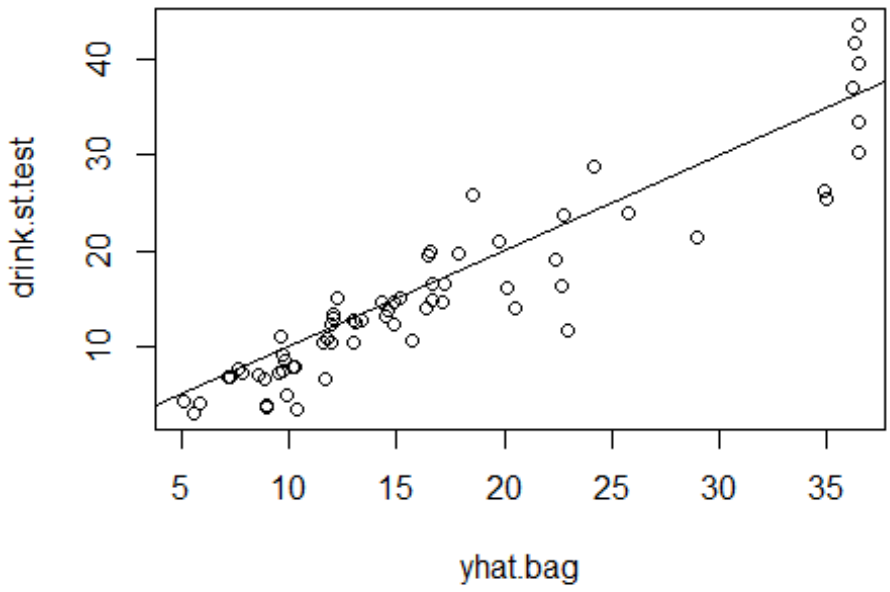
```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
set.seed(1)
bag.drink.st=randomForest(Value~.,data=drink.st,subset=train,mtry=4,ntree = 5
```

```
00, importance=TRUE)
bag.drink.st

##
## Call:
##  randomForest(formula = Value ~ ., data = drink.st, mtry = 4,      ntree =
500, importance = TRUE, subset = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 22.83371
##                     % Var explained: 69.75

yhat.bag = predict(bag.drink.st,newdata=drink.st[-train,])
plot(yhat.bag, drink.st.test)
abline(0,1)
```




```
mean((yhat.bag-drink.st.test)^2)

## [1] 14.58996
```

Bagging reduces the error rate significantly and it is computed as 14.59.

Random Forest:

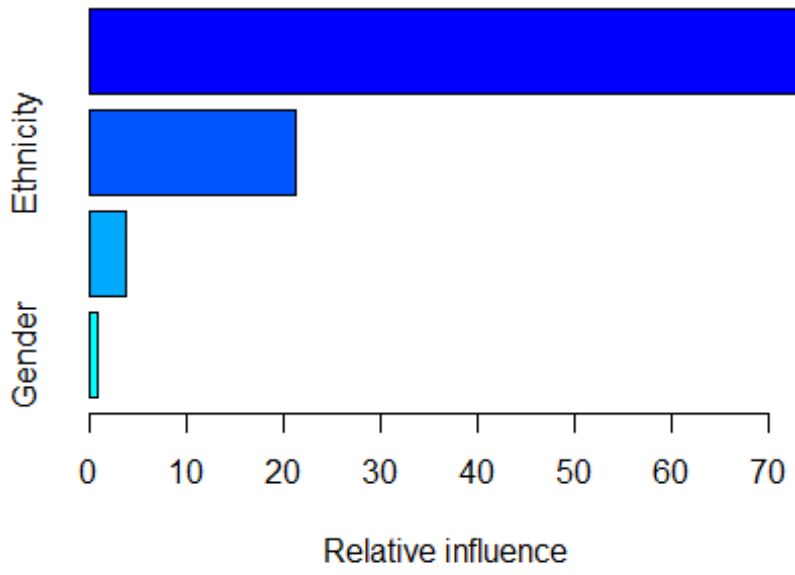
```
set.seed(1)
rf.drink.st=randomForest(Value~.,data=drink.st,subset=train,mtry=4,importance
=TRUE )
```

```
yhat.rf = predict(rf.drink.st,newdata=drink.st[-train,])
mean((yhat.rf-drink.st.test)^2)

## [1] 14.58996
```

Random Forest gives the same MSE as Bagging because both are equivalent in this case (due to same value of mtry).

Boosting

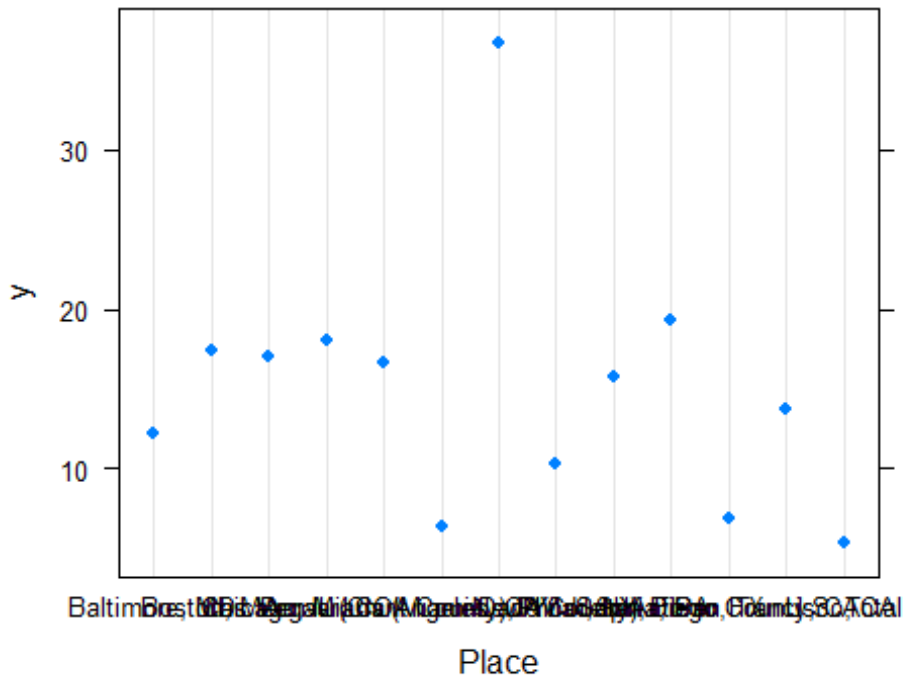
```
require(gbm)

## Loading required package: gbm

## Warning: package 'gbm' was built under R version 3.6.3

## Loaded gbm 2.1.5

set.seed(1)
boost.drink.st=gbm(Value~.,data=drink.st[train,],distribution="gaussian",n.trees=5000,interaction.depth=4)

summary(boost.drink.st)
```



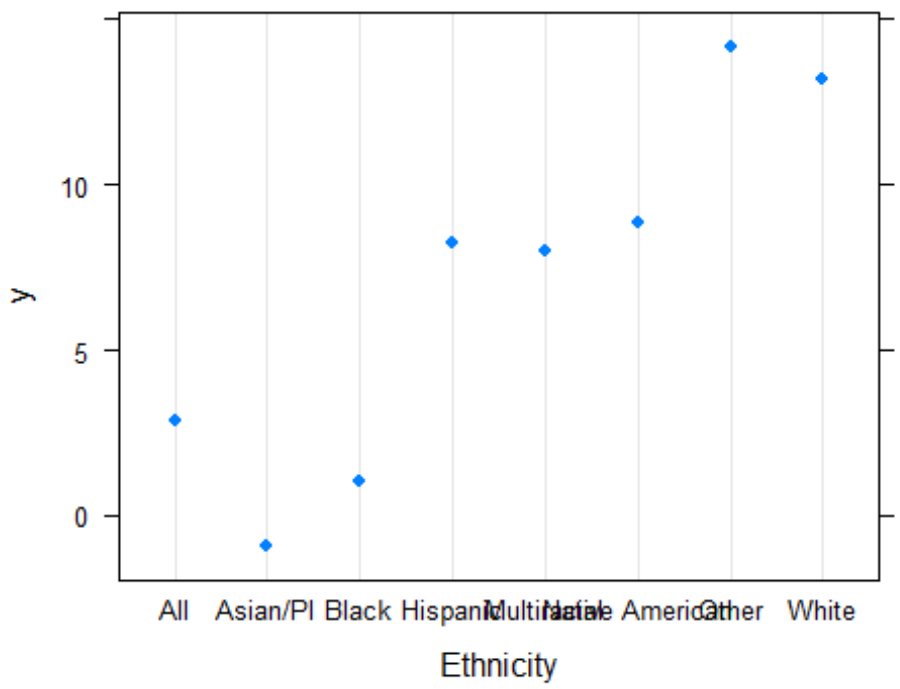

```
##                 var    rel.inf
## Place         Place 74.3072253
## Ethnicity Ethnicity 21.2226217
## Year           Year  3.6839787
## Gender       Gender  0.7861744
```

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.

```
par(mfrow=c(1,2))
plot(boost.drink.st,i="Place", type = "l")
```




```
plot(boost.drink.st,i="Ethnicity", type = "l")
```




```
yhat.boost=predict(boost.drink.st,newdata=drink.st[-train,],n.trees=5000)
mean((yhat.boost-drink.st.test)^2)

## [1] 23.06857

boost.drink.st=gbm(Value~.,data=drink.st[train,],distribution="gaussian",n.trees=5000,interaction.depth=4,shrinkage=0.2,verbose=F)
yhat.boost=predict(boost.drink.st,newdata=drink.st[-train,],n.trees=5000)
mean((yhat.boost-drink.st.test)^2)

## [1] 40.94359
```

The MSE when we perform Boosting is more than that of Bagging, 23.07 when we use default Shrinkage Parameter and 40.94 when the Shrinkage Parameter is increased to 0.2 .

Therefore, we choose Regression Tree with Bagging as the best model as it generates the least MSE.

