ASSIGNMENT

Input

```
#### question 1------------------------
### this data is in the R base installation.
library(datasets)
data_df<-state
head(state)
```

Output

```
           Murder Population Illiteracy Income Frost
Alabama      15.1       3615        2.1   3624    20
Alaska       11.3        365        1.5   6315   152
Arizona       7.8       2212        1.8   4530    15
Arkansas     10.1       2110        1.9   3378    65
California   10.3      21198        1.1   5114    20
Colorado      6.8       2541        0.7   4884   166
```

Input

```
##### question 2 ----------------------
str(data_df) ### the structure of the data
summary(data_df) ### summary statistics
```

Output

```
str(data_df) ### the struct of the data
'data.frame':   50 obs. of  5 variables:
 $ Murder    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
 $ Population: num  3615 365 2212 2110 21198 ...
 $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
 $ Income    : num  3624 6315 4530 3378 5114 ...
 $ Frost     : num  20 152 15 65 20 166 139 103 11 60 ...
>
```

Output

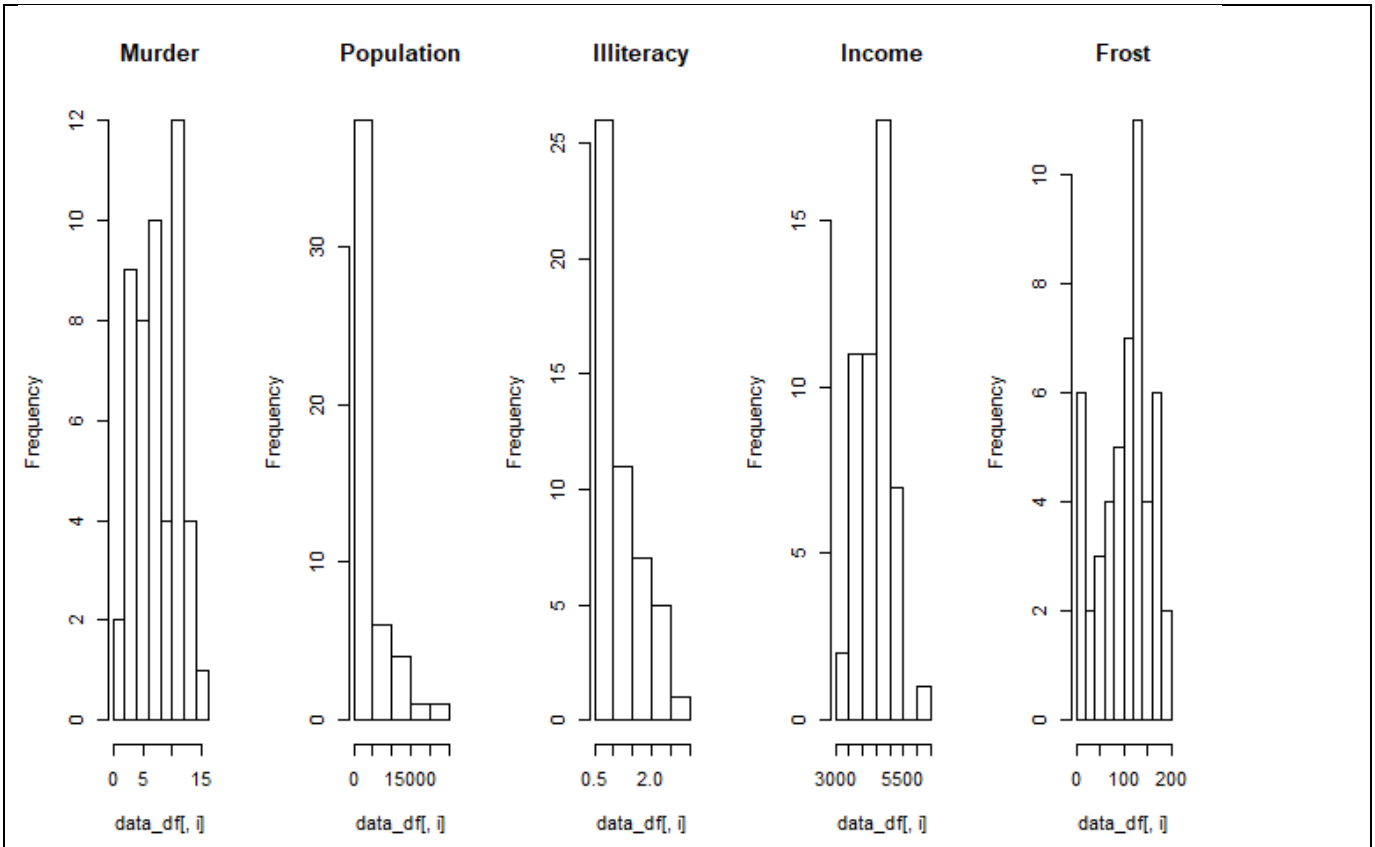
```
summary(data_df) ### summary statistics
     Murder          Population      Illiteracy        Income
 Min.   : 1.400   Min.   :  365   Min.   :0.500   Min.   :3098
 1st Qu.: 4.350   1st Qu.: 1080   1st Qu.:0.625   1st Qu.:3993
 Median : 6.850   Median : 2838   Median :0.950   Median :4519
 Mean   : 7.378   Mean   : 4246   Mean   :1.170   Mean   :4436
 3rd Qu.:10.675   3rd Qu.: 4968   3rd Qu.:1.575   3rd Qu.:4814
 Max.   :15.100   Max.   :21198   Max.   :2.800   Max.   :6315
     Frost
 Min.   :  0.00
 1st Qu.: 66.25
 Median :114.50
 Mean   :104.46
```

```
  3rd Qu.:139.75
  Max.   :188.00
```

