US-Presedential-Elections-Simple-Linear-Regression-Analyisis

Vidnyani Umathe

MS2320

The 2000 US presidential election saw intense competition between Republican George W. Bush and Democrat Al Gore. Notably close, the election's outcome was determined by the Supreme Court in Bush v.s. Gore. Florida became the focal point with a recount prompted by ballot irregularities, notably in Palm Beach County, where a controversial Butterfly ballot design (Pat Buchanan) skewed results. Despite Gore winning the popular vote by over 500,000, Bush secured the presidency with a narrow victory in Florida, garnering 271 electoral votes to Gore's 266.

So in this analysis we will see if the votes of Buchanan is linearly related to the votes of Bush or not.

Reading data from csv file

```
votes <- read.csv('C:\\Users\\hp\\Desktop\\ISSC NOTES AND ASSIGNMENTS\\Sem 2\\Machine Learning\\MA Assignment\\us-presidential-elections-2000.csv', header = TRUE, comment = '#' )
head(votes)

##     County   Bush   Gore Brow Nade Har Hag Buc Mc Ph  Mo
## 1  Alachua  34124  47365  658 3226   6  42 263  4 20  21
## 2    Baker   5610   2392   17   53   0   3  73  0  3   3
## 3      Bay  38637  18850  171  828   5  18 248  3 18  27
## 4 Bradford   5414   3075   28   84   0   2  65  0  2   3
## 5  Brevard 115185  97318  643 4470  11  39 570 11 72  76
## 6  Broward 177323 386561 1212 7101  50 129 788 34 74 124

rownames( votes ) <- votes[,1]  # use first column to set row names for the data frame
votes             <- votes[,-1] # remove first column
head(votes)

##            Bush   Gore Brow Nade Har Hag Buc Mc Ph  Mo
## Alachua   34124  47365  658 3226   6  42 263  4 20  21
## Baker      5610   2392   17   53   0   3  73  0  3   3
## Bay       38637  18850  171  828   5  18 248  3 18  27
## Bradford   5414   3075   28   84   0   2  65  0  2   3
## Brevard  115185  97318  643 4470  11  39 570 11 72  76
## Broward  177323 386561 1212 7101  50 129 788 34 74 124

n.cnddts <- ncol( votes ) # of candidates
n.cnts   <- nrow( votes ) # of counties
n.cnddts

## [1] 10

n.cnts
```

```
## [1] 67
```

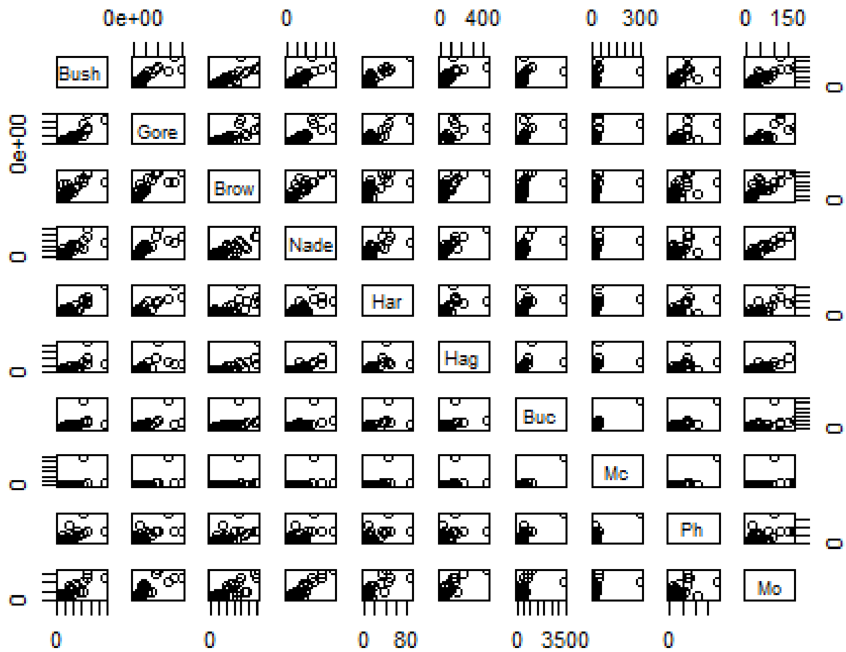
#Exploratory data analysis and visualization

```
summary(votes)
```

```
##       Bush             Gore             Brow             Nade        
##  Min.   :  1317   Min.   :   789   Min.   :   4.0   Min.   :   19.0  
##  1st Qu.:  4757   1st Qu.:  3058   1st Qu.:  23.5   1st Qu.:   95.5  
##  Median : 20206   Median : 14167   Median : 116.0   Median :  562.0  
##  Mean   : 43434   Mean   : 43420   Mean   : 244.7   Mean   : 1454.0  
##  3rd Qu.: 56547   3rd Qu.: 46015   3rd Qu.: 321.5   3rd Qu.: 1870.5  
##  Max.   :289492   Max.   :386561   Max.   :1230.0   Max.   :10022.0  
##       Har              Hag              Buc               Mc         
##  Min.   : 0.000   Min.   :  0.00   Min.   :   9.0   Min.   :  0.000  
##  1st Qu.: 1.000   1st Qu.:  3.00   1st Qu.:  46.5   1st Qu.:  1.000  
##  Median : 3.000   Median : 12.00   Median : 120.0   Median :  3.000  
##  Mean   : 8.328   Mean   : 33.93   Mean   : 260.7   Mean   :  9.224  
##  3rd Qu.: 8.000   3rd Qu.: 34.50   3rd Qu.: 285.5   3rd Qu.:  5.000  
##  Max.   :87.000   Max.   :442.00   Max.   :3407.0   Max.   :302.000  
##        Ph               Mo        
##  Min.   :  0.00   Min.   :  0.00  
##  1st Qu.:  3.00   1st Qu.:  4.00  
##  Median : 10.00   Median : 12.00  
##  Mean   : 20.42   Mean   : 26.91  
##  3rd Qu.: 20.00   3rd Qu.: 29.00  
##  Max.   :188.00   Max.   :170.00
```

