Regression & Analysis of Variance

Aymen Rumi

4/26/2020

Overview

We will analyze data from 3 distinct datasets **Abalone**, **Cigs**, **BirthWeight**, we will make hypothesis and present observations & interpret results from findings from our analysis

Abalone:

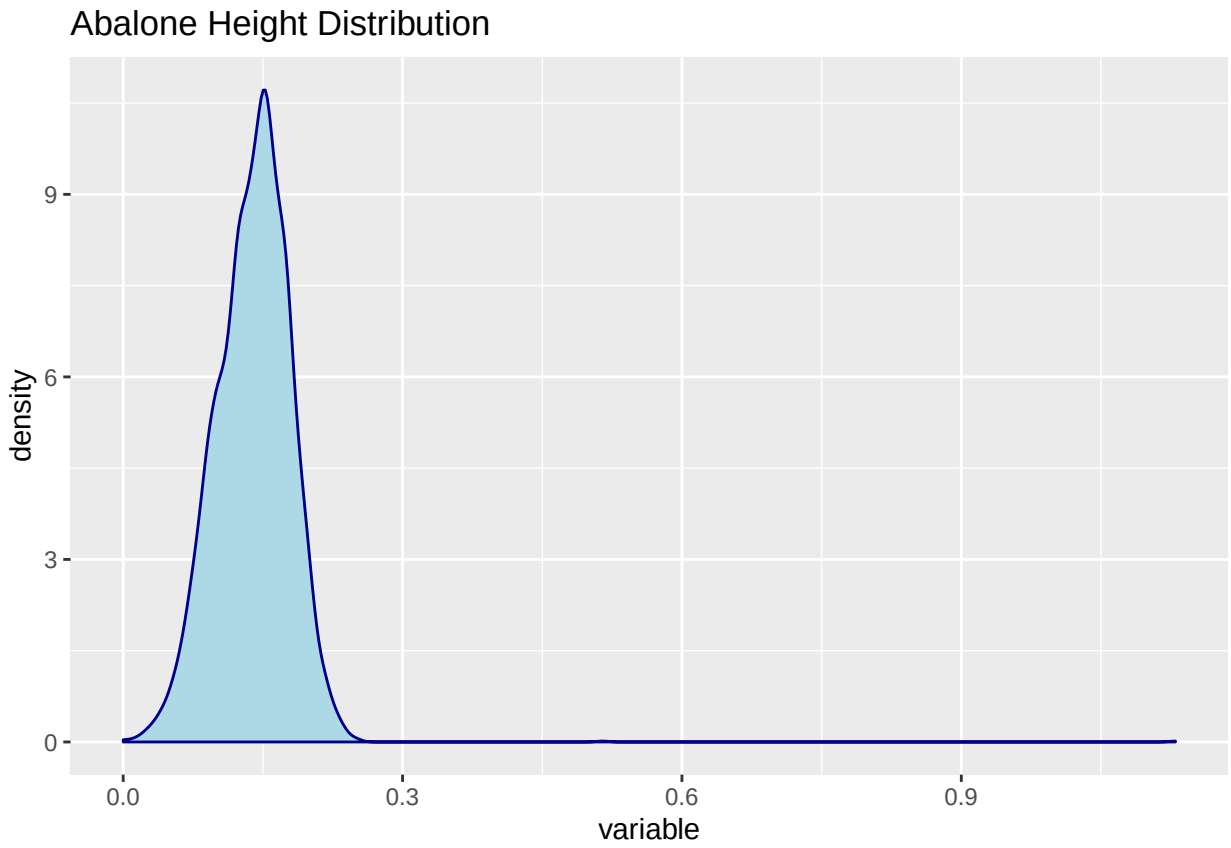
Hypothesis: I believe that a simple linear regression model with normal error assumption is appropriate to describe the relationship between the height of abalones and their ages, and particularly, that a larger height is associated with an older age, we will use data from `abalone.csv` to test this hypotheis

```
# importing data
file1 <- "http://www.math.mcgill.ca/yyang/regression/data/abalone.csv"
abalone <- read.csv(file1, header = TRUE)
```

```
# functions for summary statistics

Summary_Table<-function(data,variable)
{
    data %>% summarise(Avg = mean(variable),
        Med = median(variable),
        Q25 = quantile(variable,0.25), Q75 = quantile(variable,0.75),
        StD = sd(variable), Var=var(variable), Min=min(variable),
        Max=max(variable))%>%kable()
}

Plot_Distribution<-function(data,variable,title="")
{
  ggplot(data, aes(x=variable))+geom_density(color="darkblue", fill="lightblue")+ggtitle(title)
}
```

Univariate Analysis: Height

```
Plot_Distribution(abalone,abalone$Height,"Abalone Height Distribution")
```

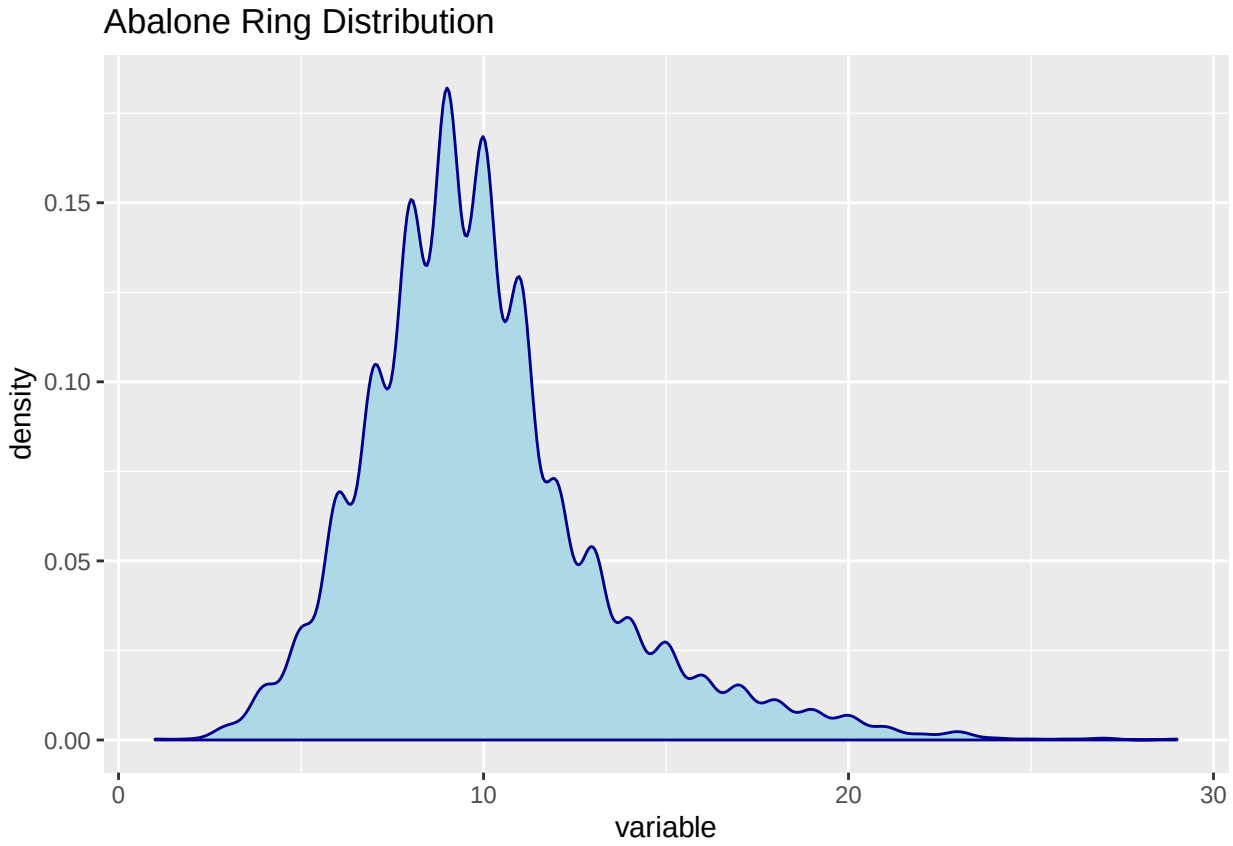



```
Summary_Table(abalone,abalone$Height)
```

Avg	Med	Q25	Q75	StD	Var	Min	Max
0.1395164	0.14	0.115	0.165	0.0418271	0.0017495	0	1.13

Univariate Analysis: Ring

```
Plot_Distribution(abalone,abalone$Rings,"Abalone Ring Distribution")
```

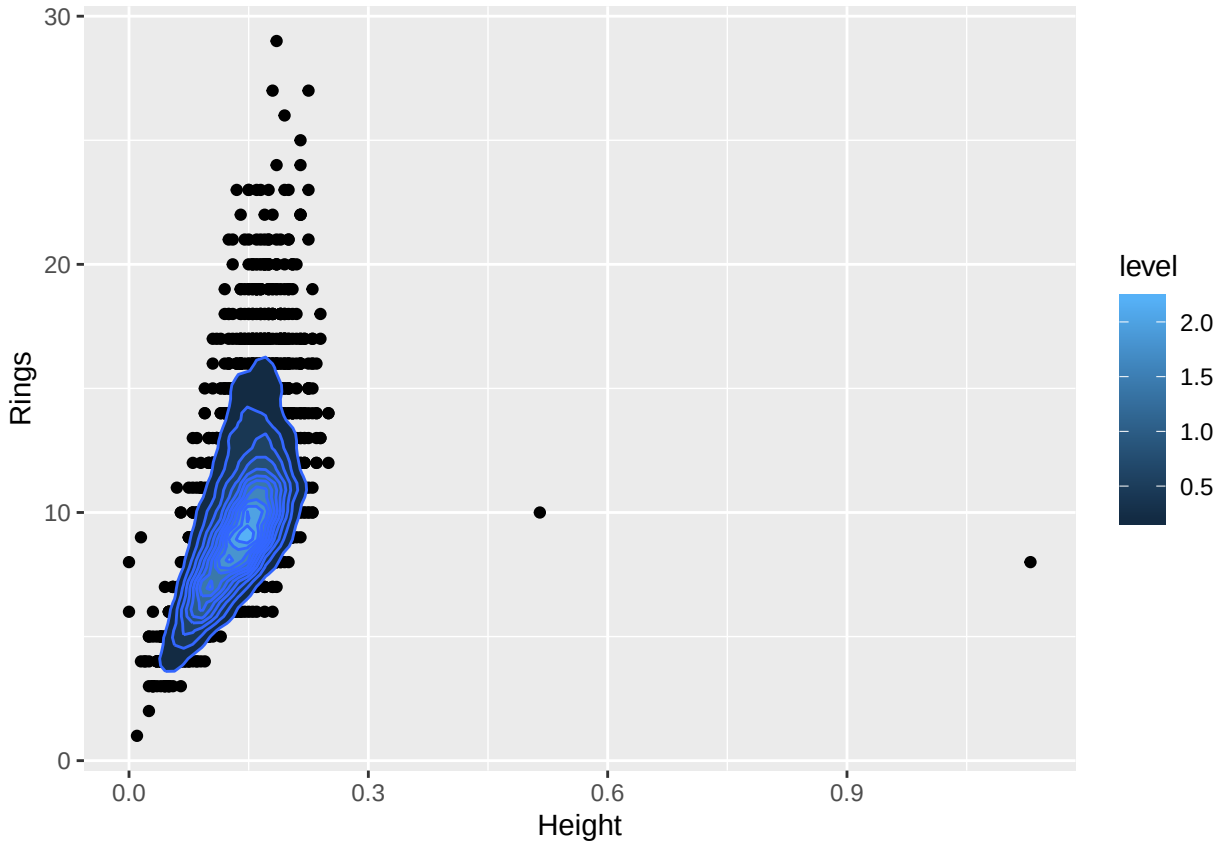



```
Summary_Table(abalone,abalone$Rings)
```