Importance of Variables:

```
importance(bag.drink.st)

##             %IncMSE IncNodePurity
## Year       1.313437      74.85846
## Gender     1.507418      33.74922
## Place     34.522380    3436.56706
## Ethnicity 24.911216    1235.12730

varImpPlot(bag.drink.st)
```

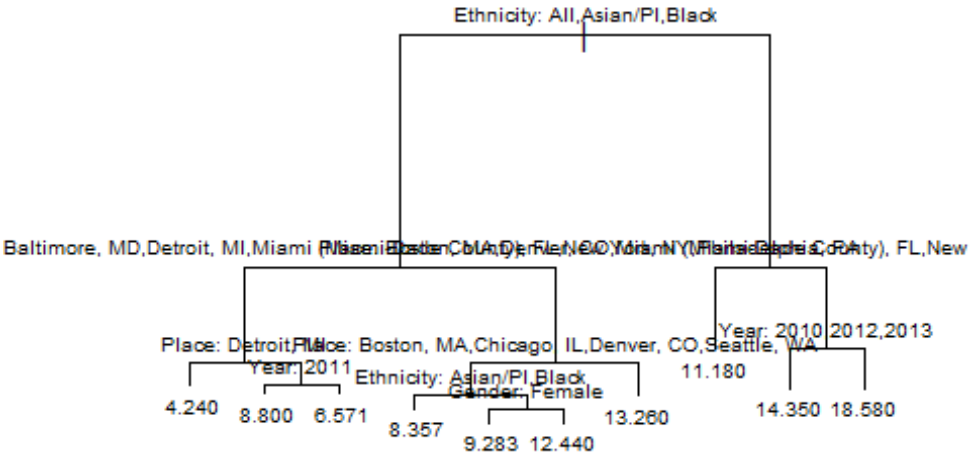
Conclusion:

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

```
set.seed(1)

## Create the Training dataset

train = sample(1:nrow(smoke.st), nrow(smoke.st)/2)
tree.smoke.st=tree(Value~.,smoke.st,subset=train)
summary(tree.smoke.st)

##
## Regression tree:
## tree(formula = Value ~ ., data = smoke.st, subset = train)
## Number of terminal nodes:  10
## Residual mean deviance:  4.48 = 286.7 / 64
```

```
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -4.2830 -1.3280 -0.3931  0.0000  1.6510  5.6430

plot(tree.smoke.st)
text(tree.smoke.st,pretty=0, cex = 0.6)
```



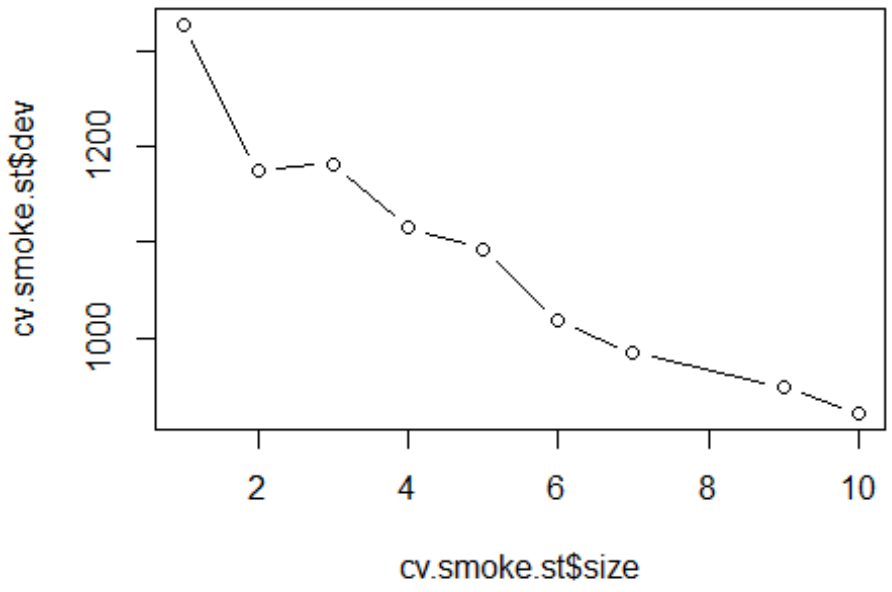

```
tree.smoke.st

## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 74 1284.000 11.100
##    2) Ethnicity: All,Asian/PI,Black 43  483.400  9.051
##      4) Place: Baltimore, MD,Detroit, MI,Miami (Miami-Dade County), FL,New
York, NY,Philadelphia, PA 18  101.700  6.667
##        8) Place: Detroit, MI 5    5.852  4.240 *
##        9) Place: Baltimore, MD,Miami (Miami-Dade County), FL,New York, NY,
Philadelphia, PA 13   55.040  7.600
##         18) Year: 2011 6   22.500  8.800 *
##         19) Year: 2013 7   16.490  6.571 *
##      5) Place: Boston, MA,Chicago, IL,Denver, CO,Houston, TX,San Antonio,
TX,Seattle, WA,U.S. Total 25  205.700 10.770
##       10) Place: Boston, MA,Chicago, IL,Denver, CO,Seattle, WA 18  118.500
9.800
##         20) Ethnicity: Asian/PI,Black 7   44.200  8.357 *
##         21) Ethnicity: All 11   50.480 10.720
```

```
##           42) Gender: Female 6   14.570  9.283 *
##           43) Gender: Both,Male 5    8.732 12.440 *
##       11) Place: Houston, TX,San Antonio, TX,U.S. Total 7   26.960 13.260
*
##    3) Ethnicity: Hispanic,Multiracial,Native American,Other,White 31  370.
700 13.940
##      6) Place: Boston, MA,Denver, CO,Miami (Miami-Dade County), FL,New Yor
k, NY,San Antonio, TX 12   33.100 11.180 *
##      7) Place: Chicago, IL,Houston, TX,Philadelphia, PA,Seattle, WA,U.S. T
otal 19  188.000 15.680
##       14) Year: 2010,2012,2013 13   65.590 14.350 *
##       15) Year: 2011 6   48.750 18.580 *
```

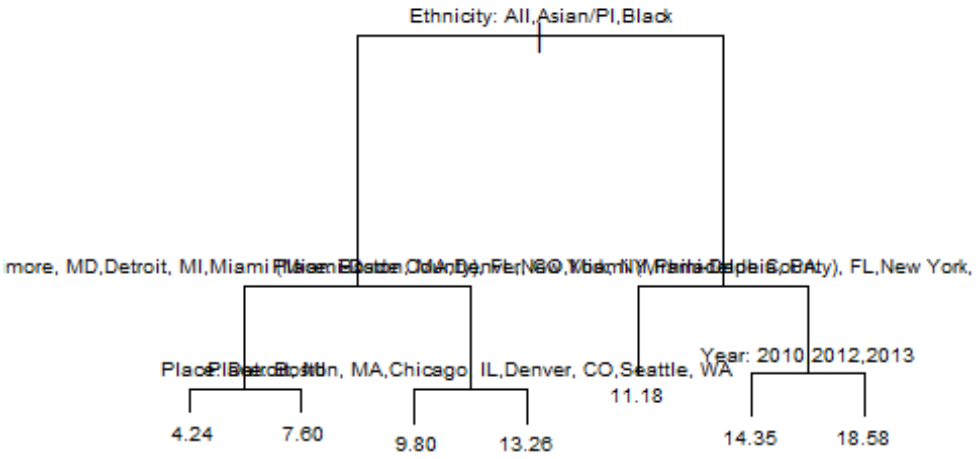
We will perform Pruning on the tree now.

Pruning:

```
cv.smoke.st=cv.tree(tree.smoke.st)
plot(cv.smoke.st$size,cv.smoke.st$dev,type='b')
```




```
prune.smoke.st=prune.tree(tree.smoke.st,best=7)
plot(prune.smoke.st)
text(prune.smoke.st,pretty=0, cex = 0.6)
```



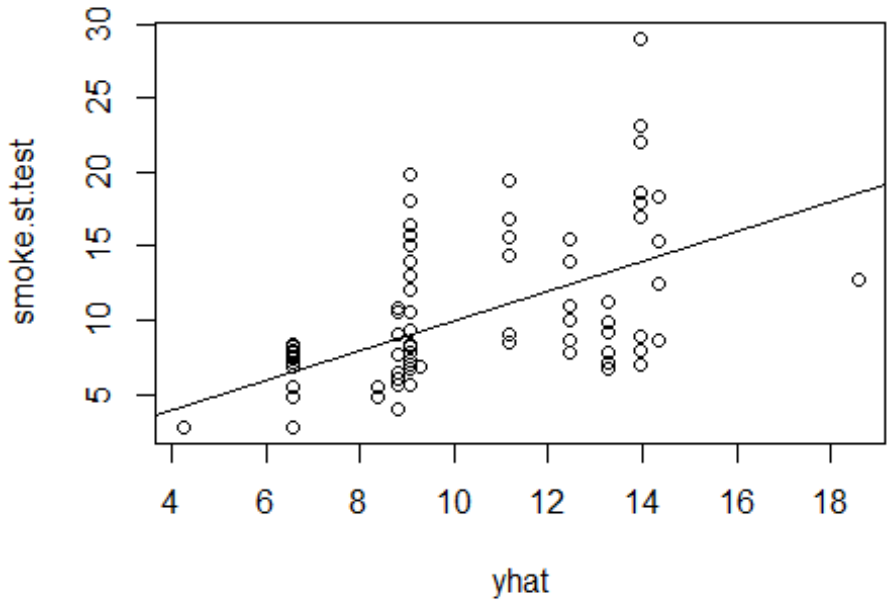

```
prune.smoke.st

## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 74 1284.000 11.100
##    2) Ethnicity: All,Asian/PI,Black 43  483.400  9.051
##      4) Place: Baltimore, MD,Detroit, MI,Miami (Miami-Dade County), FL,New
York, NY,Philadelphia, PA 18  101.700  6.667
##        8) Place: Detroit, MI 5    5.852  4.240 *
##        9) Place: Baltimore, MD,Miami (Miami-Dade County), FL,New York, NY,
Philadelphia, PA 13   55.040  7.600 *
##      5) Place: Boston, MA,Chicago, IL,Denver, CO,Houston, TX,San Antonio,
TX,Seattle, WA,U.S. Total 25  205.700 10.770
##       10) Place: Boston, MA,Chicago, IL,Denver, CO,Seattle, WA 18  118.500
9.800 *
##       11) Place: Houston, TX,San Antonio, TX,U.S. Total 7   26.960 13.260
*
##    3) Ethnicity: Hispanic,Multiracial,Native American,Other,White 31  370.
700 13.940
##      6) Place: Boston, MA,Denver, CO,Miami (Miami-Dade County), FL,New Yor
k, NY,San Antonio, TX 12   33.100 11.180 *
##      7) Place: Chicago, IL,Houston, TX,Philadelphia, PA,Seattle, WA,U.S. T
otal 19  188.000 15.680
##       14) Year: 2010,2012,2013 13   65.590 14.350 *
##       15) Year: 2011 6   48.750 18.580 *
```

Interpretation:

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.