Scatter Plot

```
plot(votes)
```

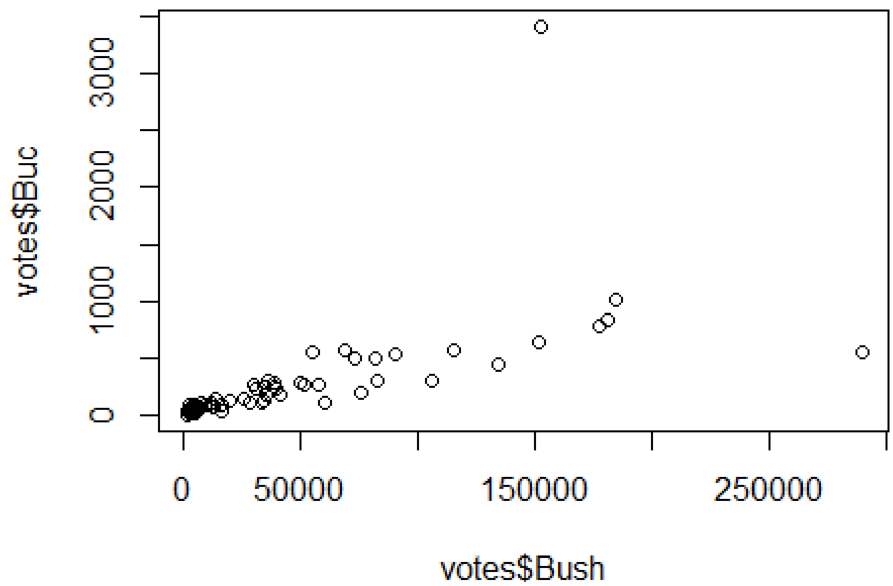



```
votes[,c("Bush","Buc")]
```

```
##                 Bush  Buc
## Alachua        34124  263
## Baker           5610   73
## Bay            38637  248
## Bradford        5414   65
## Brevard       115185  570
## Broward       177323  788
## Calhoun         2873   90
## Charlotte      35426  182
## Citrus         29765  270
## Clay           41736  186
## Collier        60433  122
## Columbia       10964   89
## Desoto          4256   36
## Dixie           2697   29
## Duval         152098  652
## Escambia       73017  502
## Flagler        12613   83
## Franklin        2454   33
## Gadsden         4767   38
## Gilchrist       3300   29
## Glades          1841    9
## Gulf            3550   71
## Hamilton        2146   23
## Hardee          3765   30
```

```
## Hendry          4747   22
## Hernando       30646  242
## Highlands      20206  127
## Hillsborough 180760  847
## Holmes          5011   76
## IndianRiver    28635  105
## Jackson         9138  102
## Jefferson       2478   29
## Lafayette       1670   10
## Lake           50010  289
## Lee           106141  305
## Leon           39053  282
## Levy            6858   67
## Liberty         1317   39
## Madison         3038   29
## Manatee        57952  271
## Marion         55141  563
## Martin         33970  112
## Miami-Dade    289492  560
## Monroe         16059   47
## Nassau         16280   90
## Okaloosa       52093  267
## Okeechobee      5057   43
## Orange        134517  446
## Osceola        26212  145
## PalmBeach     152846 3407
## Pasco          68582  570
## Pinellas      184823 1013
## Polk           90180  532
## Putnam         13447  148
## SantaRosa      36274  311
## Sarasota       83100  305
## Seminole       75677  194
## StJohns        39546  229
## StLucie        34705  124
## Sumter         12127  114
## Suwannee        8006  108
## Taylor          4056   27
## Union           2332   37
## Volusia        82214  496
## Wakulla         4512   46
## Walton         12182  120
## Washington      4994   88
```