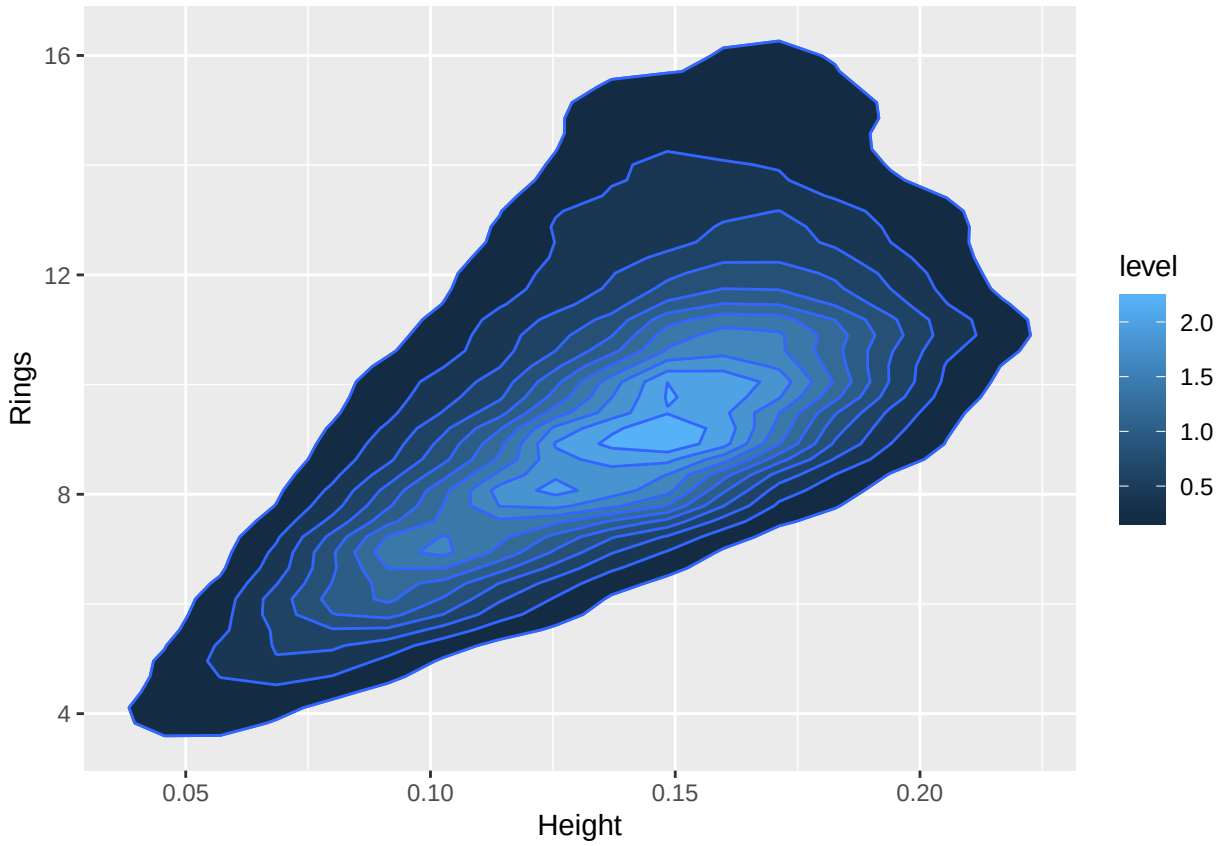
Avg	Med	Q25	Q75	StD	Var	Min	Max
9.933685	9	8	11	3.224169	10.39527	1	29

Bivariate Analysis: Height vs Rings

```
ggplot(abalone,aes(x=Height,y=Rings))+geom_point()+stat_density_2d(aes(fill = ..level..), geom = "polyg
```

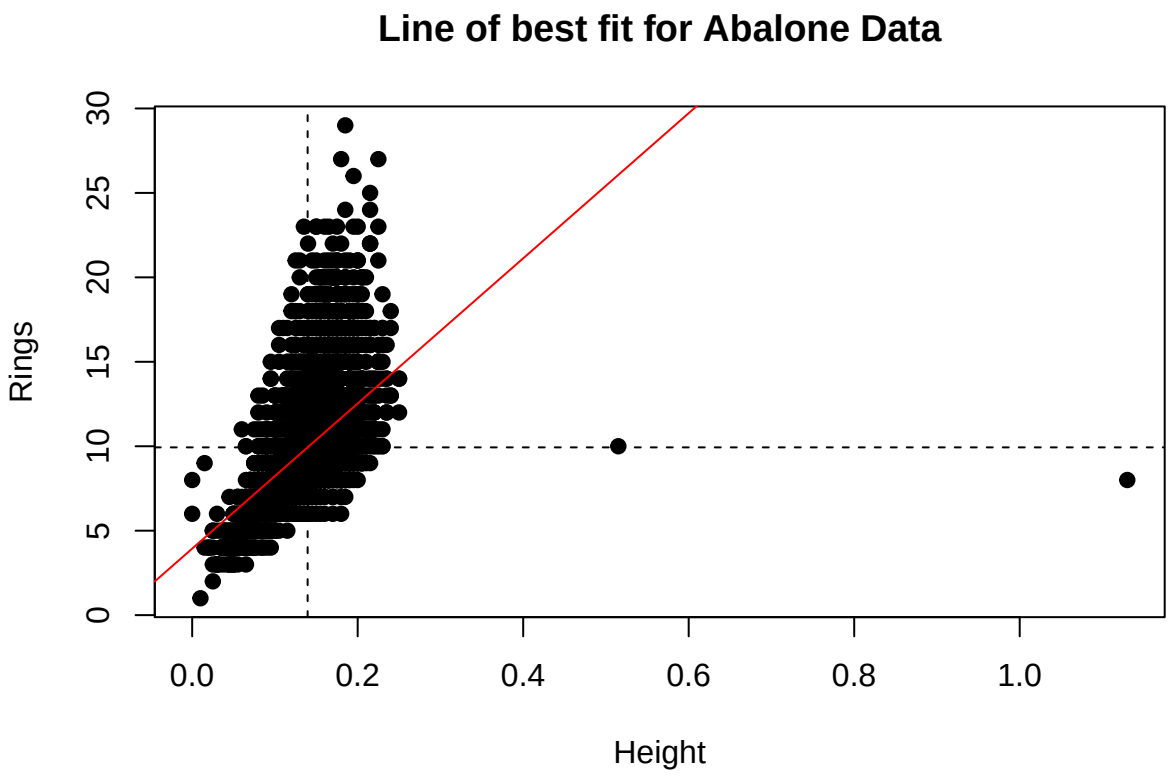



```
ggplot(abalone,aes(x=Height,y=Rings,fill = ..level..), geom = "polygon")+geom_density_2d()+stat_density
```



Fitting Linear Model

```
plot(abalone$Height,abalone$Rings,pch=19,xlab='Height',ylab='Rings')

abline(v=mean(abalone$Height),h=mean(abalone$Rings),lty=2)

fit.RP<-lm(abalone$Rings~abalone$Height)

title('Line of best fit for Abalone Data')

abline(coef(fit.RP),col='red')
```