Making Predictions on test data:

```
yhat=predict(tree.smoke.st,newdata=smoke.st[-train,])
smoke.st.test = smoke.st[-train,"Value"]
plot(yhat,smoke.st.test)
abline(0,1)
```



Mean Squared Error

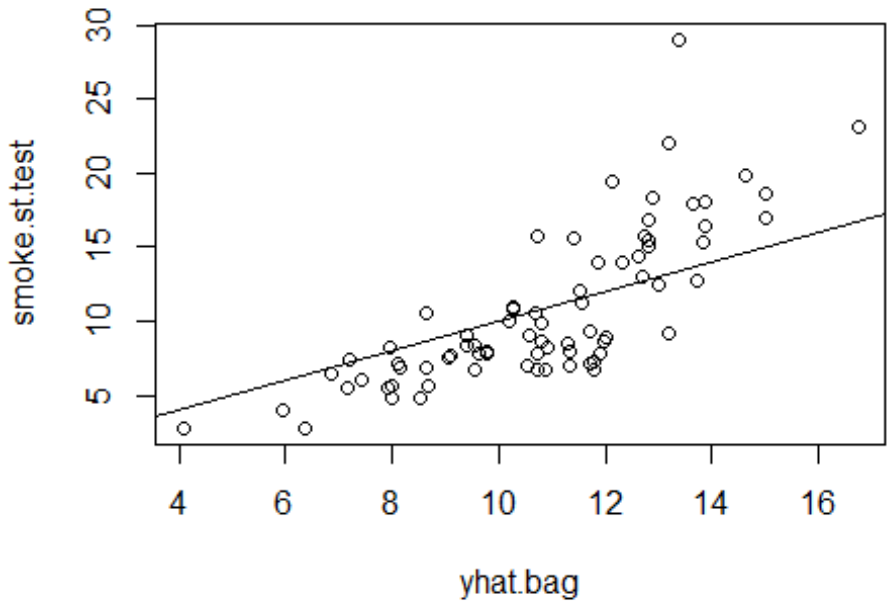
```
mean((yhat-smoke.st.test)^2)

## [1] 20.47114
```

The error rate is not bad and we can implement Bagging, Random Forest or Boosting to reduce the error and see if we can obtain a better fit.

Bagging

```
require(randomForest)
set.seed(1)
bag.smoke.st=randomForest(Value~.,data=smoke.st,subset=train,mtry=4,ntree = 5
```

```
00, importance=TRUE)
bag.smoke.st

##
## Call:
##  randomForest(formula = Value ~ ., data = smoke.st, mtry = 4,      ntree =
500, importance = TRUE, subset = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 7.397056
##                     % Var explained: 57.38

yhat.bag = predict(bag.smoke.st,newdata=smoke.st[-train,])
plot(yhat.bag, smoke.st.test)
abline(0,1)
```




```
mean((yhat.bag-smoke.st.test)^2)

## [1] 12.81416
```

Bagging reduces the error rate significantly and it is computed as 12.81.

Random Forest:

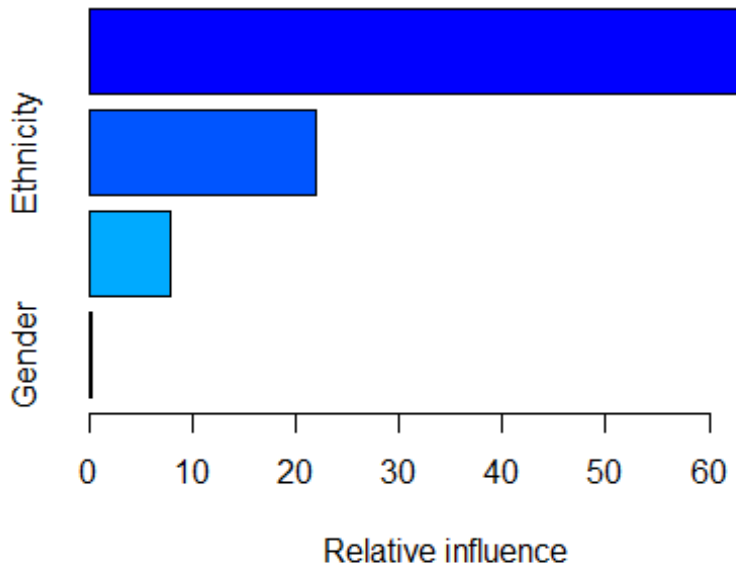
```
set.seed(1)
rf.smoke.st=randomForest(Value~.,data=smoke.st,subset=train,mtry=4,importance
=TRUE )
```

```
yhat.rf = predict(rf.smoke.st,newdata=smoke.st[-train,])
mean((yhat.rf-smoke.st.test)^2)

## [1] 12.81416
```

Random Forest gives the same MSE as Bagging because both are equivalent in this case (due to same value of mtry).

Boosting

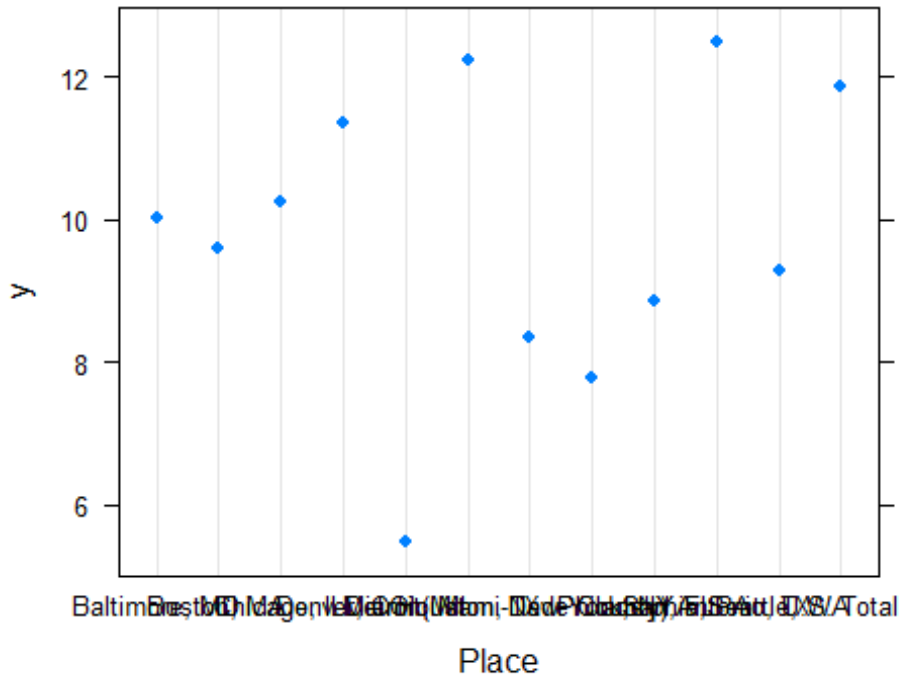
```
require(gbm)
set.seed(1)
boost.smoke.st=gbm(Value~.,data=smoke.st[train,],distribution="gaussian",n.trees=5000,interaction.depth=4)

summary(boost.smoke.st)
```



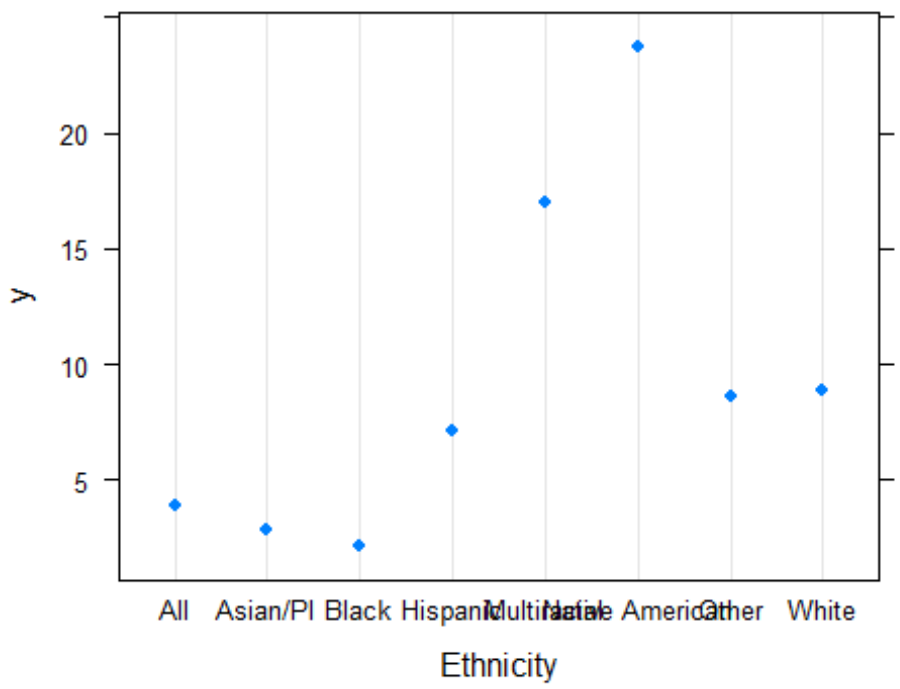

```
##                  var    rel.inf
## Place          Place 69.9009452
## Ethnicity Ethnicity 22.0367867
## Year           Year  7.8870199
## Gender       Gender  0.1752482
```

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.

```
par(mfrow=c(1,2))
plot(boost.smoke.st,i="Place", type = "l")
```




```
plot(boost.smoke.st,i="Ethnicity", type = "l")
```



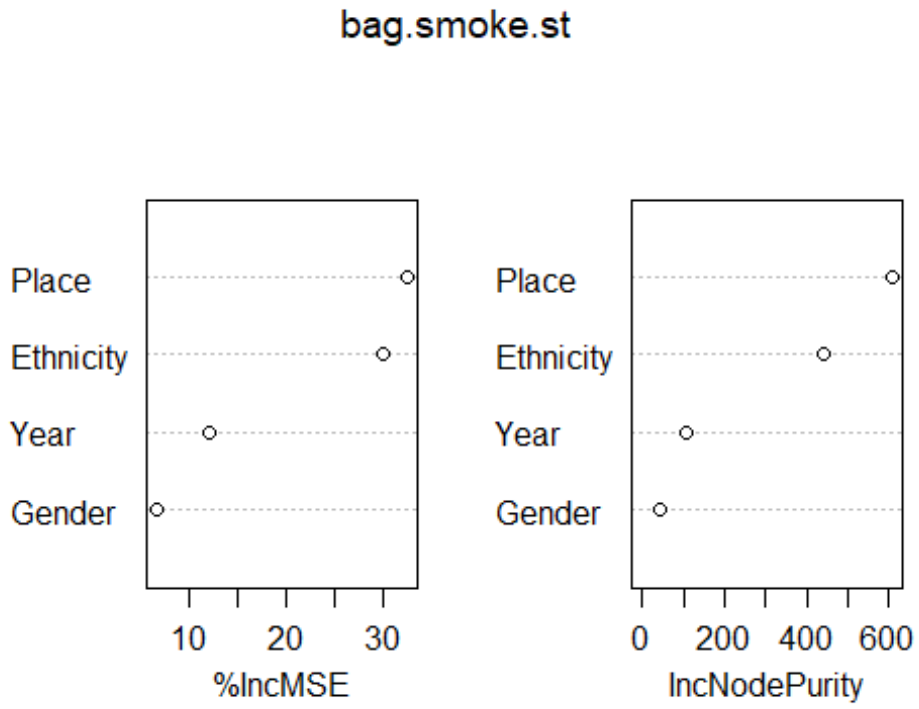

```
yhat.boost=predict(boost.smoke.st,newdata=smoke.st[-train,],n.trees=5000)
mean((yhat.boost-smoke.st.test)^2)

## [1] 16.08601

boost.smoke.st=gbm(Value~.,data=smoke.st[train,],distribution="gaussian",n.trees=5000,interaction.depth=4,shrinkage=0.2,verbose=F)
yhat.boost=predict(boost.smoke.st,newdata=smoke.st[-train,],n.trees=5000)
mean((yhat.boost-smoke.st.test)^2)

## [1] 34.99435
```

The MSE when we perform Boosting is more than that of Bagging, 16.09 when we use default Shrinkage Parameter and 34.99 when the Shrinkage Parameter is increased to 0.2 .