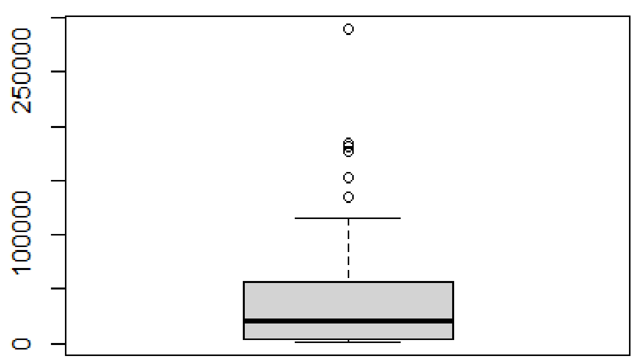
```
plot(votes$Bush,votes$Buc)
```



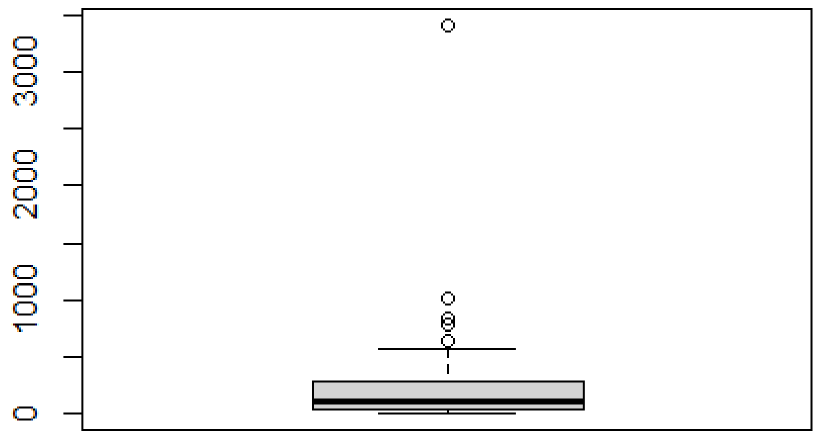
The scatter plot shows positive linear relation with two outliers.

BoxPlot

```
boxplot(votes[,"Bush"])
```



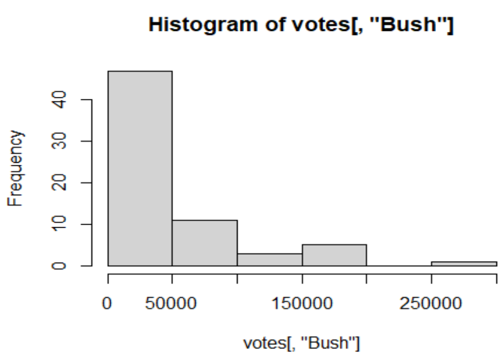
```r
boxplot(votes[,"Buc"])
```



The range of Buchanan's votes lies below 1000 with some outliers and one outlier being above 3000.

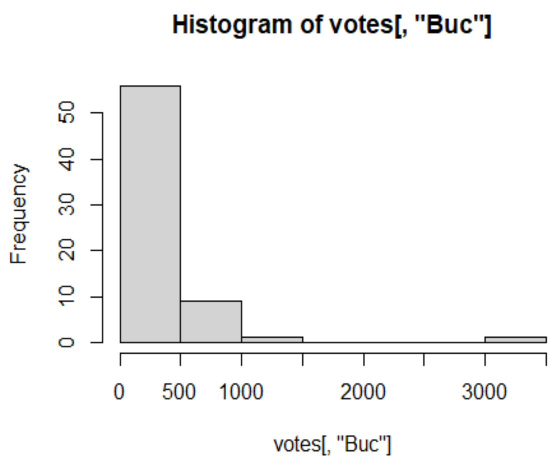
Histogram

```r
hist(votes[,"Bush"])
```



The histogram is positively skewed for Bush's votes this means most of the votes count are around 50000.

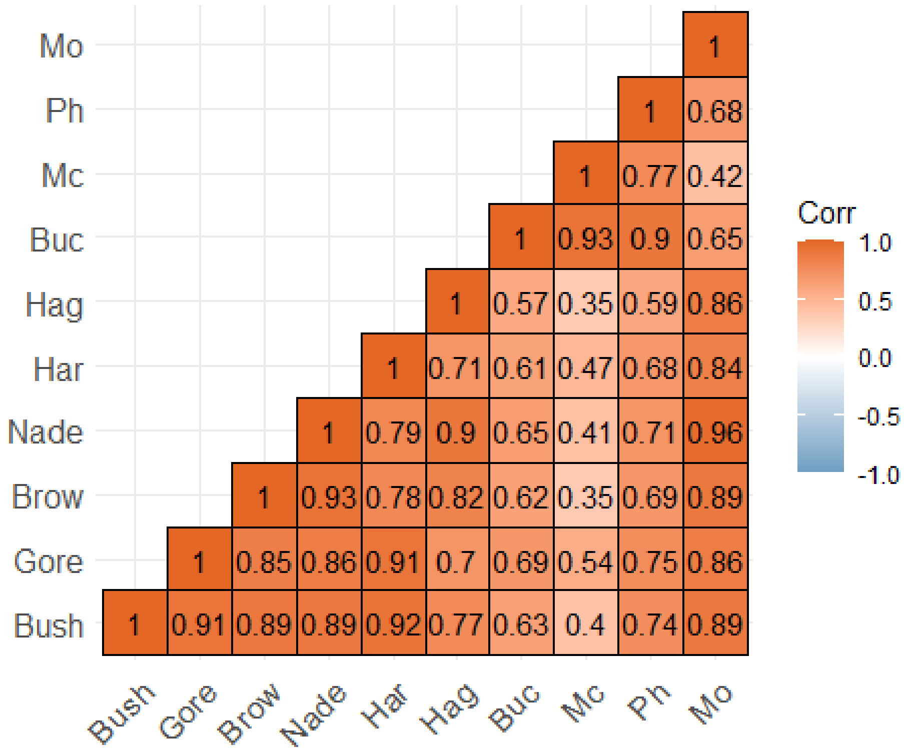
```r
hist(votes[,"Buc"])
```



The histogram is positively skewed for Buc's votes as well this means most of the votes count are around 1000.