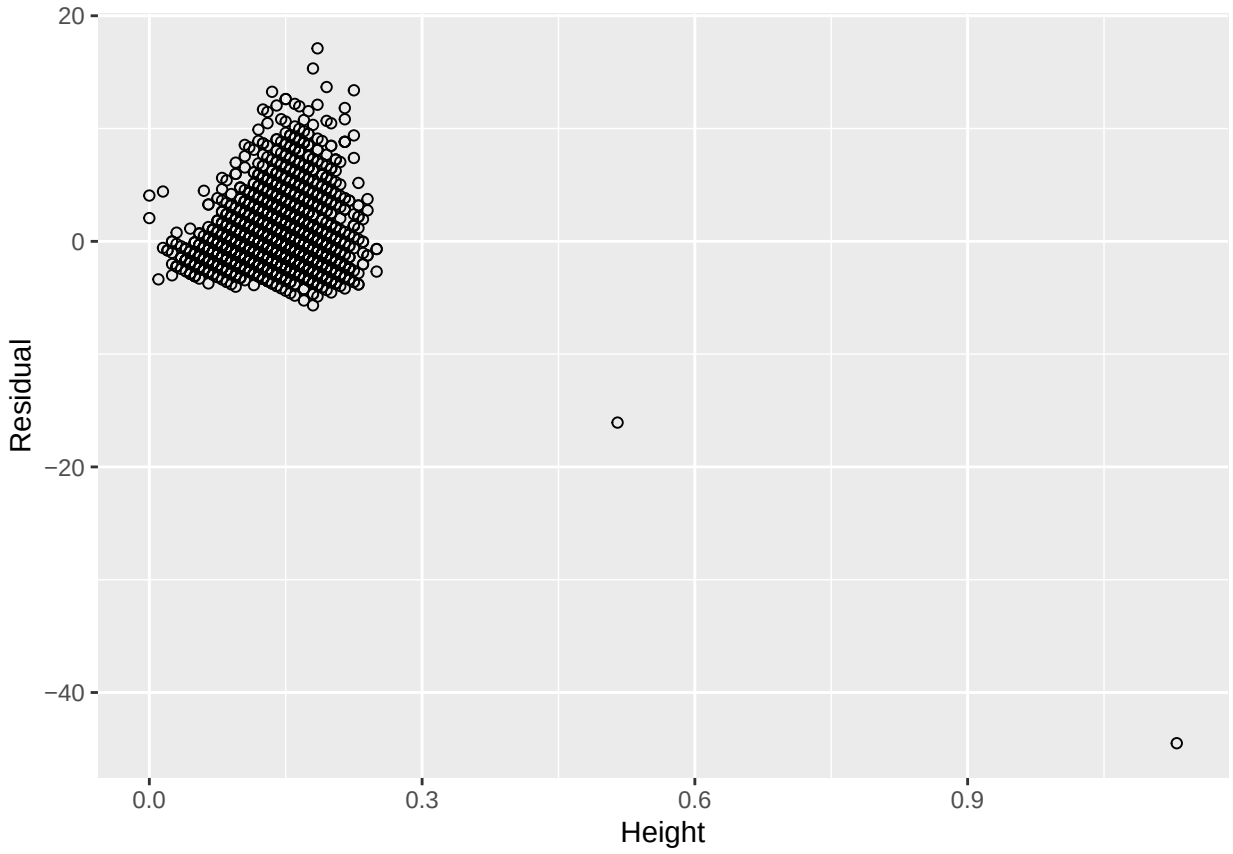



```
summary(fit.RP)
```

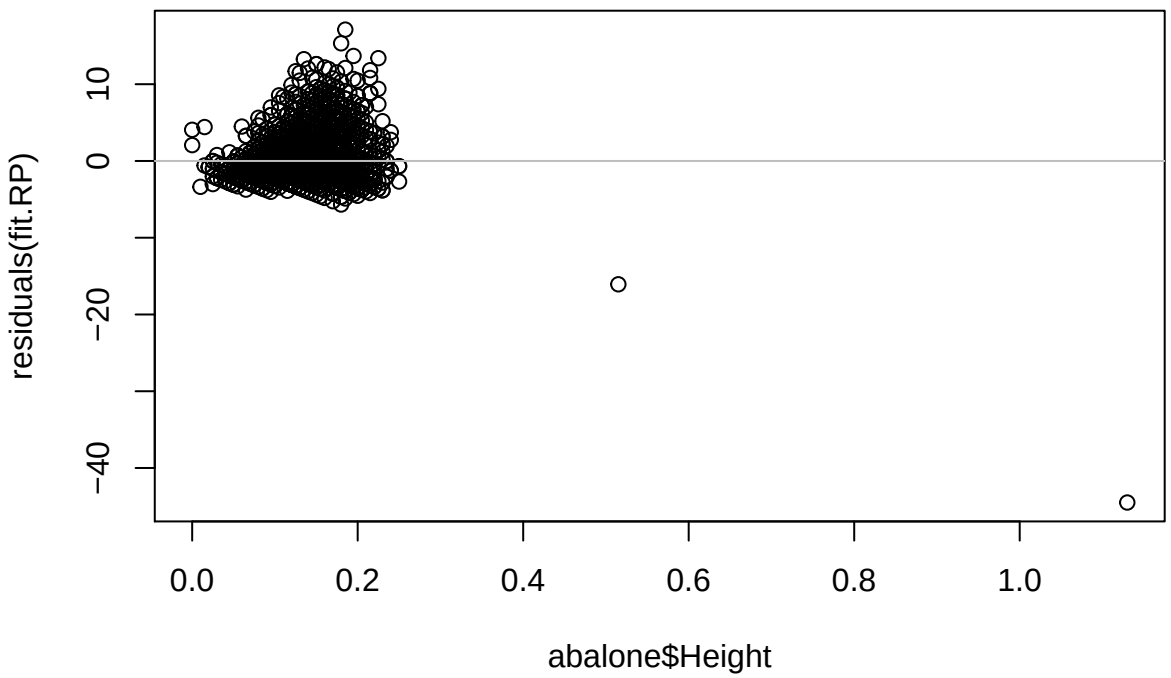
```
##
## Call:
## lm(formula = abalone$Rings ~ abalone$Height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.496  -1.657  -0.607   0.839  17.112
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.9385     0.1443   27.30   <2e-16 ***
## abalone$Height  42.9714     0.9904   43.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.677 on 4175 degrees of freedom
## Multiple R-squared:  0.3108, Adjusted R-squared:  0.3106
## F-statistic:  1882 on 1 and 4175 DF,  p-value: < 2.2e-16
```

Model Aquecuacy Checking (Residual Analysis) & Diagnostic

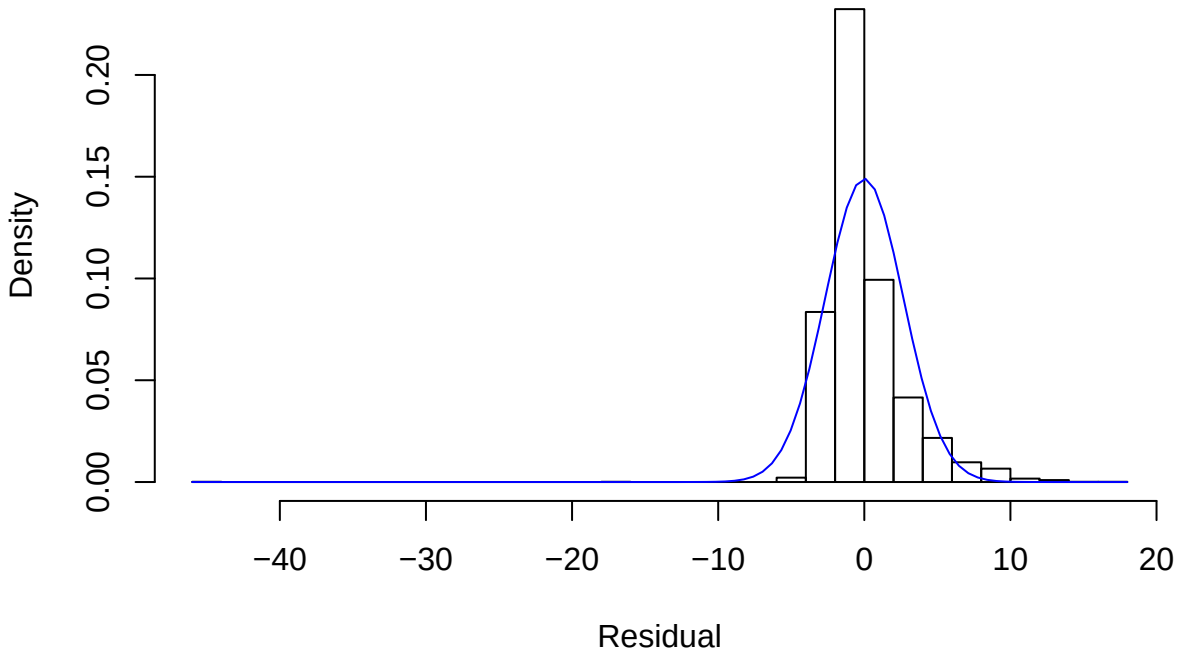
```
ggplot(data = data.frame(x = abalone$Height, y = residuals(fit.RP)), aes(x = x, y = y))+geom_point(shap
```




```
plot(abalone$Height,residuals(fit.RP))
abline(h=0,col="gray")
```

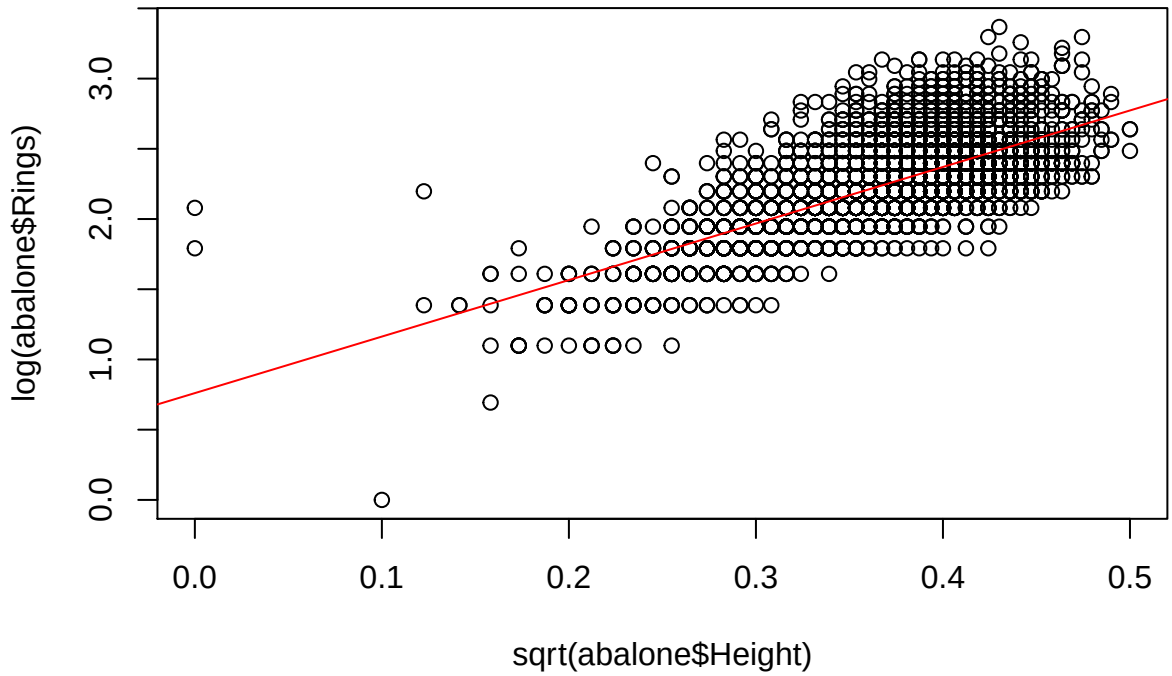


```r
hist(residuals(fit.RP),breaks=40,freq=FALSE,xlab="Residual",main="")
curve(dnorm(x,mean=0,sd=sd(residuals(fit.RP))),add=TRUE,col="blue")
```