Therefore, we choose Regression Tree with Bagging as the best model as it generates the least MSE.

Importance of Variables:

```
importance(bag.smoke.st)

##             %IncMSE IncNodePurity
## Year      12.064898      104.7582
## Gender     6.776782       43.5807
## Place     32.525871      609.6558
## Ethnicity 30.024516      439.0861

varImpPlot(bag.smoke.st)
```



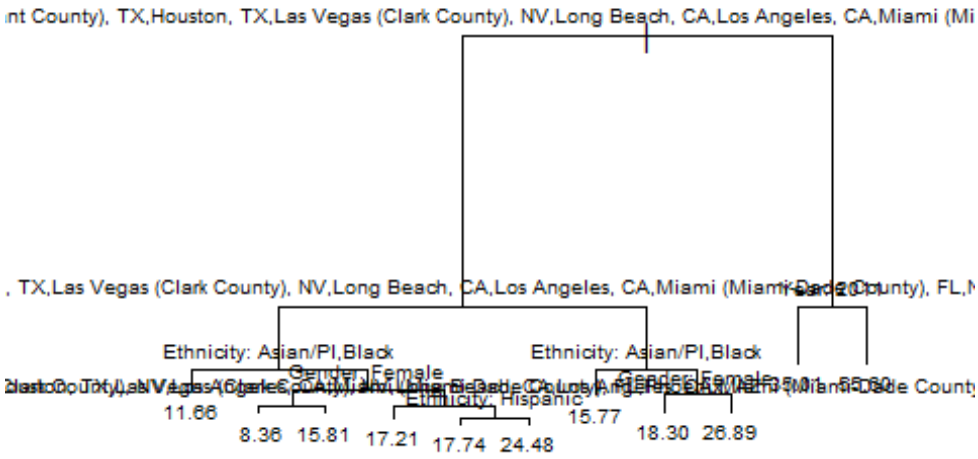
As seen above the most important predictor is **Place** and the next best predictor is **Ethnicity**.

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

```
require(tree)

set.seed(1)

## Create the Training dataset

train = sample(1:nrow(drink.ad), nrow(drink.ad)/2)
tree.drink.ad=tree(Value~.,drink.ad,subset=train)
summary(tree.drink.ad)

##
## Regression tree:
## tree(formula = Value ~ ., data = drink.ad, subset = train)
## Number of terminal nodes:  11
## Residual mean deviance:  27.09 = 3495 / 129
## Distribution of residuals:
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -14.7100  -2.5510   0.4019   0.0000   2.3660  14.5100

plot(tree.drink.ad)
text(tree.drink.ad,pretty=0, cex = 0.6)
```



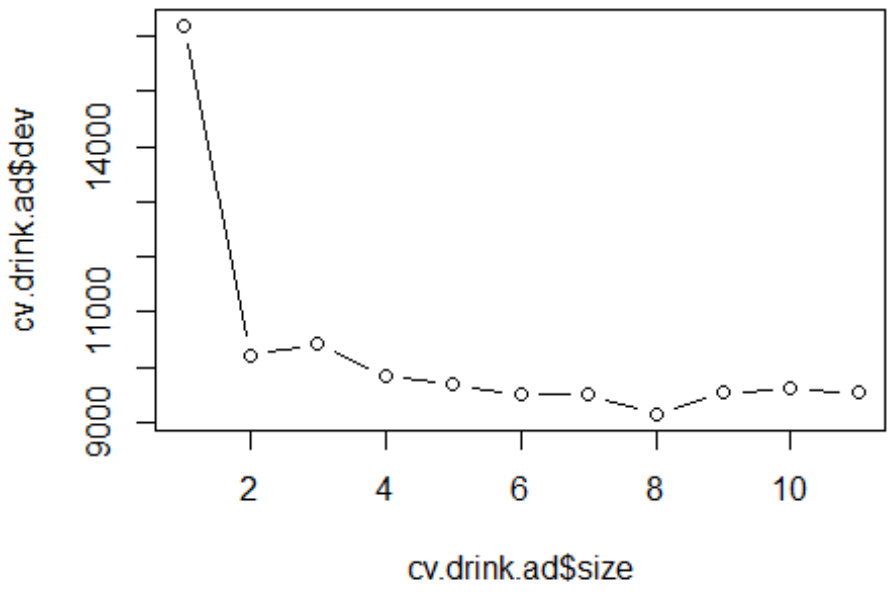

```
tree.drink.ad

## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 140 15910.000 21.69
##    2) Place: Atlanta (Fulton County), GA,Baltimore, MD,Boston, MA,Denver,
CO,Fort Worth (Tarrant County), TX,Houston, TX,Las Vegas (Clark County), NV,L
ong Beach, CA,Los Angeles, CA,Miami (Miami-Dade County), FL,New York, NY,Phil
adelphia, PA,Phoenix, AZ,San Antonio, TX,Seattle, WA,U.S. Total,Washington, D
C 125  6753.000 19.27
##      4) Place: Baltimore, MD,Fort Worth (Tarrant County), TX,Houston, TX,L
as Vegas (Clark County), NV,Long Beach, CA,Los Angeles, CA,Miami (Miami-Dade
County), FL,New York, NY,Philadelphia, PA,Phoenix, AZ,Seattle, WA 77  2530.00
0 16.45
##        8) Ethnicity: Asian/PI,Black 17   251.000 11.66 *
##        9) Ethnicity: All,Hispanic,Other,White 60  1778.000 17.81
##         18) Gender: Female 12   242.600 12.71
##           36) Place: Las Vegas (Clark County), NV,Los Angeles, CA,Miami (M
iami-Dade County), FL,Phoenix, AZ 5     9.352  8.36 *
##           37) Place: Baltimore, MD,New York, NY,Philadelphia, PA,Seattle,
```

```
WA 7    71.150 15.81 *
##         19) Gender: Both,Male 48  1145.000 19.09
##           38) Place: Fort Worth (Tarrant County), TX,Houston, TX,Las Vegas
(Clark County), NV,Long Beach, CA,Los Angeles, CA,Miami (Miami-Dade County),
FL,Phoenix, AZ,Seattle, WA 31   398.000 17.21 *
##           39) Place: Baltimore, MD,New York, NY,Philadelphia, PA 17   440.
300 22.50
##             78) Ethnicity: Hispanic 5    94.970 17.74 *
##             79) Ethnicity: All,Other,White 12   184.800 24.48 *
##      5) Place: Atlanta (Fulton County), GA,Boston, MA,Denver, CO,San Anton
io, TX,U.S. Total,Washington, DC 48  2632.000 23.79
##       10) Ethnicity: Asian/PI,Black 8   196.400 15.77 *
##       11) Ethnicity: All,Hispanic,Other,White 40  1820.000 25.39
##         22) Gender: Female 7    54.420 18.30 *
##         23) Gender: Both,Male 33  1339.000 26.89 *
##    3) Place: Chicago, IL,San Diego County, CA,San Jose, CA 15  2309.000 41
.87
##      6) Year: 2011 10   390.400 35.01 *
##      7) Year: 2013 5   505.200 55.60 *
```

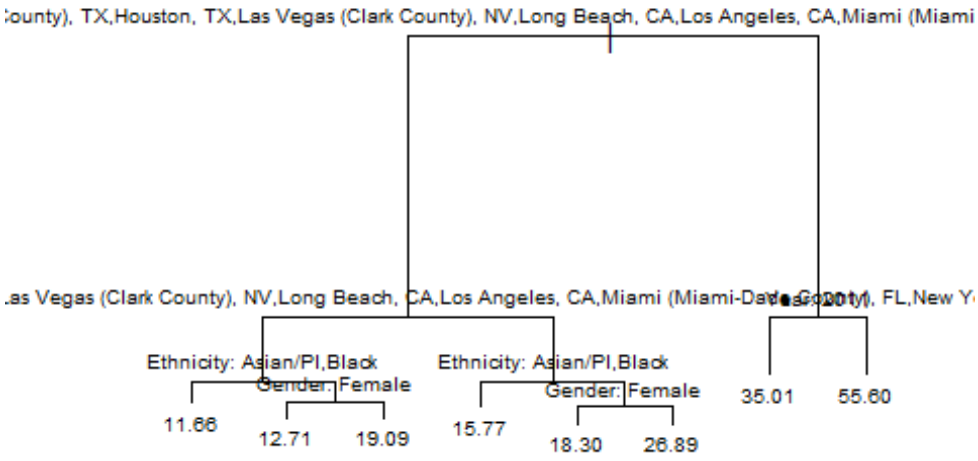
We will prune the tree now.