Correlation matrix

```r
library(ggcorrplot)

## Loading required package: ggplot2

corr<-cor(votes)
head(corr)

##           Bush      Gore      Brow      Nade       Har       Hag       Buc
## Bush 1.0000000 0.9128003 0.8945743 0.8921811 0.9213028 0.7708869 0.6250012
## Gore 0.9128003 1.0000000 0.8468092 0.8646948 0.9132500 0.6999896 0.6903406
## Brow 0.8945743 0.8468092 1.0000000 0.9301379 0.7785174 0.8214665 0.6175750
## Nade 0.8921811 0.8646948 0.9301379 1.0000000 0.7942726 0.8958150 0.6540457
## Har  0.9213028 0.9132500 0.7785174 0.7942726 1.0000000 0.7136787 0.6087694
## Hag  0.7708869 0.6999896 0.8214665 0.8958150 0.7136787 1.0000000 0.5738050
##             Mc        Ph        Mo
## Bush 0.4048047 0.7431776 0.8948772
## Gore 0.5381894 0.7473092 0.8606041
## Brow 0.3546431 0.6900022 0.8865643
## Nade 0.4118872 0.7078291 0.9578779
## Har  0.4746005 0.6785084 0.8430439
## Hag  0.3507640 0.5936235 0.8620299

#ggcorrplot(corr)
ggcorrplot(corr,type = "lower", outline.color = "black",show.diag =TRUE,lab =
TRUE, ggtheme = ggplot2::theme_minimal , colors =
c("#6D9EC1","white","#E46726"))
```



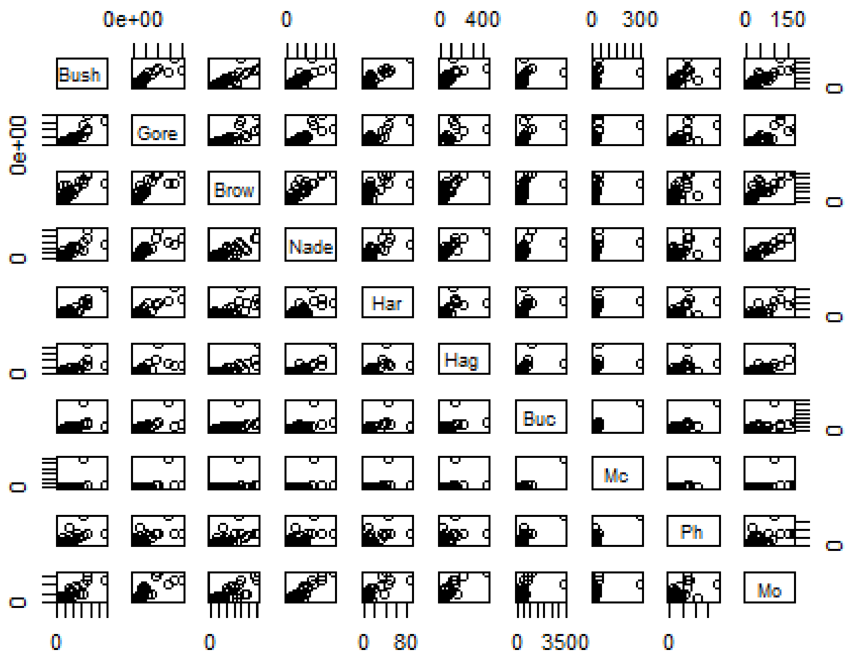
There is heteroskedasticity in the data so let's transform it with log and make it uniform.