Removing Outliers & Data Transformation

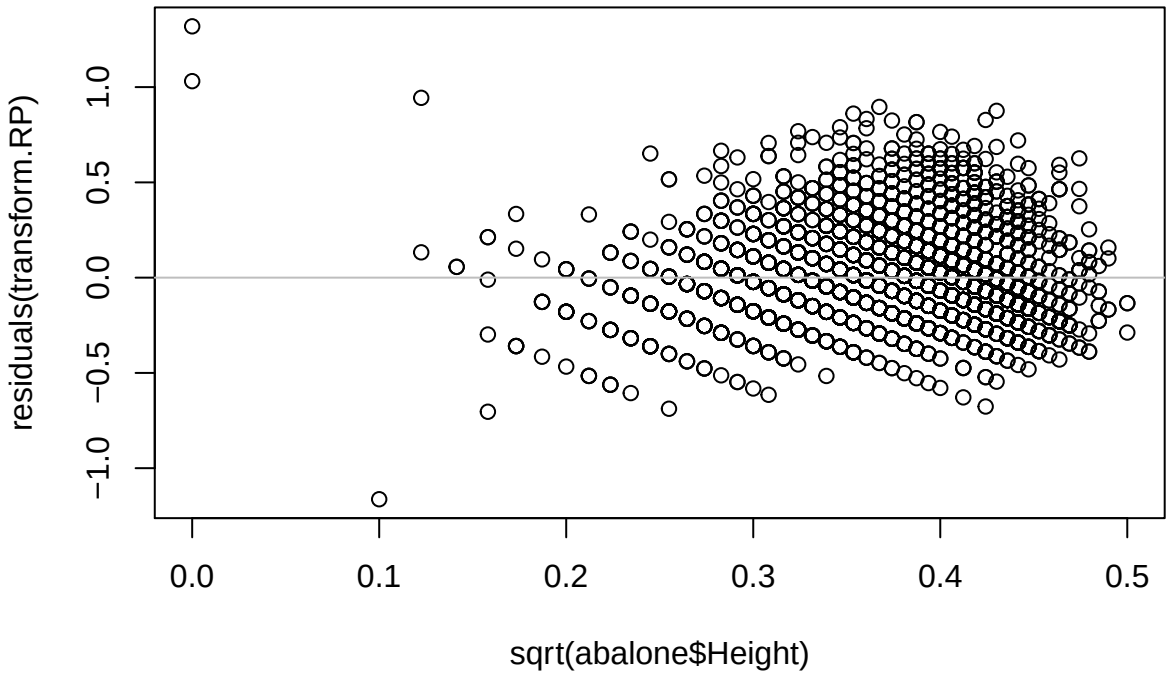
```r
abalone<-filter(abalone,Height<0.4)
plot(y=log(abalone$Rings),x=sqrt(abalone$Height))
transform.RP<-lm(log(abalone$Rings)~sqrt(abalone$Height))
abline(coef(transform.RP),col='red')
```



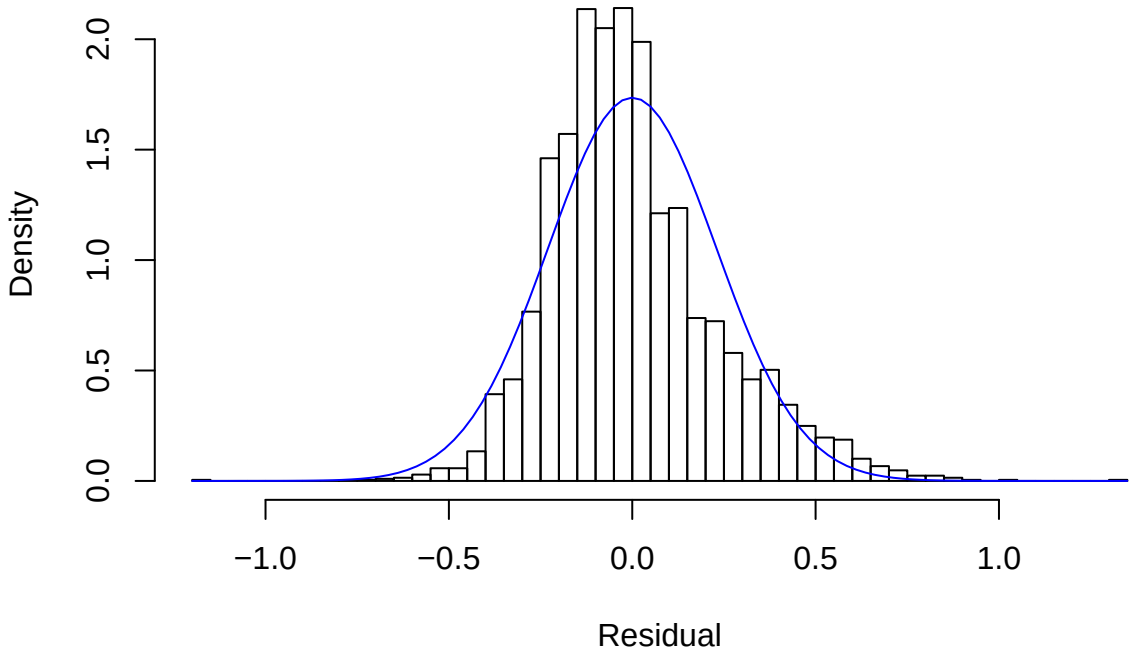
```r
summary(transform.RP)
```

```
##
## Call:
## lm(formula = log(abalone$Rings) ~ sqrt(abalone$Height))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16301 -0.14958 -0.03369  0.11898  1.31895
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.76049    0.02407   31.59   <2e-16 ***
## sqrt(abalone$Height)   4.02516    0.06452   62.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2299 on 4173 degrees of freedom
## Multiple R-squared:  0.4826, Adjusted R-squared:  0.4824
## F-statistic:  3892 on 1 and 4173 DF,  p-value: < 2.2e-16
```

```r
plot(sqrt(abalone$Height),residuals(transform.RP))
abline(h=0,col='gray')
```

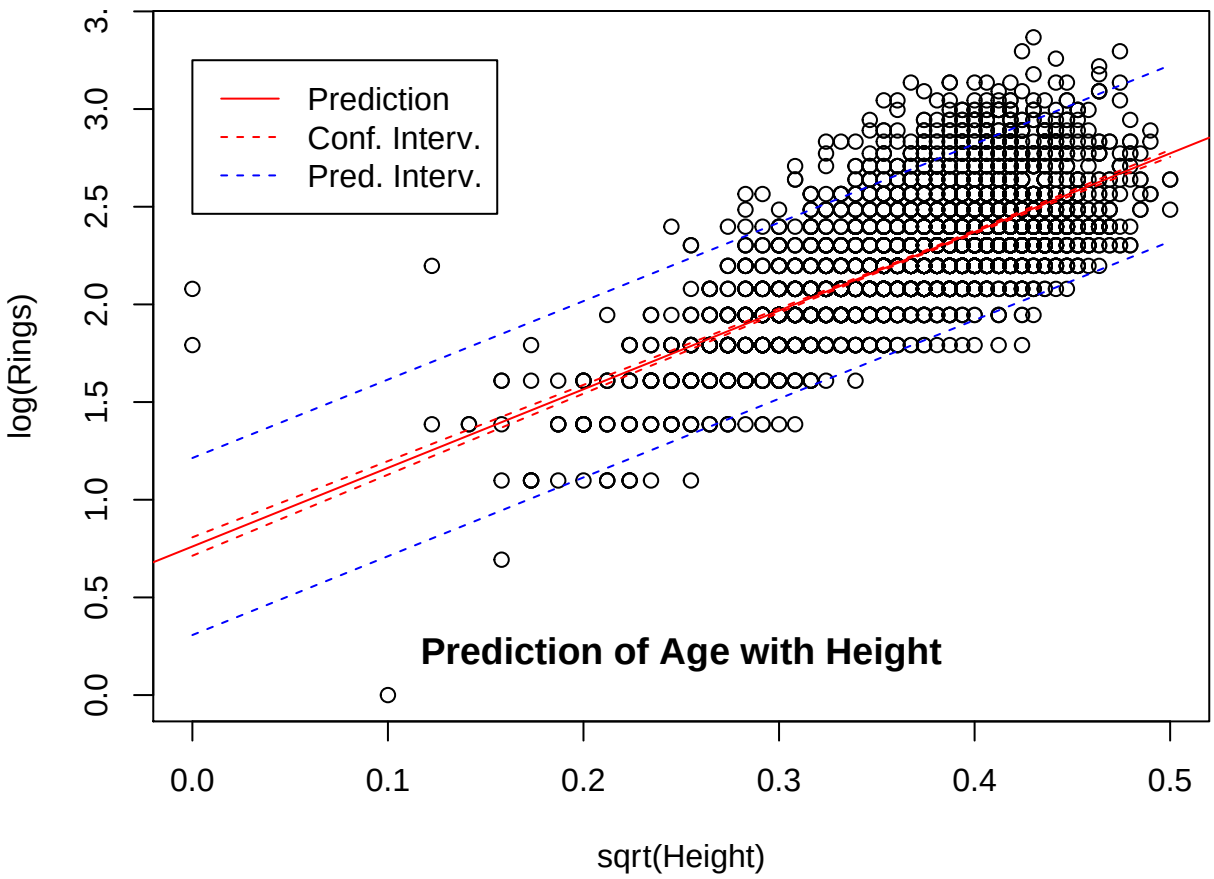



```r
hist(residuals(transform.RP),breaks=40,freq=FALSE,xlab="Residual",main="")
curve(dnorm(x,mean=0,sd=sd(residuals(transform.RP))),add=TRUE,col="blue")
```