Input

```
#### histogram ----------
# load the data
data(data_df)

# create histograms for each attribute
par(mfrow=c(1,5))
for(i in 1:5) {
  hist(data_df[,i], main=names(data_df)[i])
}
```

Output



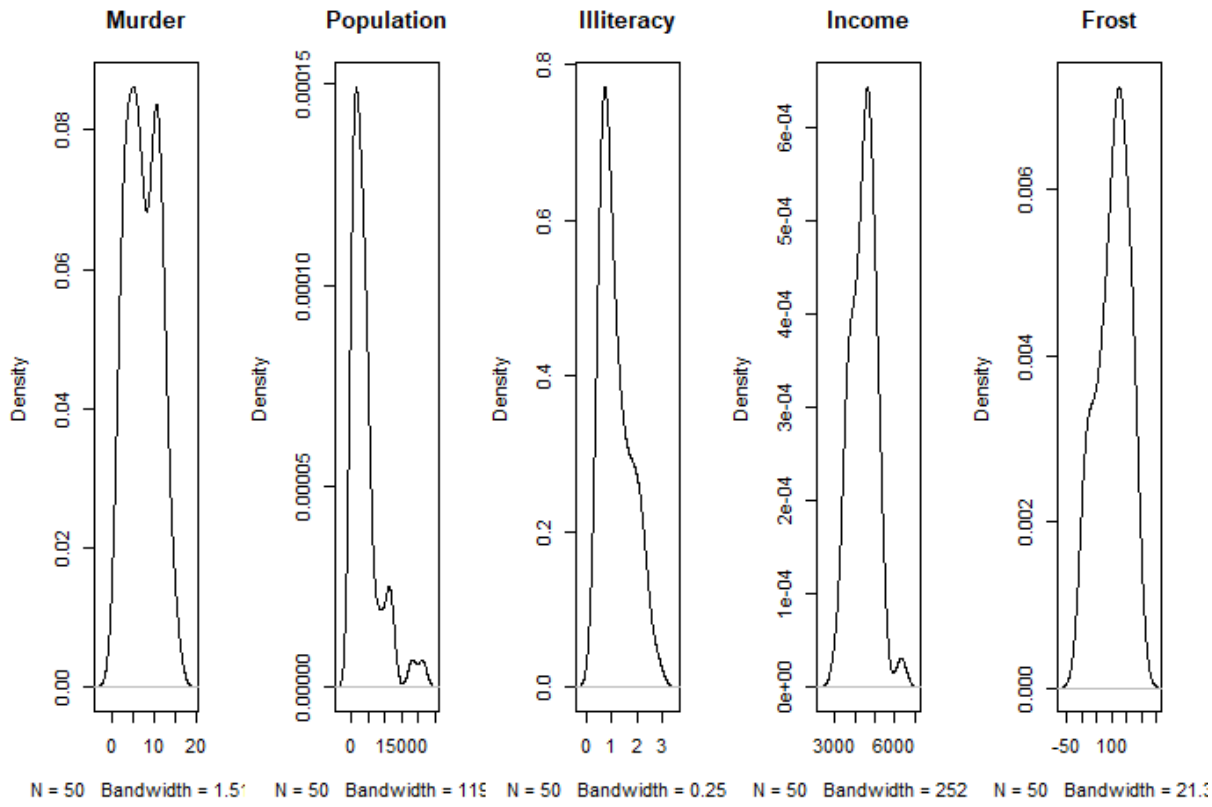
all features in the data set are skewed

Input

```
# load dataset
data(data_df)
```

```
# create a panel of simpler density plots by attribute
par(mfrow=c(1,5))
for(i in 1:5) {
  plot(density(data_df[,i]), main=names(data_df)[i])
}
```