Log Transformation

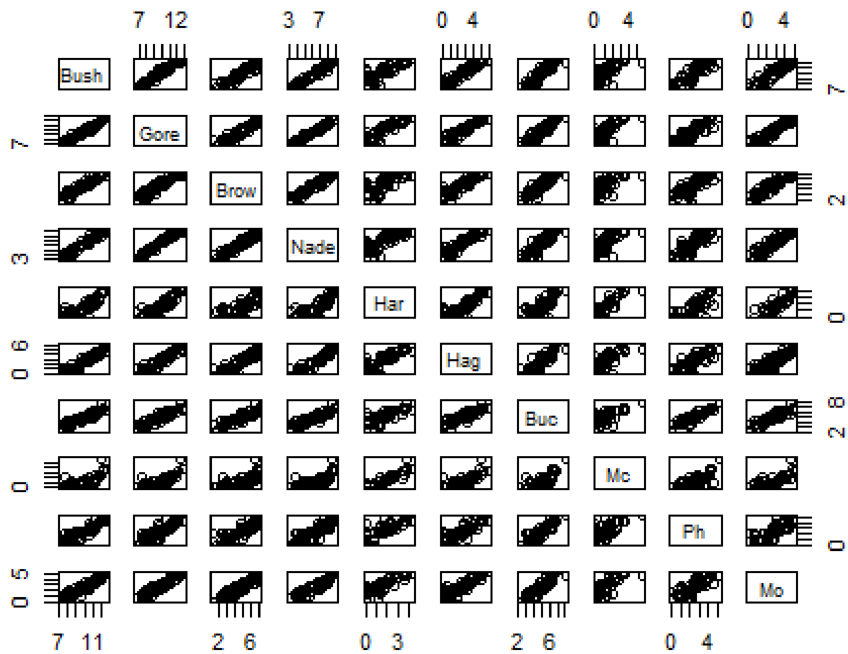
```
log.votes <- log( votes )
head(log.votes)

##                Bush      Gore     Brow     Nade      Har       Hag      Buc
## Alachua   10.437756 10.765639 6.489205 8.078998 1.791759 3.7376696 5.572154
## Baker      8.632306  7.779885 2.833213 3.970292     -Inf 1.0986123 4.290459
## Bay       10.561966  9.844268 5.141664 6.719013 1.609438 2.8903718 5.513429
## Bradford   8.596743  8.031060 3.332205 4.430817     -Inf 0.6931472 4.174387
## Brevard   11.654295 11.485739 6.466145 8.405144 2.397895 3.6635616 6.345636
## Broward   12.085728 12.865045 7.100027 8.867991 3.912023 4.8598124 6.669498
##                 Mc        Ph       Mo
## Alachua   1.386294 2.9957323 3.044522
## Baker         -Inf 1.0986123 1.098612
## Bay       1.098612 2.8903718 3.295837
## Bradford      -Inf 0.6931472 1.098612
## Brevard   2.397895 4.2766661 4.330733
## Broward   3.526361 4.3040651 4.820282
```

pair plots

```
pairs( votes )
```




```
pairs( log.votes )
```

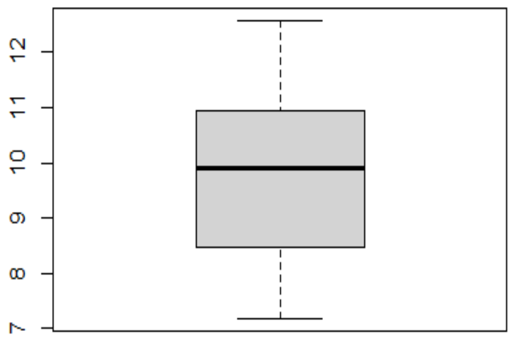



```r
summary(log.votes)
```

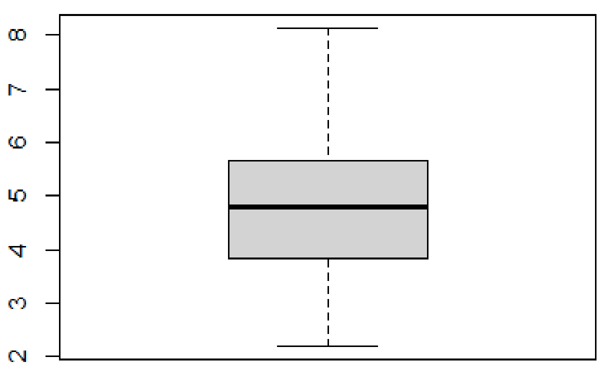
```
##        Bush             Gore             Brow            Nade      
##  Min.   : 7.183   Min.   : 6.671   Min.   :1.386   Min.   :2.944  
##  1st Qu.: 8.467   1st Qu.: 8.026   1st Qu.:3.157   1st Qu.:4.559  
##  Median : 9.914   Median : 9.559   Median :4.754   Median :6.332  
##  Mean   : 9.782   Mean   : 9.521   Mean   :4.564   Mean   :6.174  
##  3rd Qu.:10.943   3rd Qu.:10.736   3rd Qu.:5.773   3rd Qu.:7.533  
##  Max.   :12.576   Max.   :12.865   Max.   :7.115   Max.   :9.213  
##       Har              Hag              Buc              Mc        
##  Min.   : -Inf   Min.   : -Inf   Min.   :2.197   Min.   : -Inf  
##  1st Qu.:0.000   1st Qu.:1.099   1st Qu.:3.839   1st Qu.:0.000  
##  Median :1.099   Median :2.485   Median :4.787   Median :1.099  
##  Mean   : -Inf   Mean   : -Inf   Mean   :4.846   Mean   : -Inf  
##  3rd Qu.:2.079   3rd Qu.:3.541   3rd Qu.:5.654   3rd Qu.:1.609  
##  Max.   :4.466   Max.   :6.091   Max.   :8.134   Max.   :5.710  
##        Ph               Mo       
##  Min.   : -Inf   Min.   : -Inf  
##  1st Qu.:1.099   1st Qu.:1.386  
##  Median :2.303   Median :2.485  
##  Mean   : -Inf   Mean   : -Inf  
##  3rd Qu.:2.996   3rd Qu.:3.367  
##  Max.   :5.236   Max.   :5.136  
```