Building Confidence & Prediction Interval

```
par(mar=c(4,4,0,1))
x<-sqrt(abalone$Height)
y<-log(abalone$Rings)
fit.RP<-lm(y~x)

xnew<-seq(0,0.5,by=0.01)


#Confidence interval
ynew.interval<-predict(fit.RP,newdata=data.frame(x=xnew),interval='confidence')
plot(x,y,xlab='sqrt(Height)',ylab='log(Rings)')
abline(coef(fit.RP),col='red')

lines(xnew,ynew.interval[,2],lty=2,col='red')
lines(xnew,ynew.interval[,3],lty=2,col='red')

#Prediction interval

yonew.interval<-predict(fit.RP,newdata=data.frame(x=xnew),interval='prediction')
lines(xnew,yonew.interval[,2],lty=2,col='blue')
lines(xnew,yonew.interval[,3],lty=2,col='blue')
legend(0,3.25,c('Prediction','Conf. Interv.','Pred. Interv.'),col=c('red','red','blue'),lty=c(1,2,2))

title('Prediction of Age with Height',line=-17)
```



Data Analysis Conclusions:

After fitting a standard linear model and analyzing residuals, it was clear that there was not a full linear relationship between the two variable as there was sign of non-constant variance which slightly violated linear model assumptions we had made about the residuals being mean 0 with constant variance our residuals tended to have a right skew. Once log transformation of the response variable was made as well as squared root transformation of the predictor, our model tended to perform a lot better

Cigs:

Hypothesis:

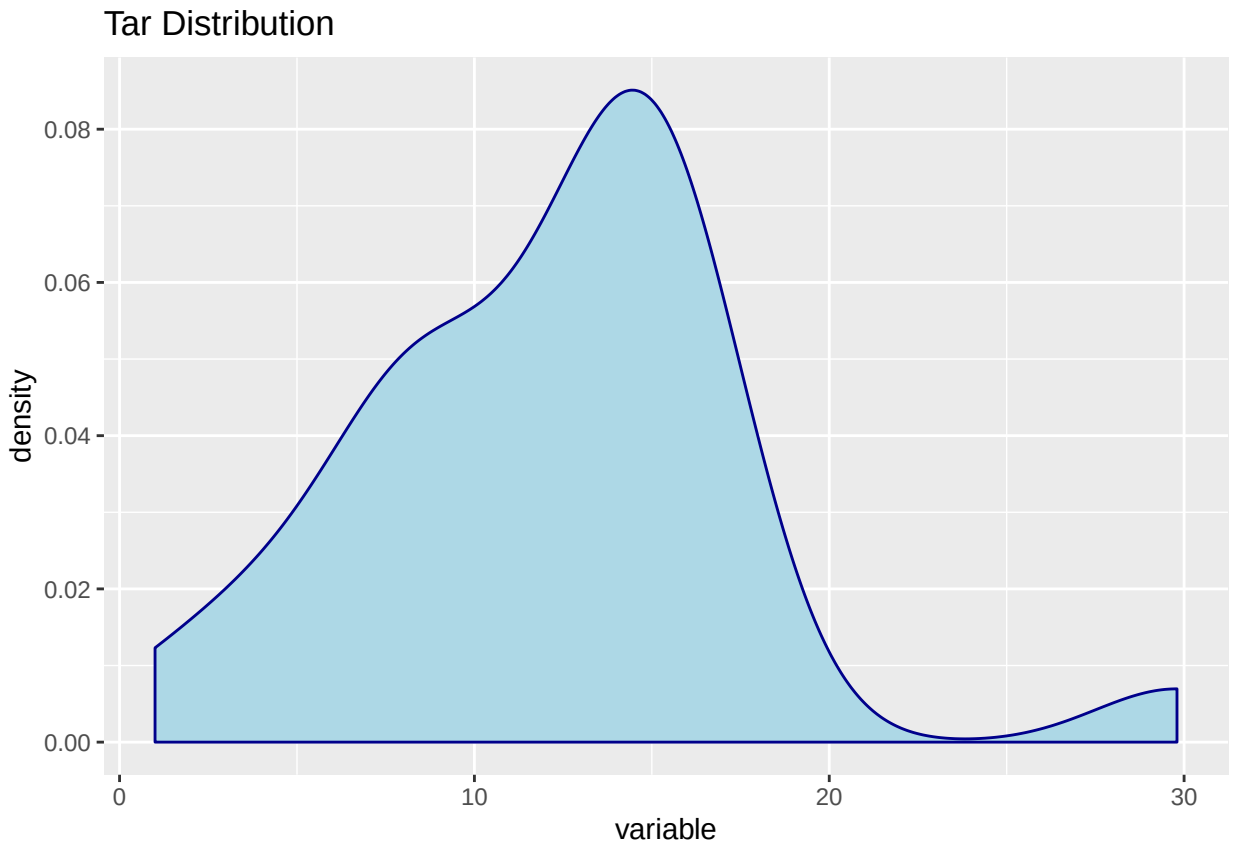
We will investigate and build a model for the relationship between response CO2 produced for cigarette and predictors of Tar, Weight & Nicotine. I believe a Multiple Linear Regression Model will help explain the relationship and will be an adequate enough model with significant relationships between Tar and Weight

```
# importing data
file1 <- "http://www.math.mcgill.ca/yyang/regression/data/cigs.csv"
cigs <- read.csv(file1, header = TRUE)
```

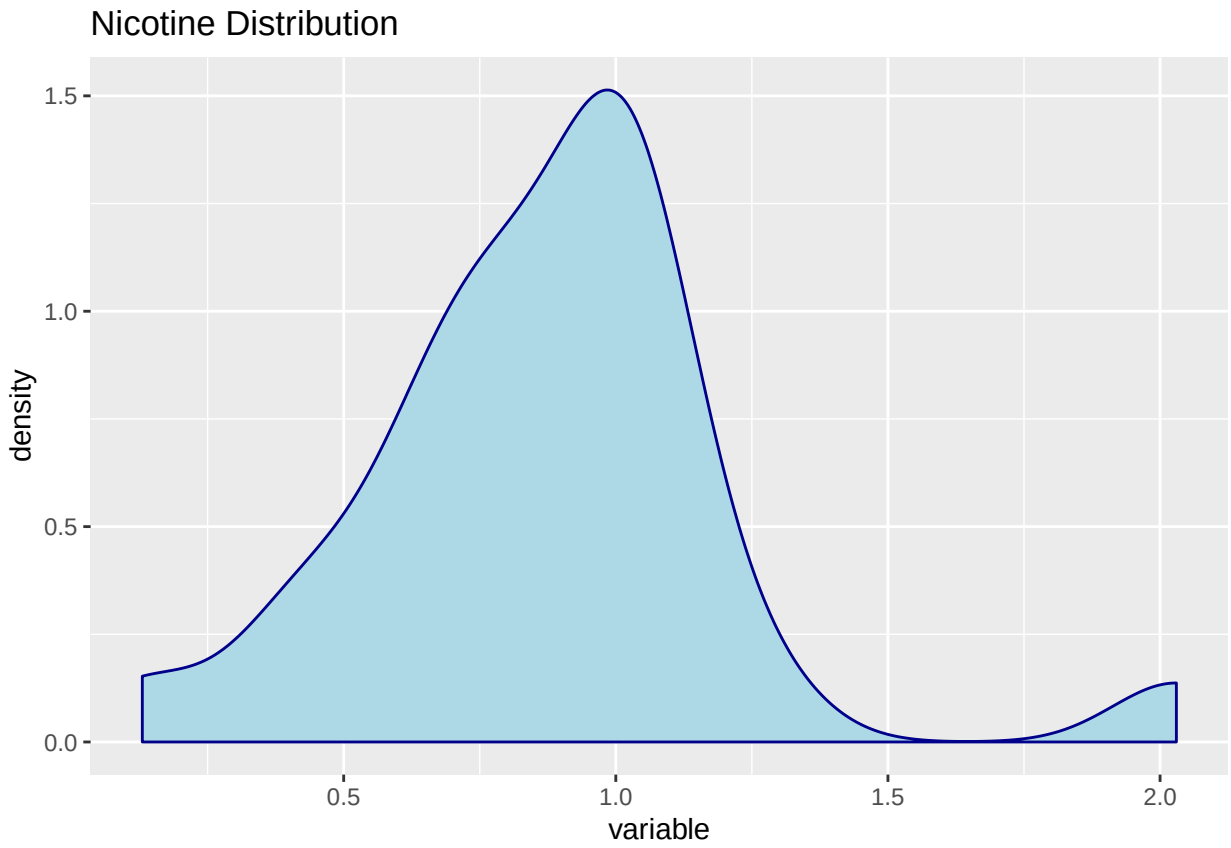
```
CO<-cigs$CO
TAR<-cigs$TAR
NICOTINE<-cigs$NICOTINE
WEIGHT<-cigs$WEIGHT
```