Pruning:

```
cv.drink.ad=cv.tree(tree.drink.ad)
plot(cv.drink.ad$size,cv.drink.ad$dev,type='b')
```



```
prune.drink.ad=prune.tree(tree.drink.ad,best=8)
plot(prune.drink.ad)
text(prune.drink.ad,pretty=0, cex = 0.6)
```




```
prune.drink.ad
```

```
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 140 15910.00 21.69
##    2) Place: Atlanta (Fulton County), GA,Baltimore, MD,Boston, MA,Denver,
CO,Fort Worth (Tarrant County), TX,Houston, TX,Las Vegas (Clark County), NV,L
ong Beach, CA,Los Angeles, CA,Miami (Miami-Dade County), FL,New York, NY,Phil
adelphia, PA,Phoenix, AZ,San Antonio, TX,Seattle, WA,U.S. Total,Washington, D
C 125  6753.00 19.27
##      4) Place: Baltimore, MD,Fort Worth (Tarrant County), TX,Houston, TX,L
as Vegas (Clark County), NV,Long Beach, CA,Los Angeles, CA,Miami (Miami-Dade
County), FL,New York, NY,Philadelphia, PA,Phoenix, AZ,Seattle, WA 77  2530.00
16.45
##        8) Ethnicity: Asian/PI,Black 17   251.00 11.66 *
##        9) Ethnicity: All,Hispanic,Other,White 60  1778.00 17.81
##         18) Gender: Female 12   242.60 12.71 *
##         19) Gender: Both,Male 48  1145.00 19.09 *
##      5) Place: Atlanta (Fulton County), GA,Boston, MA,Denver, CO,San Anton
io, TX,U.S. Total,Washington, DC 48  2632.00 23.79
##       10) Ethnicity: Asian/PI,Black 8   196.40 15.77 *
##       11) Ethnicity: All,Hispanic,Other,White 40  1820.00 25.39
```

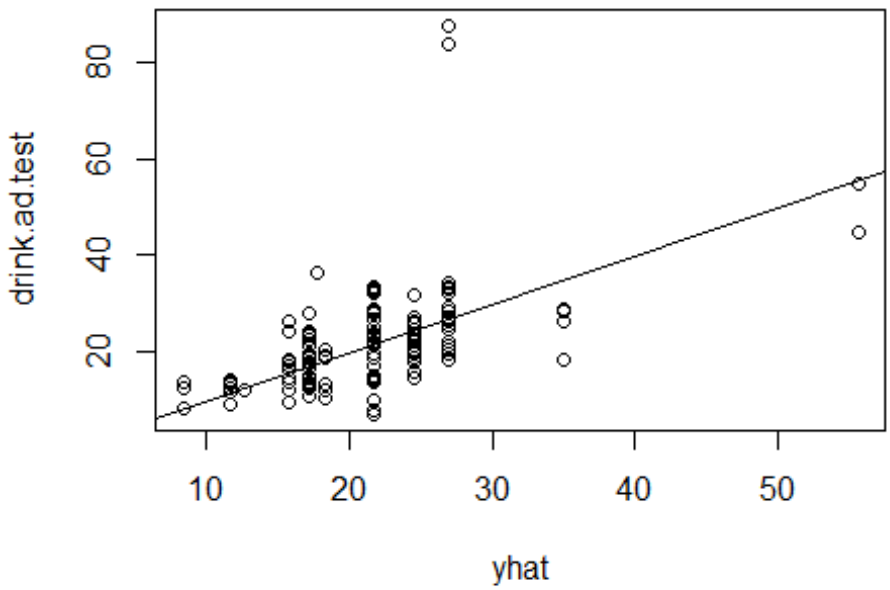
```
##         22) Gender: Female 7    54.42 18.30 *
##         23) Gender: Both,Male 33  1339.00 26.89 *
##    3) Place: Chicago, IL,San Diego County, CA,San Jose, CA 15  2309.00 41.
87
##      6) Year: 2011 10   390.40 35.01 *
##      7) Year: 2013 5   505.20 55.60 *
```

Interpretation:

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.

Making Predictions on test data:

```
yhat=predict(tree.drink.ad,newdata=drink.ad[-train,])
drink.ad.test = drink.ad[-train,"Value"]
plot(yhat,drink.ad.test)
abline(0,1)
```



Mean Squared Error

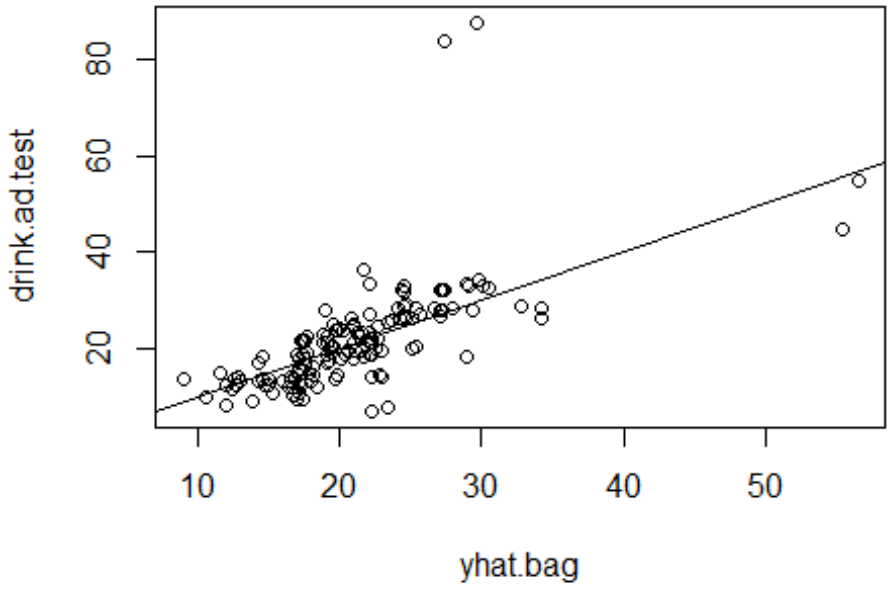
```
mean((yhat-drink.ad.test)^2)
```

```
## [1] 82.64551
```

The error rate is quite high and we need to implement Bagging, Random Forest or Boosting to reduce the error and see if we can obtain a better fit.

Bagging

```
require(randomForest)
set.seed(1)
bag.drink.ad=randomForest(Value~.,data=drink.ad,subset=train,mtry=4,ntree = 5
00, importance=TRUE)
bag.drink.ad

##
## Call:
##  randomForest(formula = Value ~ ., data = drink.ad, mtry = 4,      ntree =
500, importance = TRUE, subset = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 31.81882
##                     % Var explained: 71.99

yhat.bag = predict(bag.drink.ad,newdata=drink.ad[-train,])
plot(yhat.bag, drink.ad.test)
abline(0,1)
```




```
mean((yhat.bag-drink.ad.test)^2)

## [1] 67.5561
```

Bagging reduces the error rate significantly and it is computed as 67.55 which is still high.

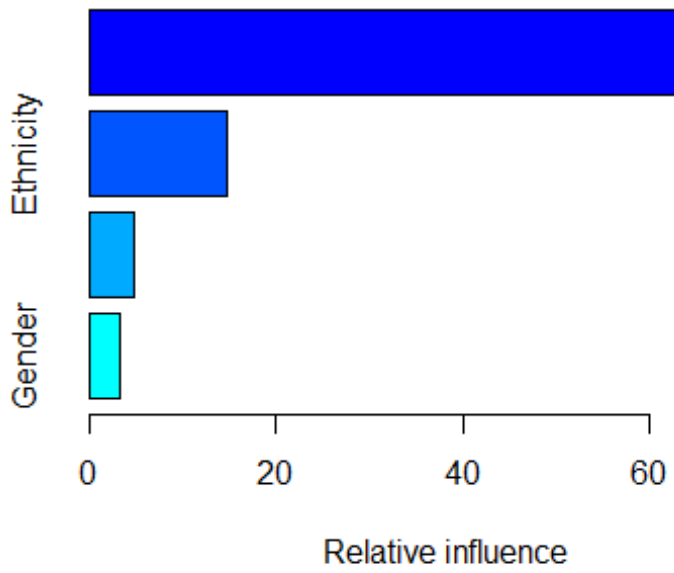
Random Forest:

```
set.seed(1)
rf.drink.ad=randomForest(Value~.,data=drink.ad,subset=train,mtry=4,importance
=TRUE )
yhat.rf = predict(rf.drink.ad,newdata=drink.ad[-train,])
mean((yhat.rf-drink.ad.test)^2)

## [1] 67.5561
```

Random Forest gives the same MSE as Bagging because both are equivalent in this case (due to same value of mtry).

Boosting

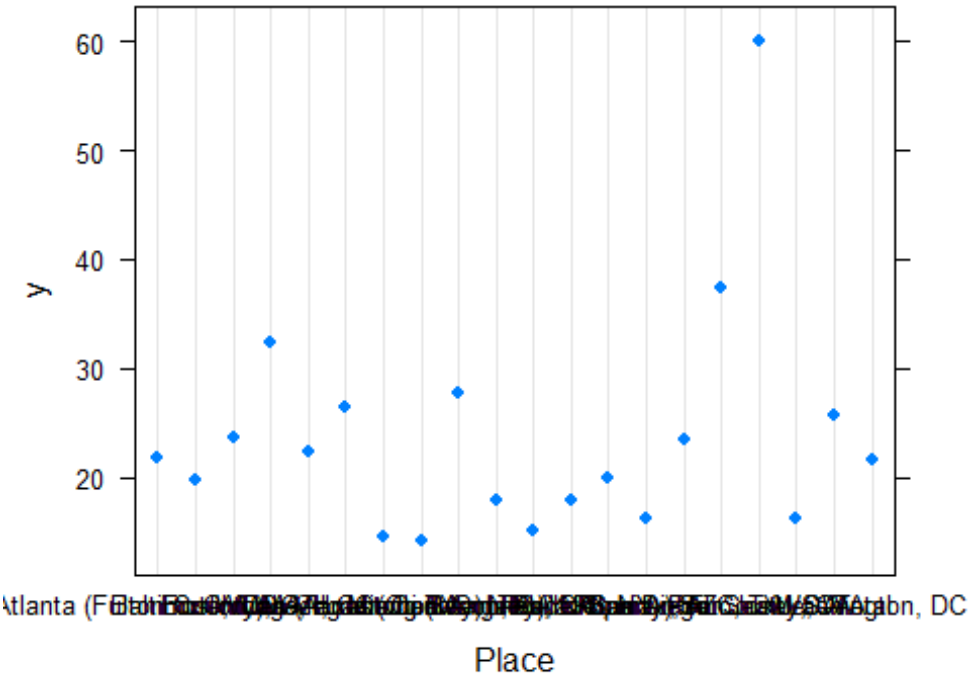
```
require(gbm)
set.seed(1)
boost.drink.ad=gbm(Value~.,data=drink.ad[train,],distribution="gaussian",n.tr
ees=5000,interaction.depth=4)

summary(boost.drink.ad)
```