Output

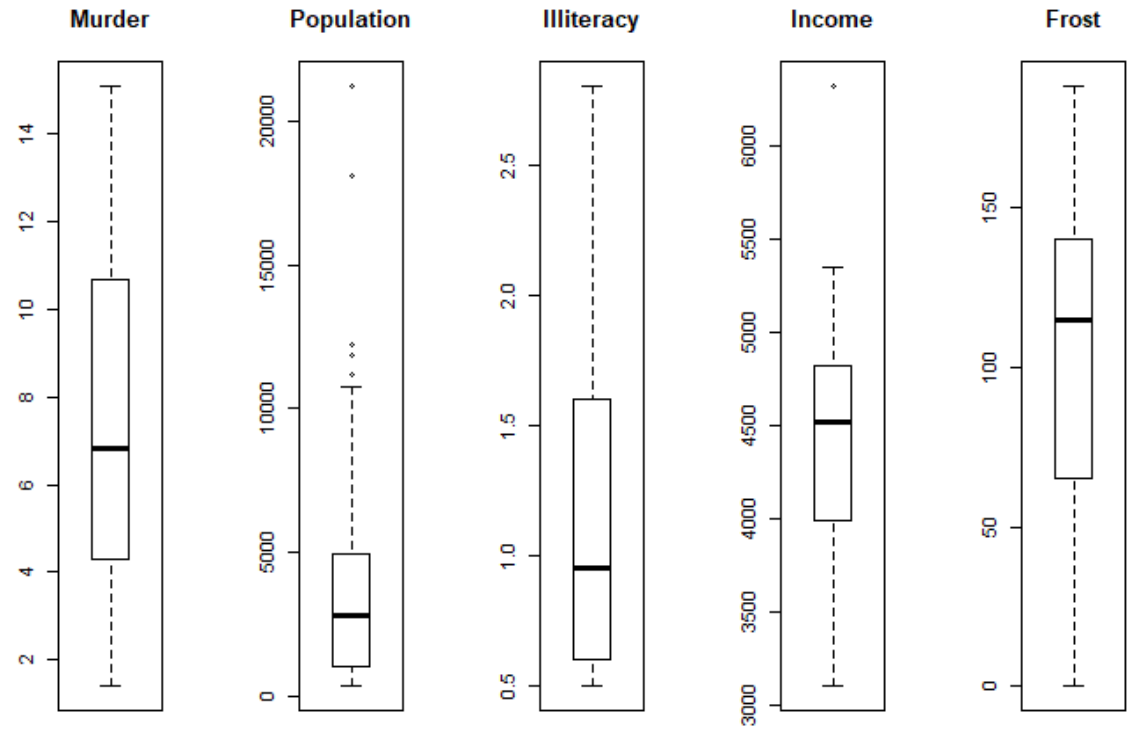



```
### one feature has bimodal
## the rest of the features are skewed
## for algorithms to perform well it will be good to transform the features
```

Input

```
###  Box And Whisker Plots----------
# load dataset
data(data_df)
# Create separate boxplots for each attribute
par(mfrow=c(1,5))
for(i in 1:5) {
  boxplot(data_df[,i], main=names(data_df)[i])
}
```