Univariate Analysis

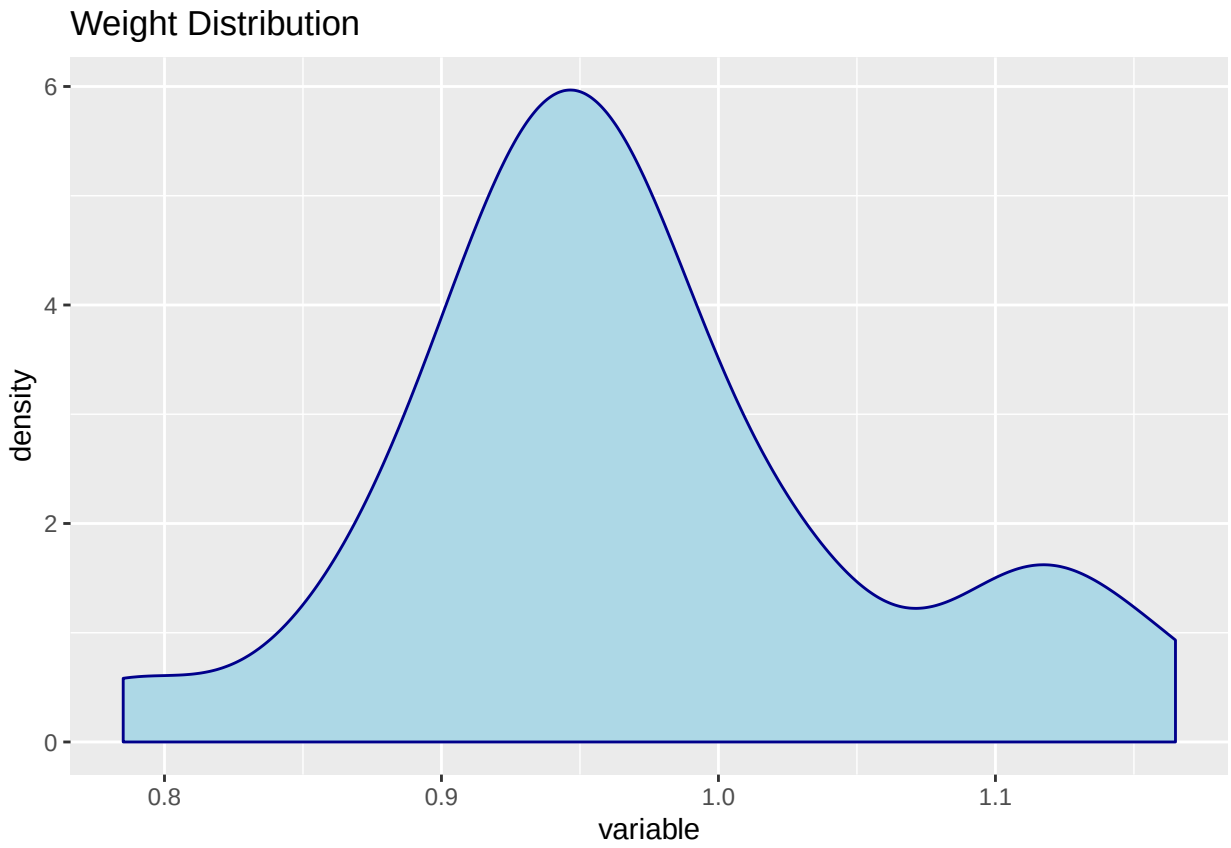
```
Plot_Distribution(cigs,TAR,"Tar Distribution")
```



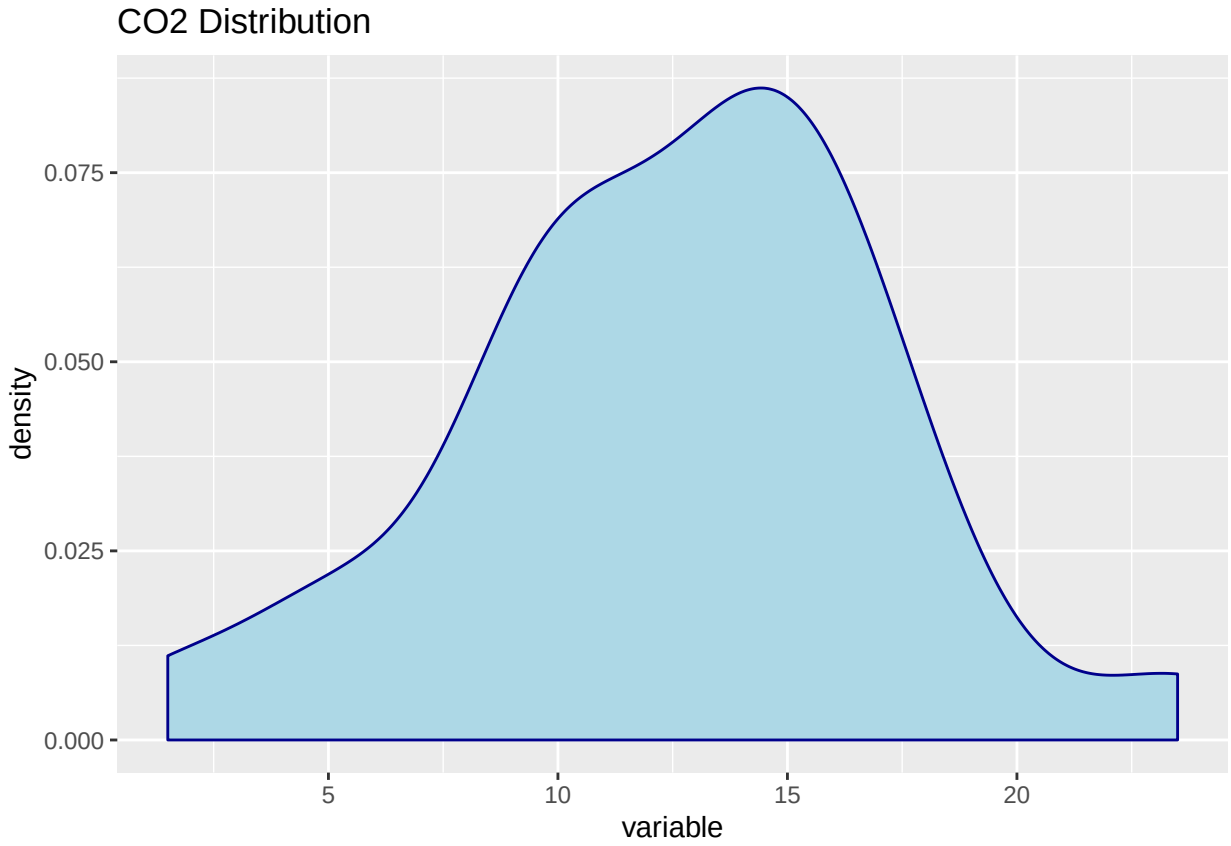

```
Plot_Distribution(cigs,NICOTINE,"Nicotine Distribution")
```




```
Plot_Distribution(cigs,WEIGHT,"Weight Distribution")
```

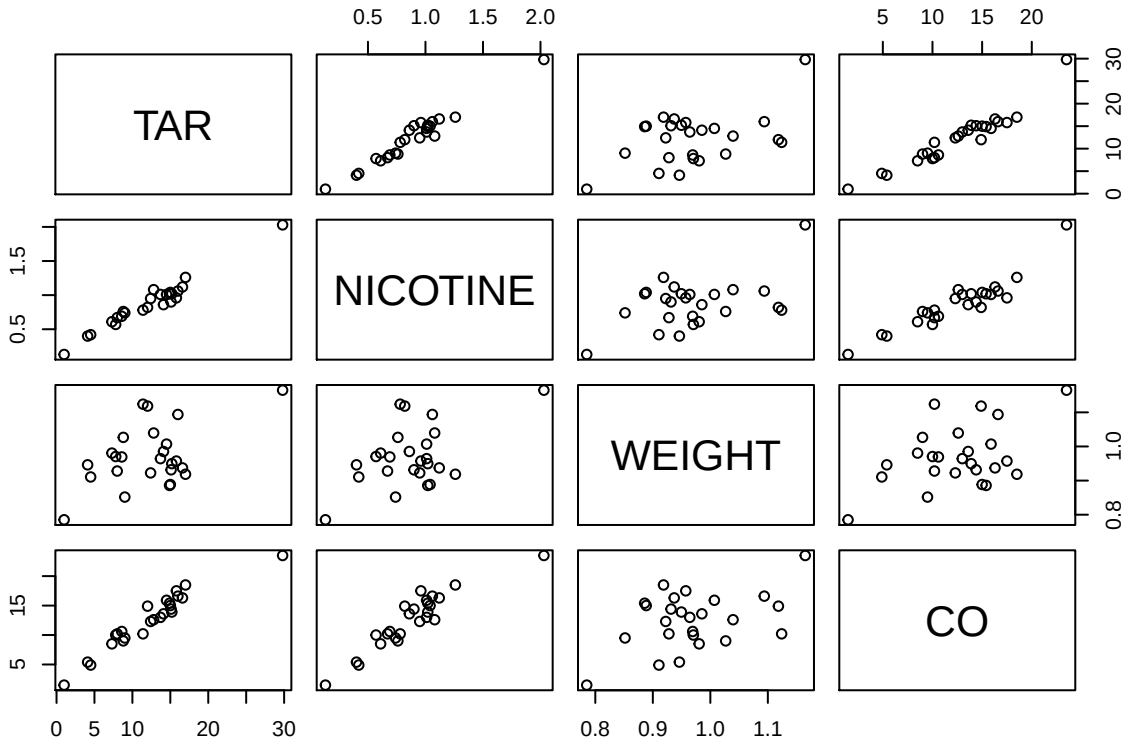



```
Plot_Distribution(cigs,CO,"CO2 Distribution")
```