BoxPlot

```
boxplot(log.votes[,"Bush"])
```

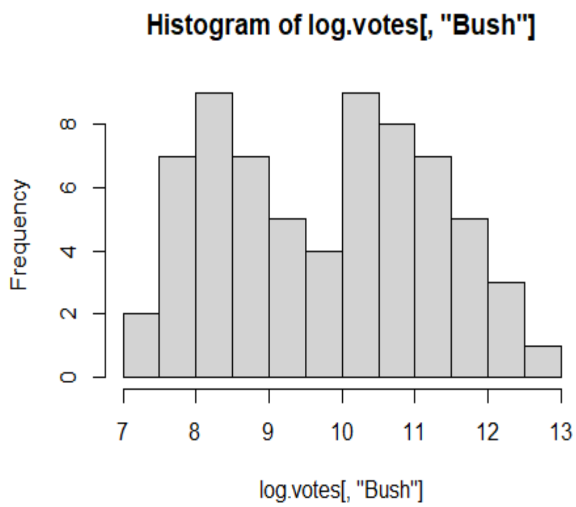



```
boxplot(log.votes[,"Buc"])
```

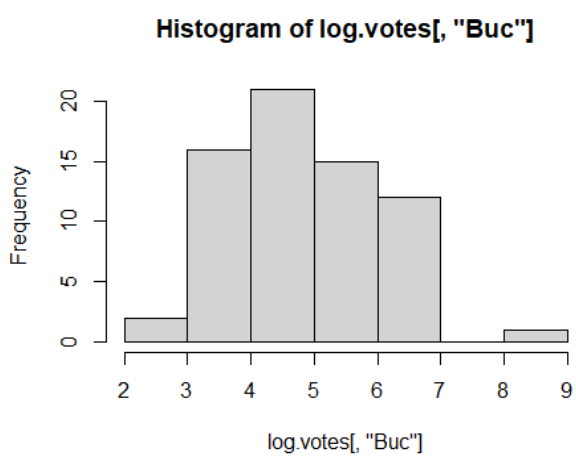


Histogram

```
hist(log.votes[,"Bush"])
```




```
hist(log.votes[,"Buc"])
```



From the box-plots and the histograms we can see that the data is now somewhere normally distributed as it gives a kind of bell-shaped curve.

Buc~Bush

```r
bush     <- votes[,'Bush']
buc      <- votes[,'Buc']
log.bush <- log( bush )
log.buc  <- log( buc )

# to find indices of two extreme outliers in the buc ~ bush data
out <- c( which( bush > 200000 ), which( buc  > 2000 ) )
out

## [1] 43 50
```

Correlations

```r
#with outliers
cor.test( bush, buc )

##
##  Pearson's product-moment correlation
##
## data:  bush and buc
## t = 6.455, df = 65, p-value = 1.574e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4527668 0.7522709
## sample estimates:
##       cor
## 0.6250012

cor.test( log.bush, log.buc )

##
##  Pearson's product-moment correlation
##
## data:  log.bush and log.buc
## t = 19.222, df = 65, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8760098 0.9515894
## sample estimates:
##       cor
## 0.9221706



#without outliers
cor.test( bush[-out], buc[-out] )

##
##  Pearson's product-moment correlation
##
## data:  bush[-out] and buc[-out]
```

```
## t = 20.106, df = 63, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8876142 0.9569506
## sample estimates:
##       cor
## 0.9301473
```

```r
cor.test( log.bush[-out], log.buc[-out] )
```

```
##
##  Pearson's product-moment correlation
##
## data:  log.bush[-out] and log.buc[-out]
## t = 20.094, df = 63, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8874909 0.9569016
## sample estimates:
##       cor
## 0.9300689
```

The correlation between Buc and Bush is **0.6250012** for original data and **0.9221706** for log transformed data if we consider outliers. Whereas, if we do not consider outliers then the correlation comes out to be almost same for original data and log data which is **0.9301473** and **0.9300689** respectively.

Therefore, we reject the null-hypothesis and interpret that votes of Buc and Bush are highly positively correlated.