Output

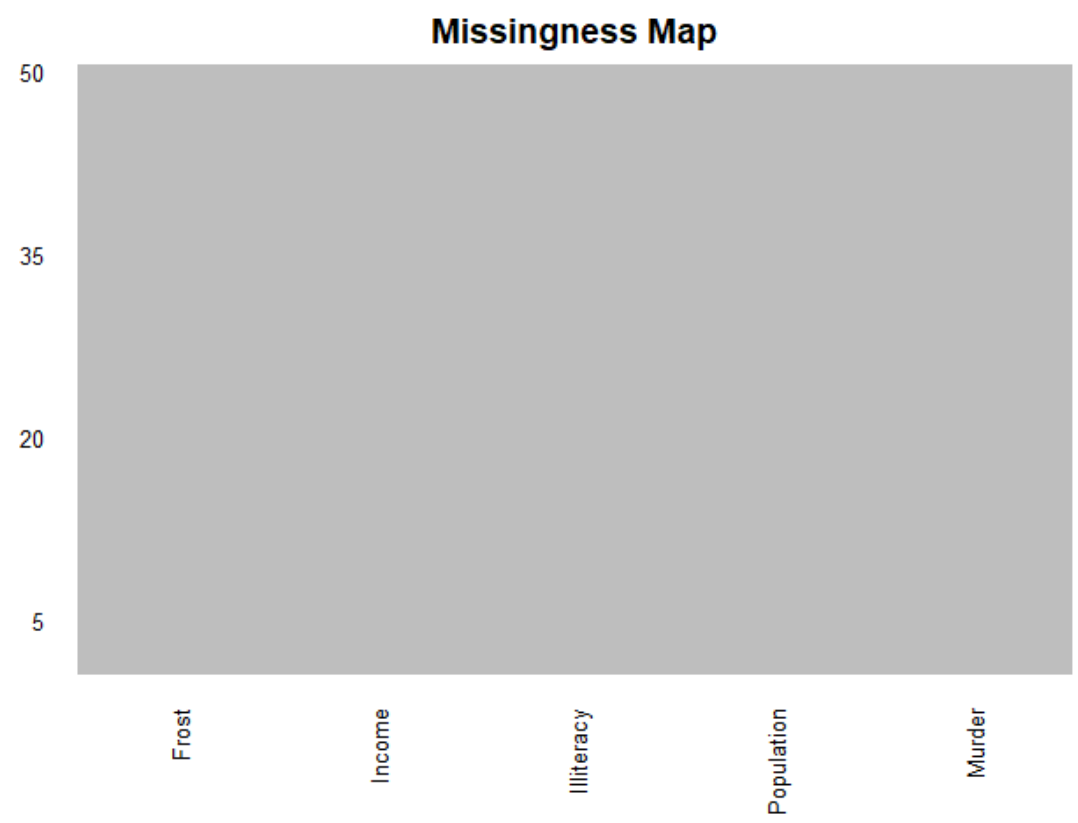


it clear from the the boxplot that population is quite normal
and rest of the features are all skewed

Input

```
#### Missing Plot----------
### the presence of missing values can have a  negative impact on a model

# load packages
install.packages('Amelia')
library(Amelia)
# load dataset
data(data_df)
# create a missing map
missmap(data_df, col=c("black", "grey"), legend=FALSE)
```

Output



the data is clean, there are no missing values
otherwise going forward, we would have to impute the missing values.

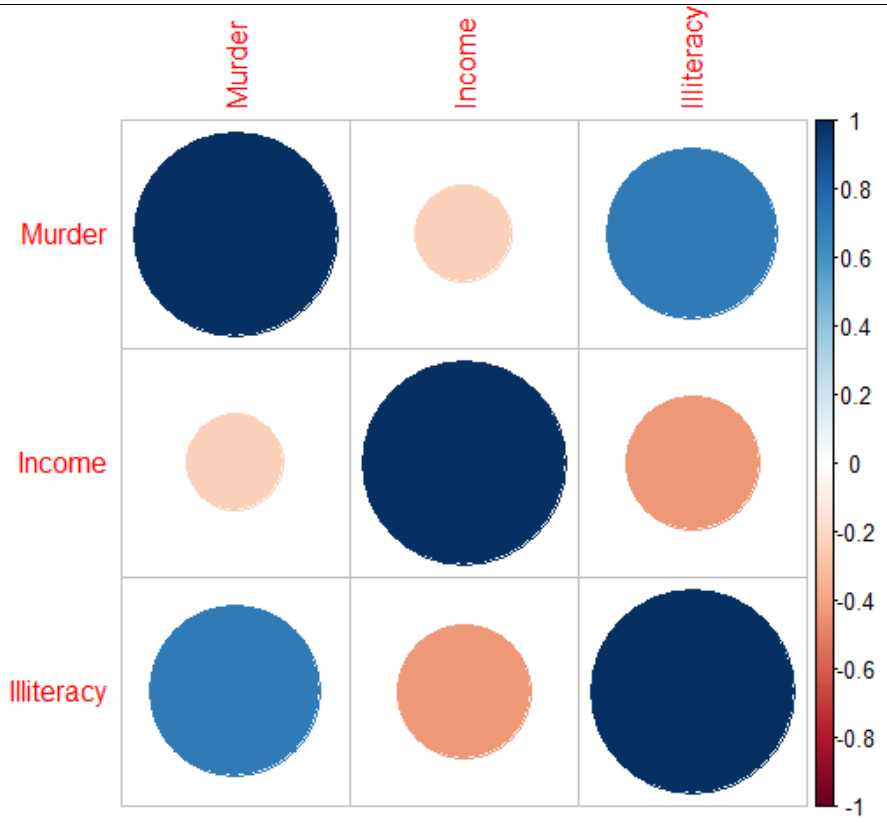
Input

```
##### question 3 and 4----------------------------------------------------

###questions to be answered using regression analysis in this study------------------
### Q1 how does illiteracy rate influence murder
#### Q2 how does income also influence murder

#### hypothesis to be tested from regression analysis using t-test--------------------

### for illiteracy
### null hypothesis--- Illiteracy has no relationship with murder
### alternate hypothesis--- illiteracy has a relationship with murder

#### for income
### null hypothesis--- income has no relationship with murder
### alternate hypothesis--- income has a relationship with murder

#### why I THINK it is an interesting question----------------------------
#### using correlation and correlation plot


#### Correlation Plot----------------------------------
#### I would have to take a subset of the data to get my questions answered.

var=c('Murder','Income','Illiteracy')
data_df_one<-data_df[var]
```