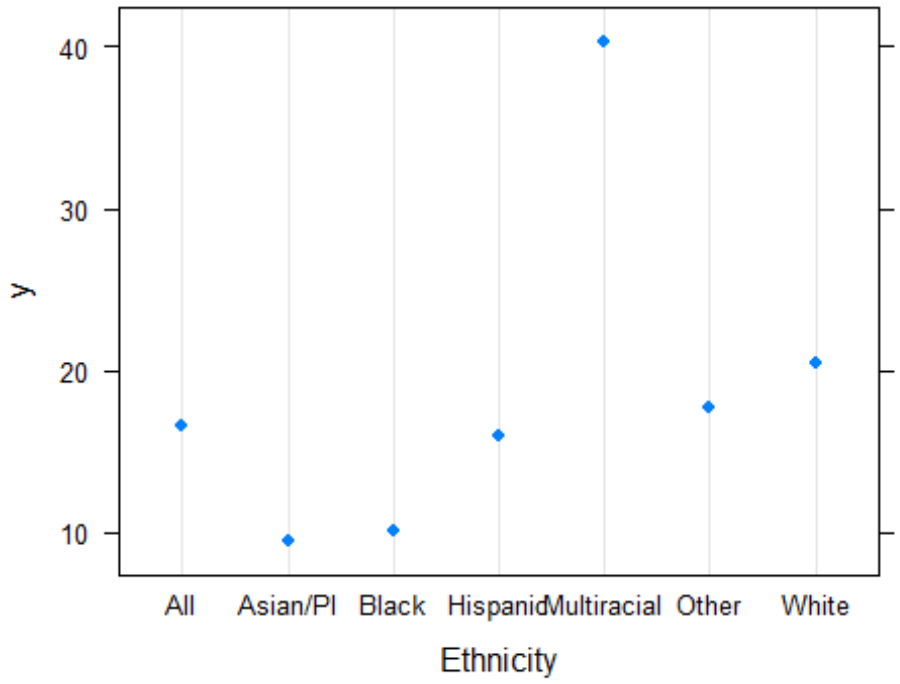

```
##                 var   rel.inf
## Place         Place 77.332858
## Ethnicity Ethnicity 14.661675
## Year           Year  4.752362
## Gender       Gender  3.253105
```

Again, we see that **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.

```
par(mfrow=c(1,2))
plot(boost.drink.ad,i="Place")
```




```
plot(boost.drink.ad,i="Ethnicity")
```




```
yhat.boost=predict(boost.drink.ad,newdata=drink.ad[-train,],n.trees=5000)
mean((yhat.boost-drink.ad.test)^2)
```

```
## [1] 77.60509
```

```
boost.drink.ad=gbm(Value~.,data=drink.ad[train,],distribution="gaussian",n.trees=5000,interaction.depth=4,shrinkage=0.2,verbose=F)
yhat.boost=predict(boost.drink.ad,newdata=drink.ad[-train,],n.trees=5000)
mean((yhat.boost-drink.ad.test)^2)
```

```
## [1] 79.02369
```

There is no improvement in MSE when we perform Boosting. It is more than that of Bagging, 79.02 when we use default Shrinkage Parameter and 83.89 when the Shrinkage Parameter is increased to 0.2 .

Therefore, we choose Regression Tree with Bagging as the best model as it generates the least MSE.

Importance of Variables:

```
importance(bag.drink.ad)
```

```
##             %IncMSE IncNodePurity
## Year       11.04616      881.5873
## Gender     28.94206     1037.5100
## Place      45.94340    11259.2414
## Ethnicity 29.90383     2124.0443
```

```
varImpPlot(bag.drink.ad)
```

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

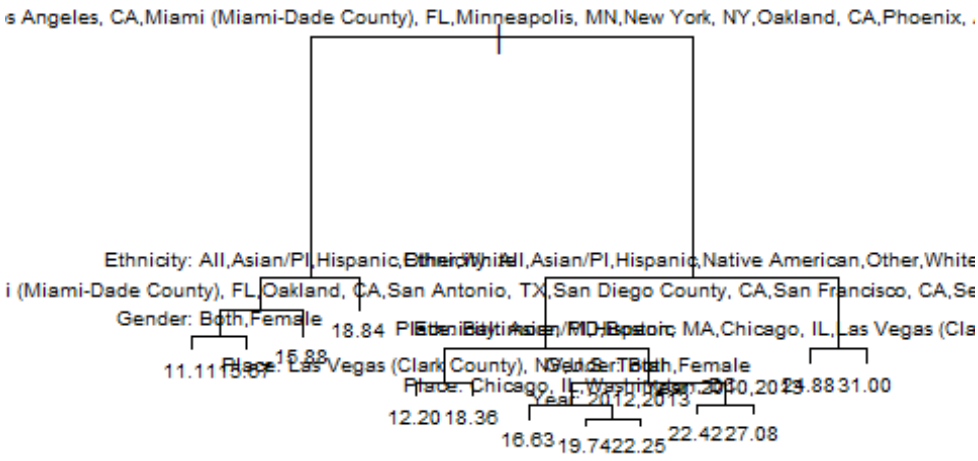
When we check the plots, it is clear that **San Jose** has the highest rate of binge drinking among adults and Las Vegas and Houston seem to have lower binge drinking rates.

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

```
require(tree)

set.seed(1)

## Create the Training dataset

train = sample(1:nrow(smoke.ad), nrow(smoke.ad)/2)
tree.smoke.ad=tree(Value~.,smoke.ad,subset=train)
summary(tree.smoke.ad)

##
## Regression tree:
## tree(formula = Value ~ ., data = smoke.ad, subset = train)
## Number of terminal nodes:  13
## Residual mean deviance:  9.475 = 1203 / 127
```

```
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -7.9440 -1.8160 -0.1315  0.0000  1.7280  9.5560

plot(tree.smoke.ad)
text(tree.smoke.ad,pretty=0, cex = 0.6)
```



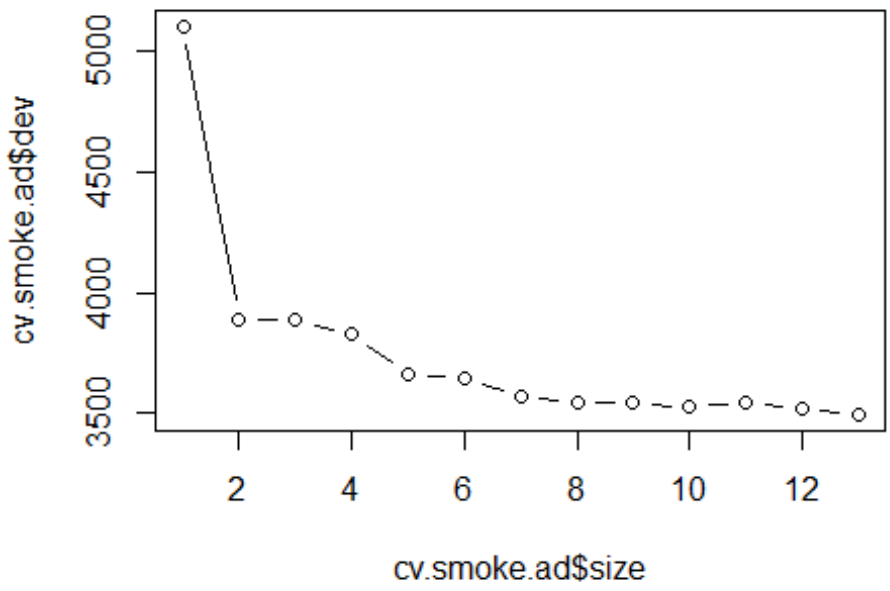

```
tree.smoke.ad

## node), split, n, deviance, yval
##       * denotes terminal node
##
##   1) root 140 5062.000 17.94
##     2) Place: Atlanta (Fulton County), GA,Fort Worth (Tarrant County), TX,
Long Beach, CA,Los Angeles, CA,Miami (Miami-Dade County), FL,Minneapolis, MN,
New York, NY,Oakland, CA,Phoenix, AZ,San Antonio, TX,San Diego County, CA,San
Francisco, CA,San Jose, CA,Seattle, WA 65 1135.000 14.00
##       4) Ethnicity: All,Asian/PI,Hispanic,Other,White 56  664.400 13.22
##         8) Place: Atlanta (Fulton County), GA,Los Angeles, CA,Miami (Miami
-Dade County), FL,Oakland, CA,San Antonio, TX,San Diego County, CA,San Franci
sco, CA,Seattle, WA 37  314.600 11.85
##          16) Gender: Both,Female 31  179.100 11.11 *
##          17) Gender: Male 6   31.090 15.67 *
##         9) Place: Fort Worth (Tarrant County), TX,Long Beach, CA,Minneapol
is, MN,New York, NY,Phoenix, AZ,San Jose, CA 19  145.400 15.88 *
##       5) Ethnicity: Black,Multiracial,Native American 9  225.500 18.84 *
##     3) Place: Baltimore, MD,Boston, MA,Chicago, IL,Denver, CO,Las Vegas (C
```

```
lark County), NV,Philadelphia, PA,U.S. Total,Washington, DC 75 2039.000 21.36
##       6) Ethnicity: All,Asian/PI,Hispanic,Native American,Other,White 62 1
258.000 20.13
##        12) Ethnicity: Asian/PI,Hispanic 10  203.900 15.28
##          24) Place: Las Vegas (Clark County), NV,U.S. Total 5   16.640 12.
20 *
##          25) Place: Boston, MA,Philadelphia, PA,Washington, DC 5   92.390
18.36 *
##        13) Ethnicity: All,Native American,Other,White 52  773.400 21.06
##          26) Gender: Both,Female 41  522.100 20.13
##            52) Place: Chicago, IL,Washington, DC 7  129.700 16.63 *
##            53) Place: Baltimore, MD,Boston, MA,Denver, CO,Las Vegas (Clark
County), NV,Philadelphia, PA,U.S. Total 34  289.000 20.85
##             106) Year: 2012,2013 19  142.300 19.74 *
##             107) Year: 2010,2011,2014 15   93.640 22.25 *
##          27) Gender: Male 11   82.750 24.54
##            54) Year: 2010,2013 6   14.510 22.42 *
##            55) Year: 2011,2012 5    8.928 27.08 *
##       7) Ethnicity: Black,Multiracial 13  239.500 27.23
##        14) Place: Baltimore, MD,Boston, MA,Chicago, IL,Las Vegas (Clark Co
unty), NV,Philadelphia, PA 8   67.270 24.88 *
##        15) Place: Denver, CO,Washington, DC 5   56.800 31.00 *
```

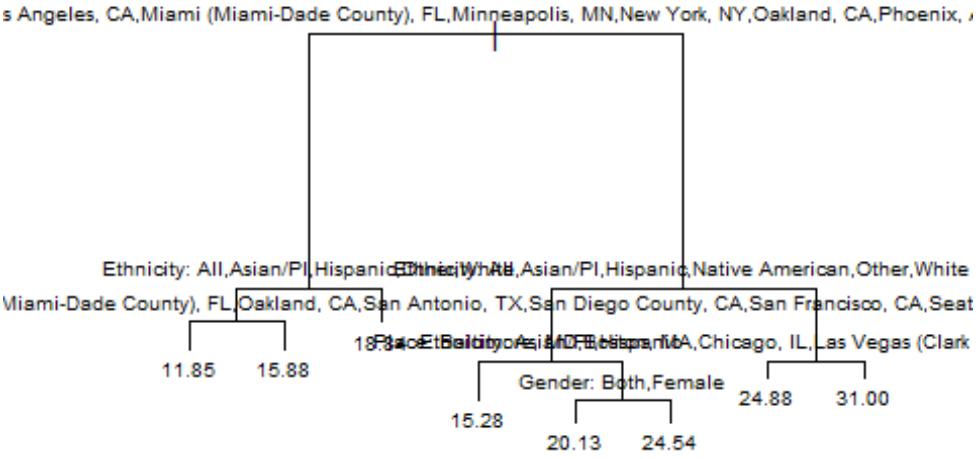
We will perform Pruning on the tree now.