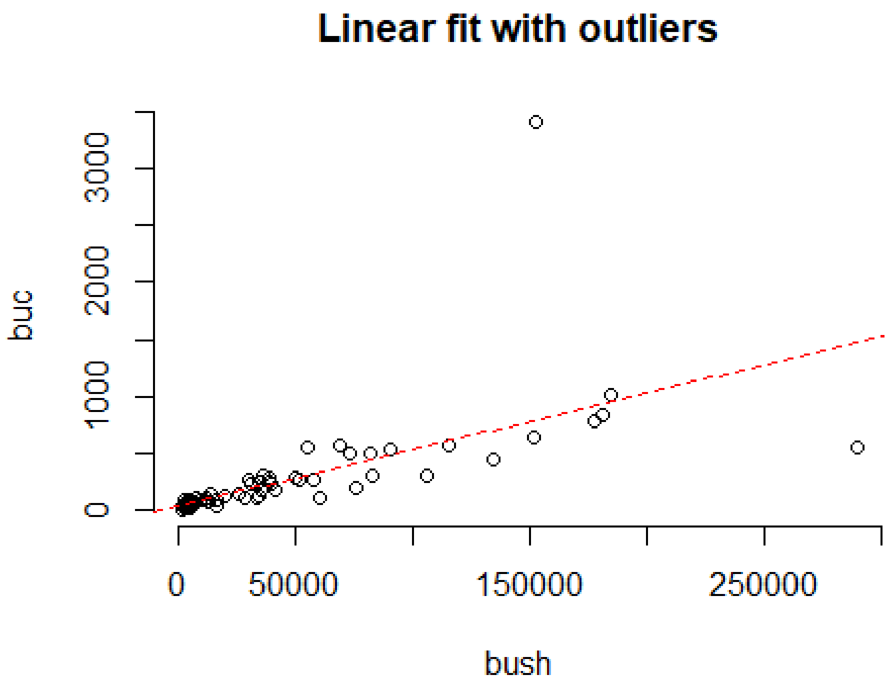
Buc~Bush Linear fits

```r
fit1 <- lm( buc ~ bush )     # with outliers
summary(fit1)
```

```
##
## Call:
## lm(formula = buc ~ bush)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -911.30  -46.11  -26.05   12.01 2608.01
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.697e+01  5.446e+01   0.863    0.392
## bush        4.920e-03  7.622e-04   6.455 1.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 353.9 on 65 degrees of freedom
## Multiple R-squared:  0.3906, Adjusted R-squared:  0.3813
## F-statistic: 41.67 on 1 and 65 DF,  p-value: 1.574e-08
```

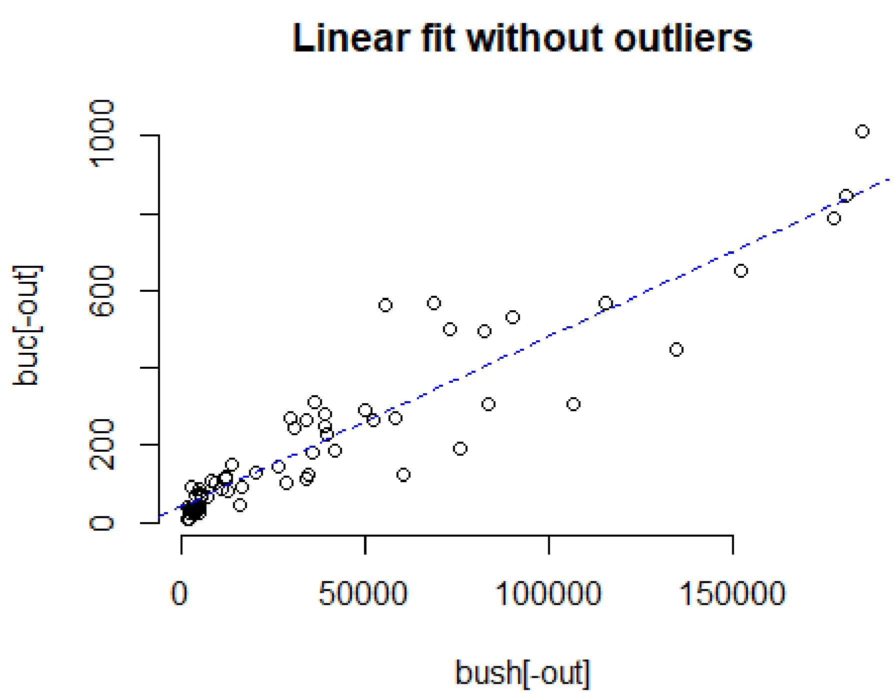
```
plot( bush, buc, bty = 'n', main = 'Linear fit with outliers' )
abline( coef( fit1 ), lty = 2 ,col='red')
```




```
fit2 <- lm( buc[-out] ~ bush[-out] )  # without outliers
summary(fit2)

##
## Call:
## lm(formula = buc[-out] ~ bush[-out])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -204.09  -26.45  -10.70   26.33  279.40
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.981e+01  1.322e+01   3.012  0.00374 **
## bush[-out]  4.421e-03  2.199e-04  20.106  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.61 on 63 degrees of freedom
```

```
## Multiple R-squared:  0.8652, Adjusted R-squared:  0.863
## F-statistic: 404.3 on 1 and 63 DF,  p-value: < 2.2e-16
```

```
plot( bush[-out], buc[-out], bty = 'n', main = 'Linear fit without outliers'
)
abline( coef( fit2 ), lty = 2 ,col='blue')
```