Input

```
# load package
library(corrplot)

# load the data
data(data_df_one)
# calculate correlations
correlations <- cor(data_df_one[,1:3])

# create correlation plot
corrplot(correlations, method="circle")
```

Output

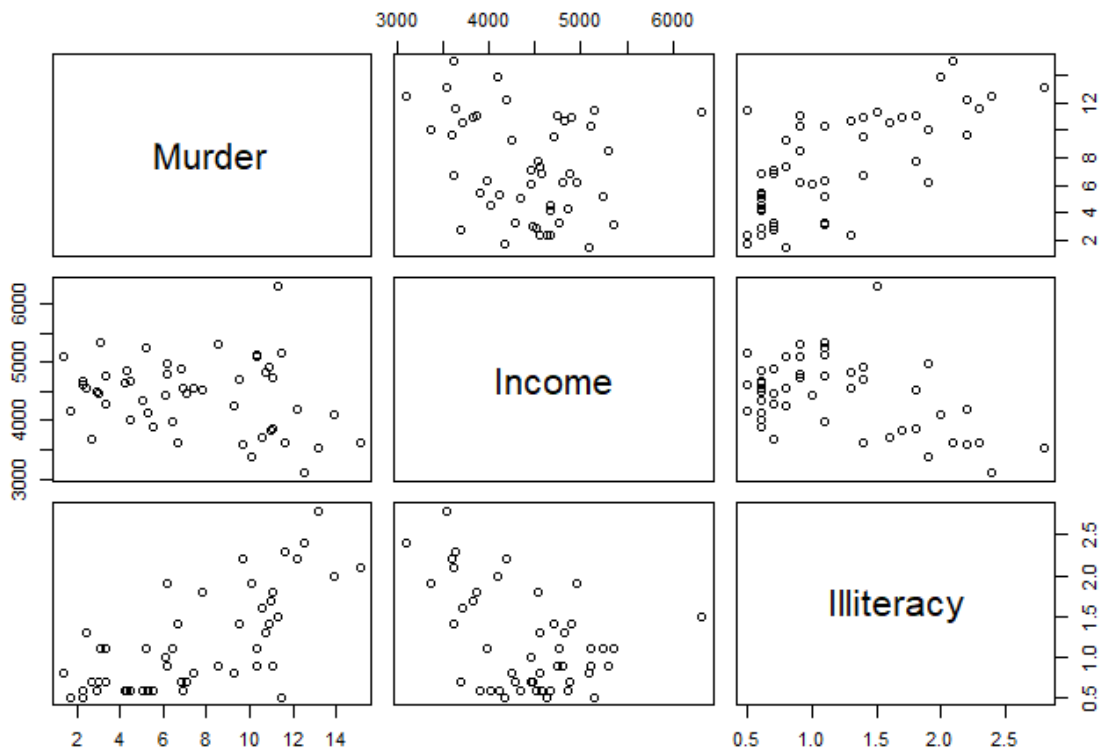


the bigger the circle the higher the correlation and vice versa
the circles tell us that income is negatively correlated with murder.
also, illiteracy rate is positively correlated with murder.