Pruning:

```
cv.smoke.ad=cv.tree(tree.smoke.ad)
plot(cv.smoke.ad$size,cv.smoke.ad$dev,type='b')
```




```
prune.smoke.ad=prune.tree(tree.smoke.ad,best=8)
plot(prune.smoke.ad)
text(prune.smoke.ad,pretty=0, cex = 0.6)
```



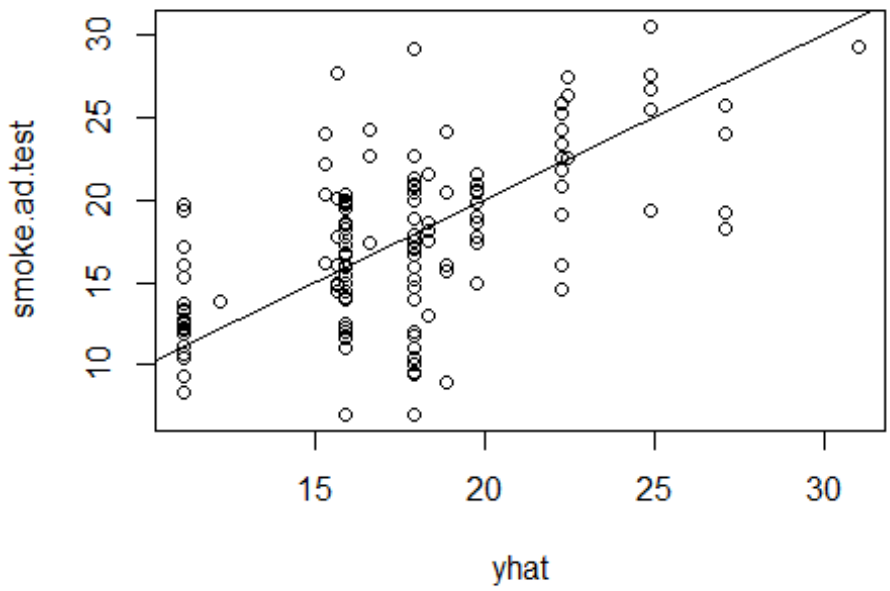
```
prune.smoke.ad

## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 140 5062.00 17.94
##    2) Place: Atlanta (Fulton County), GA,Fort Worth (Tarrant County), TX,L
ong Beach, CA,Los Angeles, CA,Miami (Miami-Dade County), FL,Minneapolis, MN,N
ew York, NY,Oakland, CA,Phoenix, AZ,San Antonio, TX,San Diego County, CA,San
Francisco, CA,San Jose, CA,Seattle, WA 65 1135.00 14.00
##      4) Ethnicity: All,Asian/PI,Hispanic,Other,White 56  664.40 13.22
##        8) Place: Atlanta (Fulton County), GA,Los Angeles, CA,Miami (Miami-
Dade County), FL,Oakland, CA,San Antonio, TX,San Diego County, CA,San Francis
co, CA,Seattle, WA 37  314.60 11.85 *
##        9) Place: Fort Worth (Tarrant County), TX,Long Beach, CA,Minneapoli
s, MN,New York, NY,Phoenix, AZ,San Jose, CA 19  145.40 15.88 *
##      5) Ethnicity: Black,Multiracial,Native American 9  225.50 18.84 *
##    3) Place: Baltimore, MD,Boston, MA,Chicago, IL,Denver, CO,Las Vegas (Cl
ark County), NV,Philadelphia, PA,U.S. Total,Washington, DC 75 2039.00 21.36
##      6) Ethnicity: All,Asian/PI,Hispanic,Native American,Other,White 62 12
58.00 20.13
##       12) Ethnicity: Asian/PI,Hispanic 10  203.90 15.28 *
##       13) Ethnicity: All,Native American,Other,White 52  773.40 21.06
##         26) Gender: Both,Female 41  522.10 20.13 *
##         27) Gender: Male 11   82.75 24.54 *
##      7) Ethnicity: Black,Multiracial 13  239.50 27.23
##       14) Place: Baltimore, MD,Boston, MA,Chicago, IL,Las Vegas (Clark Cou
nty), NV,Philadelphia, PA 8   67.27 24.88 *
##       15) Place: Denver, CO,Washington, DC 5   56.80 31.00 *
```

Interpretation:

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.

Making Predictions on test data:

```
yhat=predict(tree.smoke.ad,newdata=smoke.ad[-train,])
smoke.ad.test = smoke.ad[-train,"Value"]
plot(yhat,smoke.ad.test)
abline(0,1)
```



Mean Squared Error

```
mean((yhat-smoke.ad.test)^2)
```

```
## [1] 18.30589
```

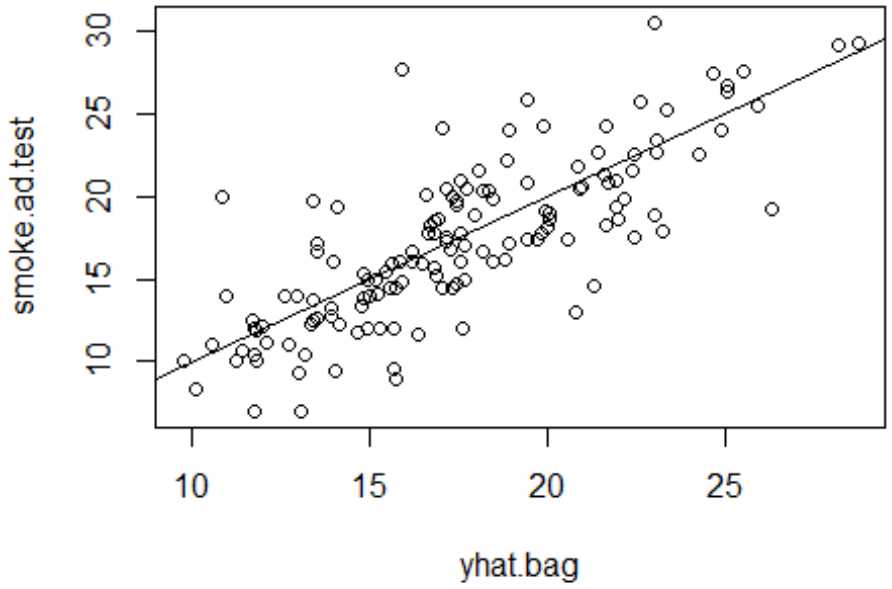
The error rate is quite good and we can also implement Bagging, Random Forest or Boosting to see if we can get n even better MSE.

Bagging

```
require(randomForest)
set.seed(1)
bag.smoke.ad=randomForest(Value~.,data=smoke.ad,subset=train,mtry=4,ntree = 5
00, importance=TRUE)
bag.smoke.ad
```

```
##
## Call:
##  randomForest(formula = Value ~ ., data = smoke.ad, mtry = 4,      ntree =
500, importance = TRUE, subset = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 15.06826
##                     % Var explained: 58.32
```

```
yhat.bag = predict(bag.smoke.ad,newdata=smoke.ad[-train,])
plot(yhat.bag, smoke.ad.test)
abline(0,1)
```




```
mean((yhat.bag-smoke.ad.test)^2)

## [1] 9.478208
```

Bagging reduces the error rate significantly and it is computed as 9.48.

Random Forest:

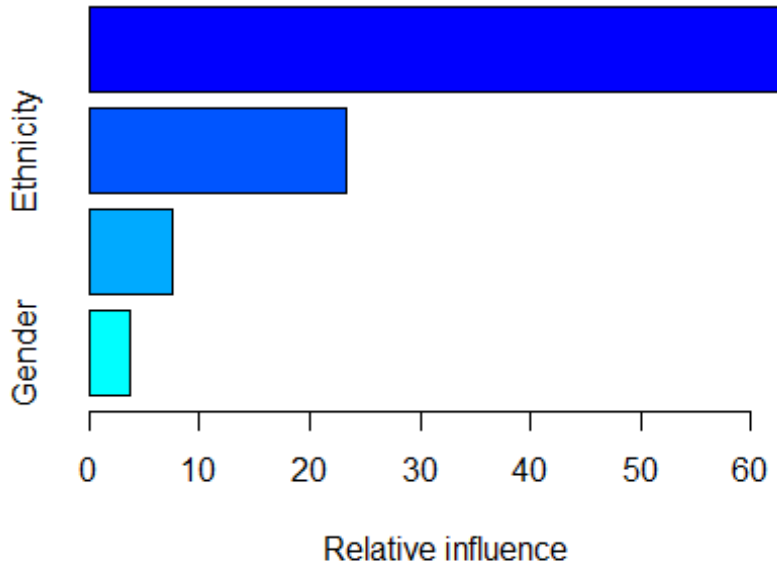
```
set.seed(1)
rf.smoke.ad=randomForest(Value~.,data=smoke.ad,subset=train,mtry=4,importance
=TRUE )
yhat.rf = predict(rf.smoke.ad,newdata=smoke.ad[-train,])
mean((yhat.rf-smoke.ad.test)^2)

## [1] 9.478208
```

Random Forest gives the same MSE as Bagging because both are equivalent in this case (due to same value of mtry).