Bivariate Analysis

```
plot(cigs)
```



Regression Analysis

```
summary(lm(CO~TAR+NICOTINE+WEIGHT))
```

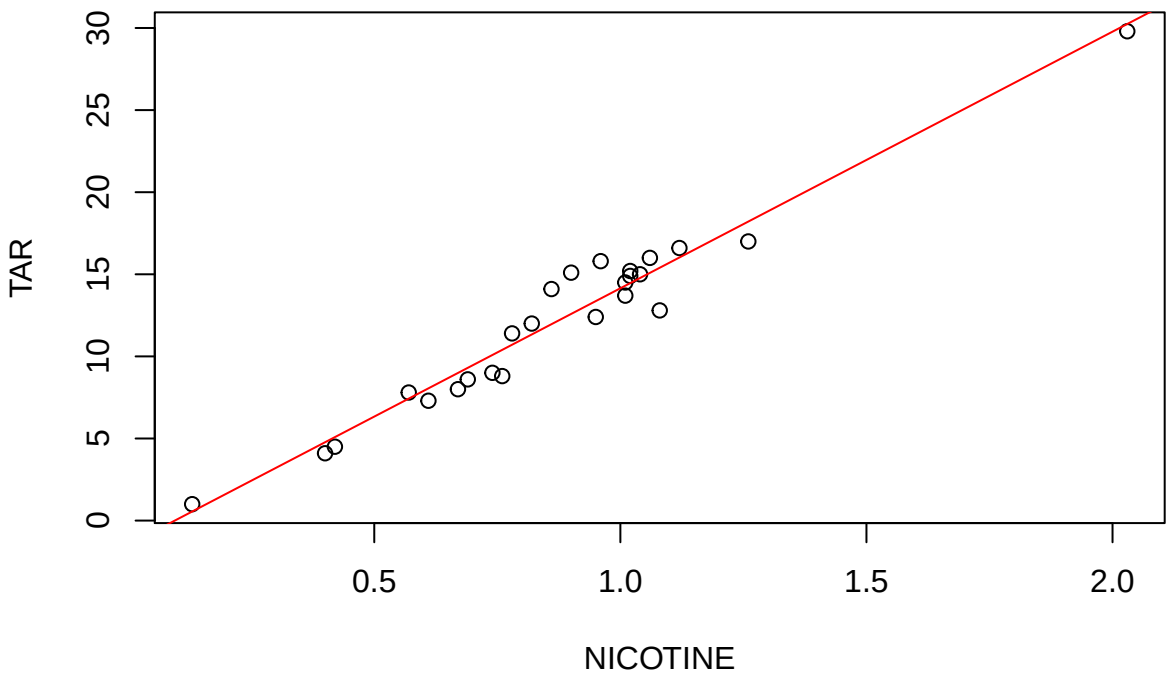
```
##
## Call:
## lm(formula = CO ~ TAR + NICOTINE + WEIGHT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89261 -0.78269  0.00428  0.92891  2.45082
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2022     3.4618   0.925 0.365464
## TAR           0.9626     0.2422   3.974 0.000692 ***
## NICOTINE     -2.6317     3.9006  -0.675 0.507234
## WEIGHT       -0.1305     3.8853  -0.034 0.973527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.446 on 21 degrees of freedom
## Multiple R-squared:  0.9186, Adjusted R-squared:  0.907
## F-statistic: 78.98 on 3 and 21 DF,  p-value: 1.329e-11
```

```r
anova(lm(CO~TAR+NICOTINE+WEIGHT))
```

```
## Analysis of Variance Table
##
## Response: CO
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## TAR        1 494.28  494.28 236.4843 6.651e-13 ***
## NICOTINE   1   0.97    0.97   0.4661    0.5023
## WEIGHT     1   0.00    0.00   0.0011    0.9735
## Residuals 21  43.89    2.09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analyzing Muticollinearity

```r
plot(y=TAR,x=NICOTINE)
abline(coef(lm(TAR~NICOTINE)),col='red')
```




```r
cor(TAR,NICOTINE)
```

```
## [1] 0.9766076
```

Picking Model

```r
anova(lm(CO~TAR+WEIGHT))
```

```
## Analysis of Variance Table
##
## Response: CO
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## TAR        1 494.28  494.28 242.4892 2.308e-13 ***
## WEIGHT     1   0.03    0.03   0.0123    0.9127
## Residuals 22  44.84    2.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

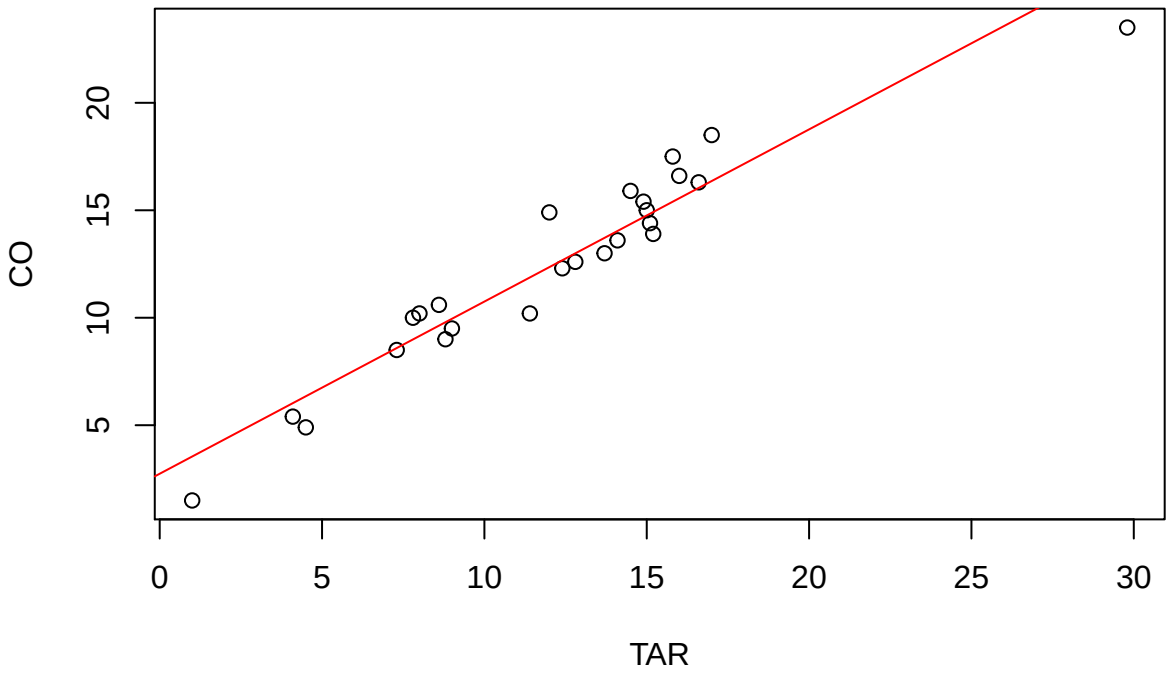
```r
sigma(lm(CO~TAR+WEIGHT))
```

```
## [1] 1.427713
```

```r
sigma(lm(CO~TAR))
```

```
## [1] 1.396721
```

```r
plot(y=CO,x=TAR)

abline(coef(lm(CO~TAR)),col='red')
```



Data Analysis Conclusions:

After analyzing predictors, conclusions were made about multicollinearity between Tar & Nicotine thus we opted to choose tar as one of the variable in our model, upon fitting a MLR model with tar, weight, nicotine it was clear that the only significant preditor was tar. Analysis on resdisual was done to confirm this finding

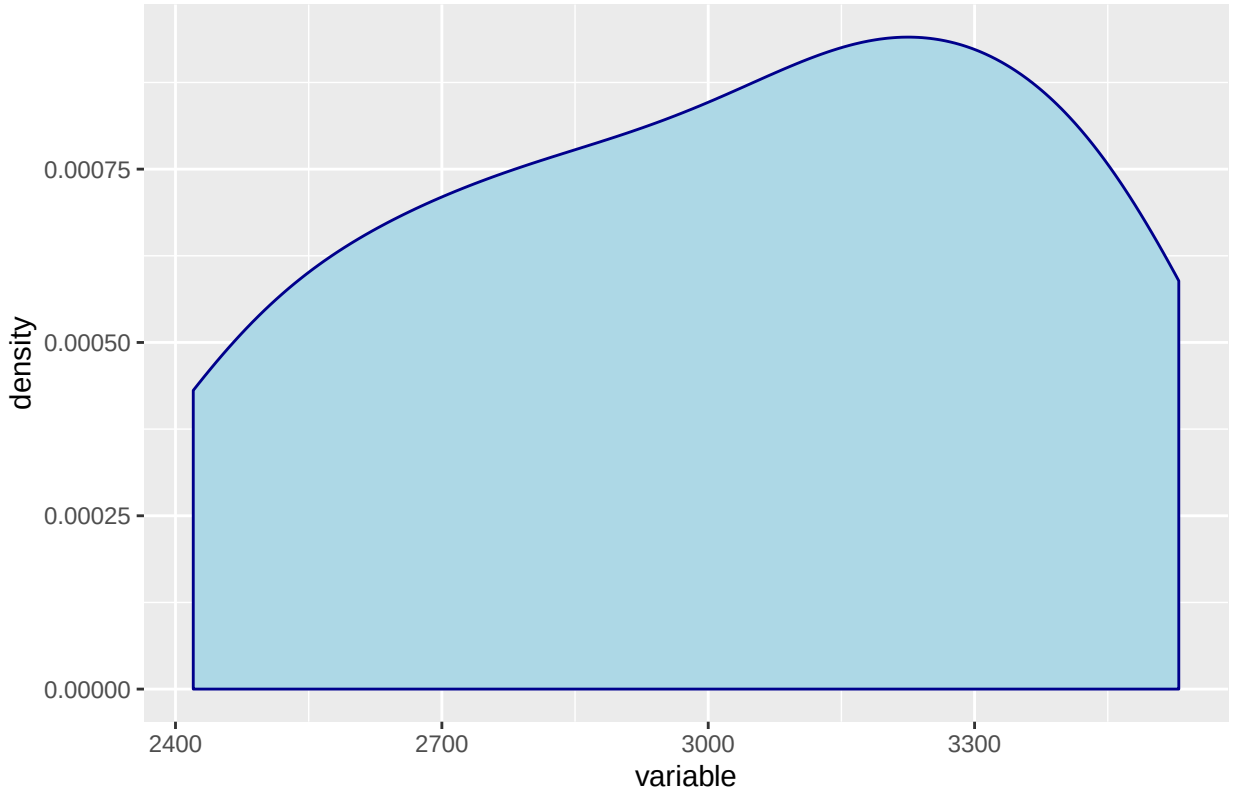
BirthWeight (Smokers vs Non-Smokers):

Hypothesis: I believe that there is a difference in Birth Weight in relation to Gestion time, in between Smokers & Non-Smokers

```
data.source<-"http://www.math.mcgill.ca/yyang/regression/data/birthsmokers.csv"
birthsmokers<-read.csv(file=data.source)
```