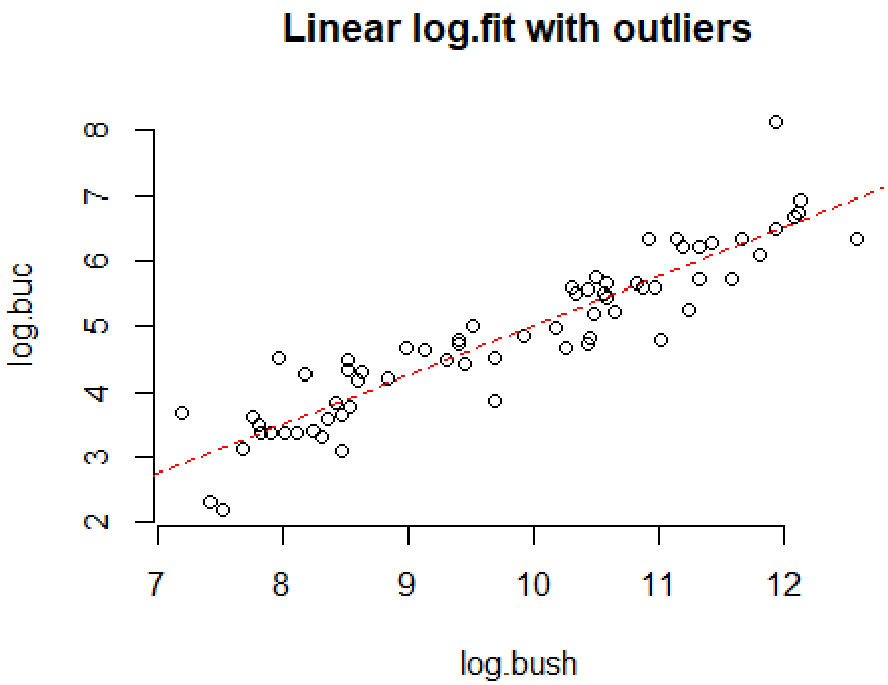
log(Buc) ~ log(Bush) linear fits

```
# with outliers
log.fit1 <- lm( log.buc ~ log.bush )
summary( log.fit1 )
```

```
##
## Call:
## lm(formula = log.buc ~ log.bush)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97038 -0.24247  0.00825  0.25452  1.65752
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.55079    0.38903  -6.557 1.04e-08 ***
## log.bush     0.75620    0.03934  19.222  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4672 on 65 degrees of freedom
## Multiple R-squared:  0.8504, Adjusted R-squared:  0.8481
## F-statistic: 369.5 on 1 and 65 DF,  p-value: < 2.2e-16
```

```
plot( log.bush, log.buc, bty = 'n', main = 'Linear log.fit with outliers' )
abline( coef( log.fit1 ), lty = 2 ,col='red')
```

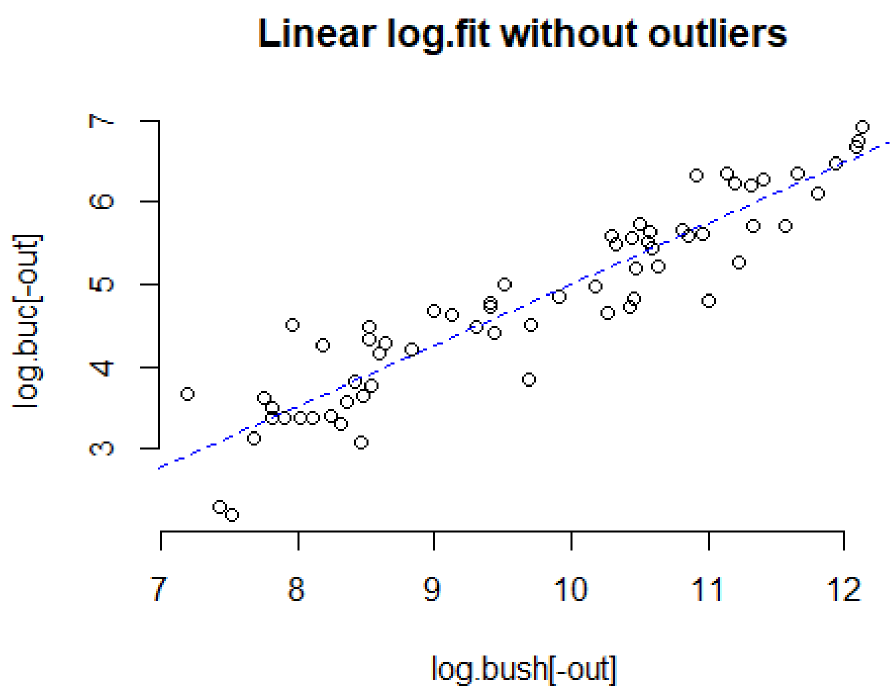



```
# without outliers
log.fit2 <- lm( log.buc[-out] ~ log.bush[-out] )
summary( log.fit2 )
```

```
##
## Call:
## lm(formula = log.buc[-out] ~ log.bush[-out])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95353 -0.21862  0.01486  0.25651  1.01906
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.4237     0.3619  -6.698 6.71e-09 ***
## log.bush[-out]   0.7415     0.0369  20.094  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4178 on 63 degrees of freedom
```

```
## Multiple R-squared:  0.865,  Adjusted R-squared:  0.8629
## F-statistic: 403.8 on 1 and 63 DF,  p-value: < 2.2e-16
```

```
plot( log.bush[-out], log.buc[-out], bty = 'n', main = 'Linear log.fit
without outliers' )
abline( coef( log.fit2 ), lty = 2 ,col='blue')
```



Conclusion:

Log transformed data gives more accurate fit than the original data.

Moreover, the votes of Bush are positively linearly related with the votes of Buchanan.