Boosting

```
require(gbm)
set.seed(1)
boost.smoke.ad=gbm(Value~.,data=smoke.ad[train,],distribution="gaussian",n.tr
ees=5000,interaction.depth=4)
```

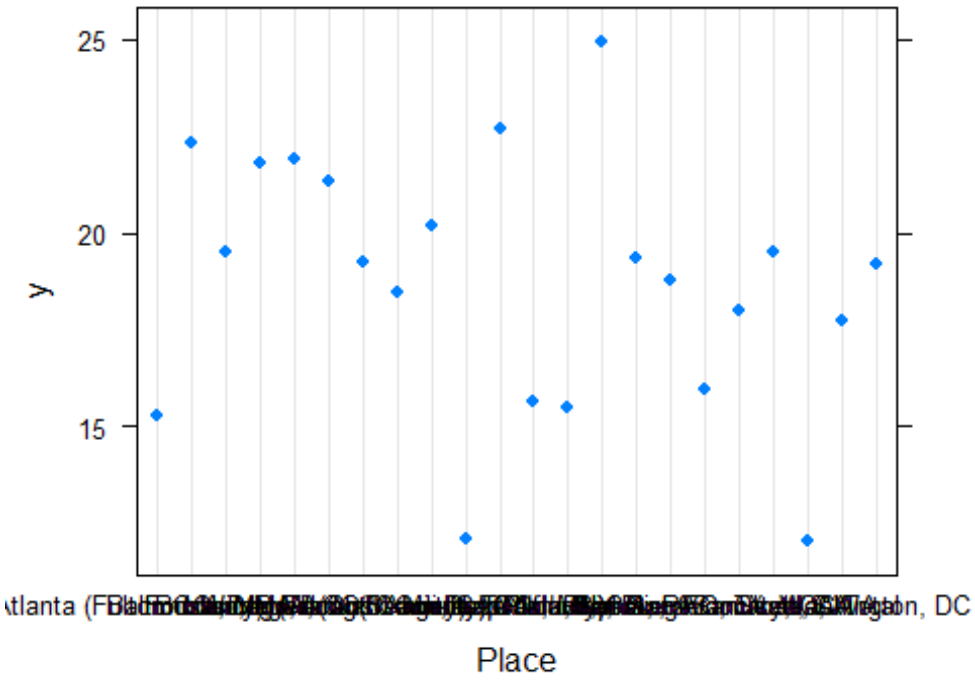
```
summary(boost.smoke.ad)
```



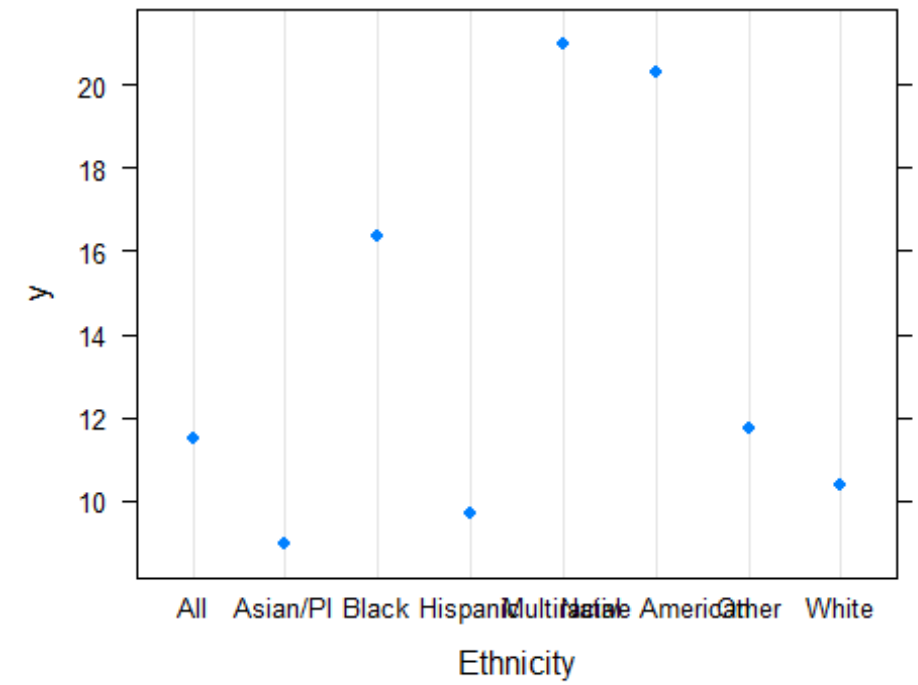

```
##                 var   rel.inf
## Place         Place 65.456142
## Ethnicity Ethnicity 23.368735
## Year           Year  7.552619
## Gender       Gender  3.622503
```

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.

```
par(mfrow=c(1,2))
plot(boost.smoke.ad,i="Place", type = "l")
```




```
plot(boost.smoke.ad,i="Ethnicity", type = "l")
```



```
yhat.boost=predict(boost.smoke.ad,newdata=smoke.ad[-train,],n.trees=5000)
mean((yhat.boost-smoke.ad.test)^2)

## [1] 17.29699

boost.smoke.ad=gbm(Value~.,data=smoke.ad[train,],distribution="gaussian",n.trees=5000,interaction.depth=4,shrinkage=0.2,verbose=F)
yhat.boost=predict(boost.smoke.ad,newdata=smoke.ad[-train,],n.trees=5000)
mean((yhat.boost-smoke.ad.test)^2)

## [1] 23.42302
```

The MSE when we perform Boosting is more than that of Bagging, 17.29 when we use default Shrinkage Parameter and 23.42 when the Shrinkage Parameter is increased to 0.2.

Therefore, we choose **Regression Tree with Bagging** as the best model as it generates the least MSE.

Importance of Variables:

```
importance(bag.smoke.ad)

##             %IncMSE IncNodePurity
## Year       3.485604      280.1958
## Gender    22.590533      339.1244
## Place     50.643009     2955.9471
## Ethnicity 32.568357     1245.3257

varImpPlot(bag.smoke.ad)
```

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.

Summary of Results:

Regression Tree analysis with Bagging gives us the best model for all datasets we examined - drink.st, smoke.st, drink.ad and smoke.ad.

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.

4. Additional Analyses

More Statistical Methods

We tried application of statistical methods apart from the ones described above like Logistic Regression, Linear Discriminant Analysis and K Nearest Neighbors. As most of the data is categorical we have not been able to utilize these methods in an effective manner.

i) Logistic Regression:

Applied logistic Regression on the High School students and Adults drinking problem data. As per initial analysis, there does not seem to be any effect of gender, value and Place on the disease. However, we need to investigate more to check for any significant relationships.

ii) Linear Discriminant Analysis: Applied Linear Discriminant analysis on the High School students and Adults drinking problem data. Looking at posterior probability, it is clear that there is uncertainty after we have sampled data. This is planned as a future step of analysis.

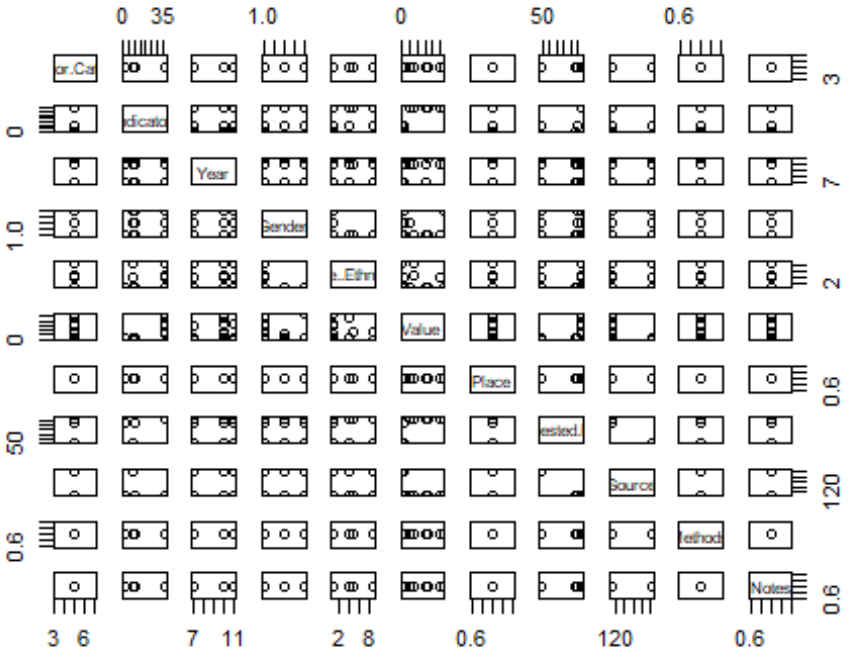
iii) KNN Classification: We employed the K-Nearest Neighbors method to check if predictions can be made for the Indicator Value given the data for previous years. However,

Value is a numeric variable and would need to be converted to categorical to make KNN classification work.

Disease Indicator Analysis

We also did some research and examined the dataset for significant factors influencing HIV/AIDS occurrences in major cities of the United States.

```
require(dplyr)
health_atlanta <- health.data %>%
  filter(Place == "Atlanta (Fulton County), GA")

health_atlanta_hiv <- health_atlanta %>%
  filter(Indicator.Category == "HIV/AIDS")

lm.fit_at_hiv <- lm(Value~Year, data=health_atlanta_hiv)
summary(lm.fit_at_hiv)

##
## Call:
## lm(formula = Value ~ Year, data = health_atlanta_hiv)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -595.5 -495.6 -202.1 -106.4 2084.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    211.2      265.4   0.796    0.433
## Year2012       386.9      328.6   1.177    0.249
## Year2013         8.8      375.3   0.023    0.981
##
## Residual standard error: 750.6 on 28 degrees of freedom
## Multiple R-squared:  0.06702,    Adjusted R-squared:  0.0003766
## F-statistic: 1.006 on 2 and 28 DF,  p-value: 0.3786

pairs(health_atlanta_hiv)
```




```
dataset<-health.data %>%
  filter(Year == 2013)%>%
  filter(Gender=="Both")%>%
  filter(Place == "Atlanta (Fulton County), GA")

lm.fit=lm(Value~Race..Ethnicity,data=dataset)
summary(lm.fit)

AIDS_mort_rate <- health.data %>%
  filter(Indicator == "HIV-Related Mortality Rate (Age-Adjusted; Per 100,000
people)")

pairs(AIDS_mort_rate)

lm.fit=lm(Value~Year,data=AIDS_mort_rate)
summary(lm.fit)

lm.fit=lm(Value~Gender,data=AIDS_mort_rate)
summary(lm.fit)

lm.fit=lm(Value~Race..Ethnicity,data=AIDS_mort_rate)
summary(lm.fit)

lm.fit=lm(Value~Place,data=AIDS_mort_rate)
summary(lm.fit)
```

5. Future Research

Analyze using more predictors: We have performed our analysis based on four variables which we feel are the most influential based on our Questions of Interest. All the remaining variables in our 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 statistical analysis.
Therefore, future research can be performed by processing these data columns and including these as explanatory variables in Linear Regression and Regression Tree analysis. Examples of these columns include Age(age of the people on whom the study has been conducted) and Methods (the methods used to obtain the Value).

Perform analysis on disease indicators: Our 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.

6. References

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

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