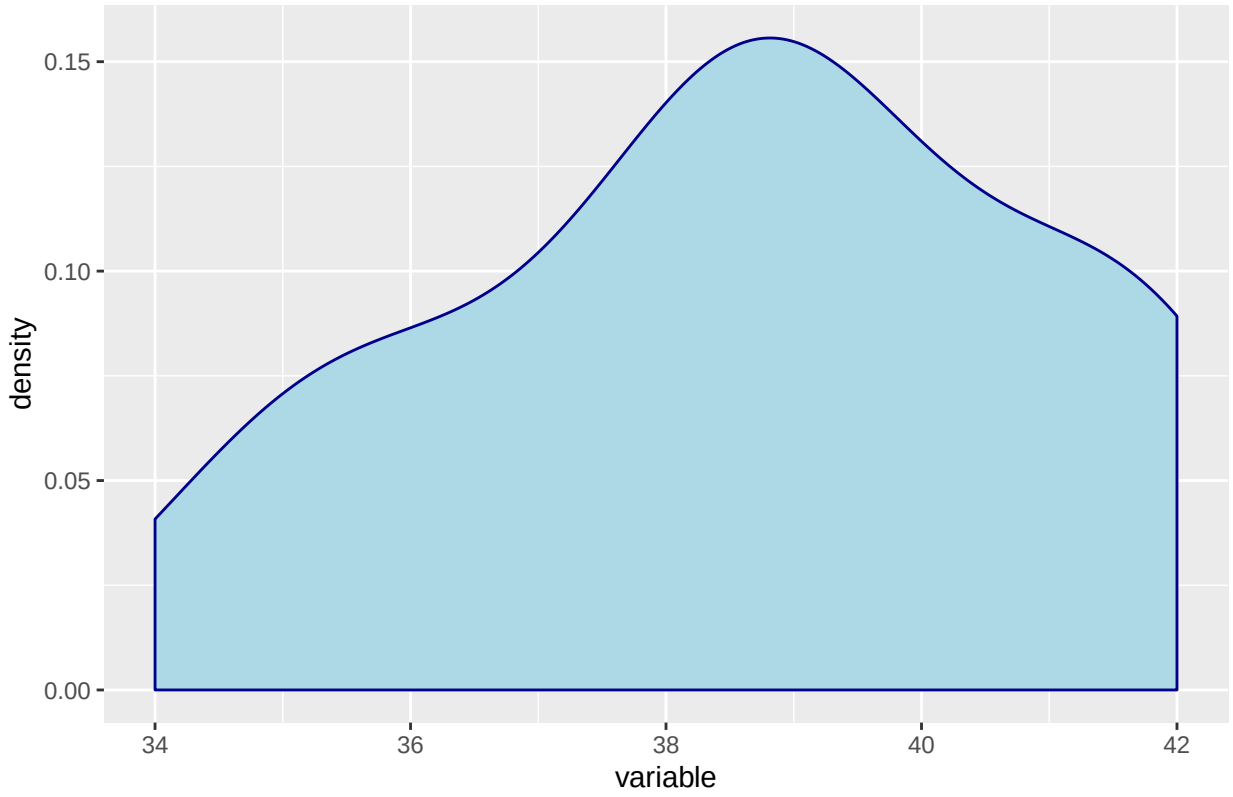
Univariate Analysis: Weight

```
Plot_Distribution(birthsmokers,birthsmokers$Wgt)
```



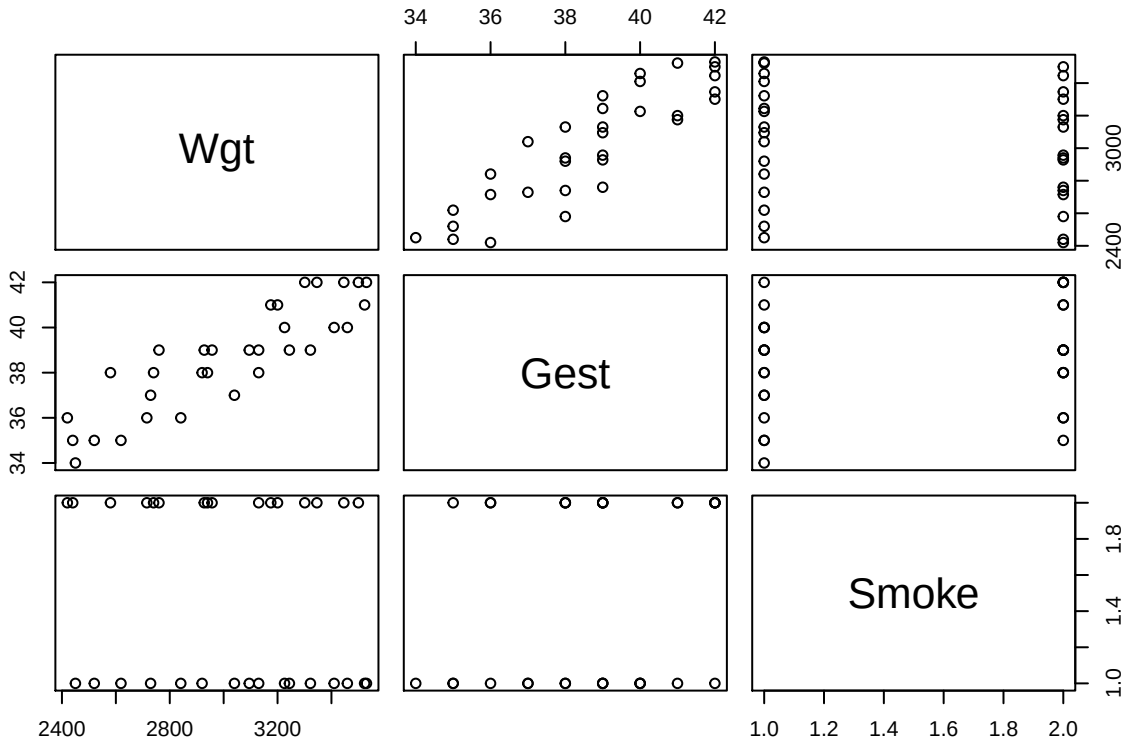
Univariate Analysis: Gestations

```
Plot_Distribution(birthsmokers,birthsmokers$Gest)
```

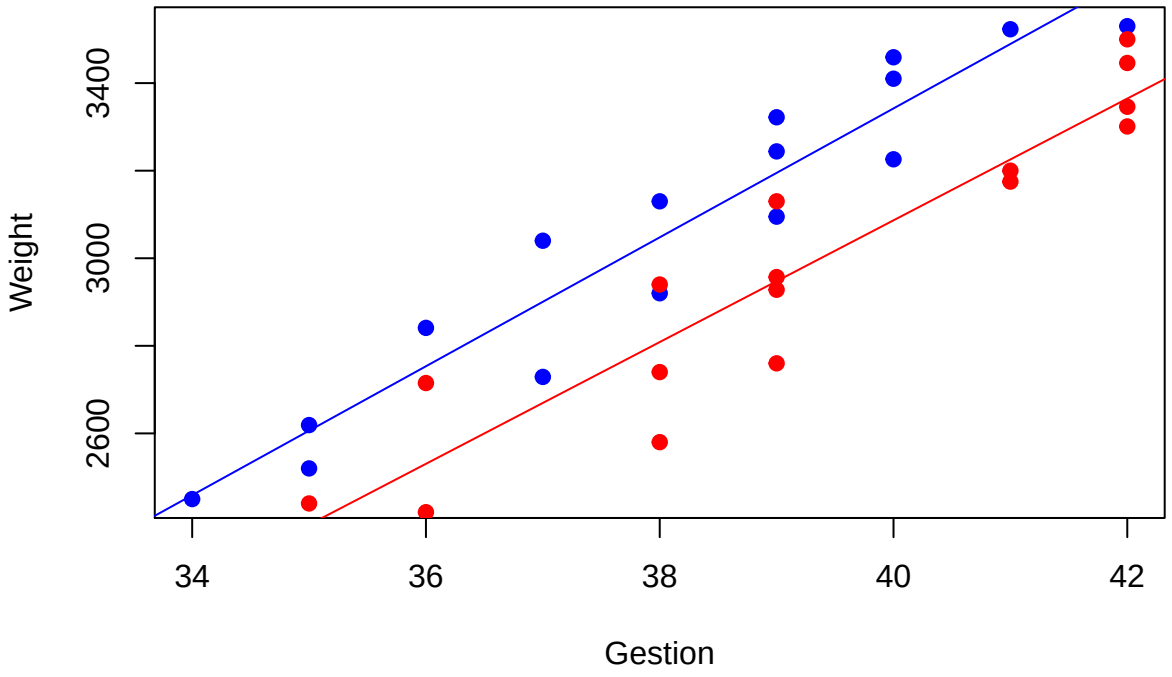


Bivariate Analysis

```
plot(birthsmokers)
```



```r
plot(x=subset(birthsmokers,Smoke=="no")$Gest,y=subset(birthsmokers,Smoke=="no")$Wgt,col="blue",pch=19,x

points(x=subset(birthsmokers,Smoke=="yes")$Gest,y=subset(birthsmokers,Smoke=="yes")$Wgt,col="red",pch=1

abline(coef(lm(Wgt~Gest,data=subset(birthsmokers,Smoke=="no"))),col='blue')
abline(coef(lm(Wgt~Gest,data=subset(birthsmokers,Smoke=="yes"))),col='red')
```



Regression Analysis

```r
summary(lm(Wgt~Gest,data=subset(birthsmokers,Smoke=="no")))
```

```
##
## Call:
## lm(formula = Wgt ~ Gest, data = subset(birthsmokers, Smoke ==
##     "no"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -171.52 -101.59   23.28   83.63  139.48
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2546.14     457.29  -5.568 6.93e-05 ***
## Gest          147.21      11.97  12.294 6.85e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 106.9 on 14 degrees of freedom
## Multiple R-squared:  0.9152, Adjusted R-squared:  0.9092
## F-statistic: 151.1 on 1 and 14 DF,  p-value: 6.852e-09
```

```r
summary(lm(Wgt~Gest,data=subset(birthsmokers,Smoke=="yes")))
```

```
##
## Call:
## lm(formula = Wgt ~ Gest, data = subset(birthsmokers, Smoke ==
##     "yes"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -228.53  -64.86  -19.10   93.89  184.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2474.56     553.97  -4.467 0.000532 ***
## Gest          139.03      14.11   9.851 1.12e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 126.6 on 14 degrees of freedom
## Multiple R-squared:  0.8739, Adjusted R-squared:  0.8649
## F-statistic: 97.04 on 1 and 14 DF,  p-value: 1.125e-07
```

```r
confint(lm(Wgt~Gest,data=subset(birthsmokers,Smoke=="no")))
```

```
##                   2.5 %     97.5 %
## (Intercept) -3526.9338 -1565.3420
## Gest          121.5251   172.8887
```

```r
confint(lm(Wgt~Gest,data=subset(birthsmokers,Smoke=="yes")))
```

```
##                   2.5 %     97.5 %
## (Intercept) -3662.7169 -1286.4114
## Gest          108.7586   169.2989
```

Data Analysis Conclusions:

There is strong evidence of difference in correlation between Non Smoker & Smokers as observed from the data. More specifically there is 8 units of difference between predictors & response for Smokers vs Non-Smokers