Input

```
###### Scatterplot Matrix---------------------------------

# load the data
data(data_df_one)
# pair-wise scatterplots of all 4 attributes
pairs(data_df_one)
```

Output



there is a downward sloping relationship between murder and income indicating
a negative relationship between murder and income

also, there is a positive relationship between illiteracy and murder, i.e.
indicating a positive relationship

Input

```
##### fitting the regression model------------------------------
fit<-lm(Murder~Illiteracy + Income,data=data_df_one)
summary(fit) ### summary of fit
```

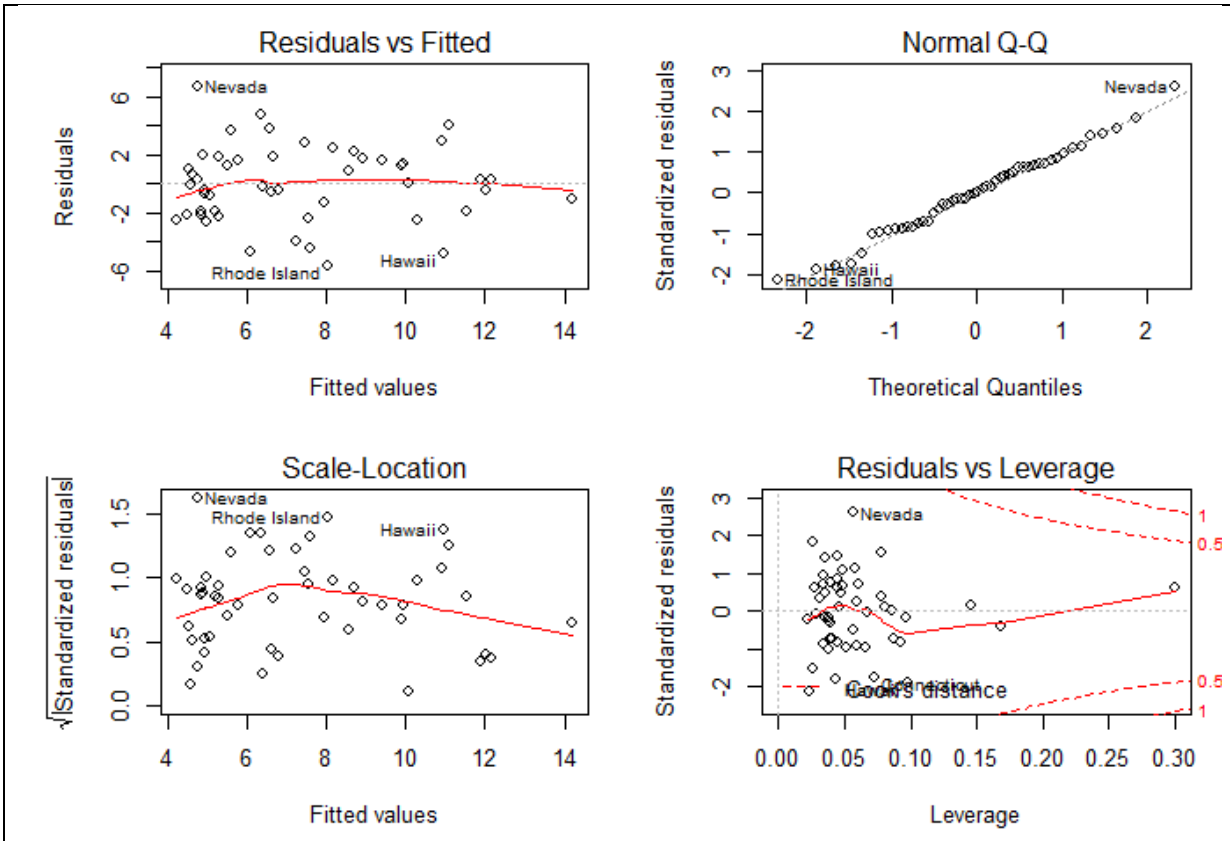
Output

```
Call:
lm(formula = Murder ~ Illiteracy + Income, data = data_df_one)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6343 -1.9289 -0.0171  1.6779  6.7349

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4409926  3.5034602  -0.126    0.900
Illiteracy   4.5099882  0.6934465   6.504 4.63e-08 ***
Income       0.0005731  0.0006879   0.833    0.409
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.661 on 47 degrees of freedom
Multiple R-squared:  0.5015,   Adjusted R-squared:  0.4803
F-statistic: 23.64 on 2 and 47 DF,  p-value: 7.841e-08
```

Input

```
##### checking model assumptions----------------------------------------
par(mfrow=c(2,2))
plot(fit)
```

Output



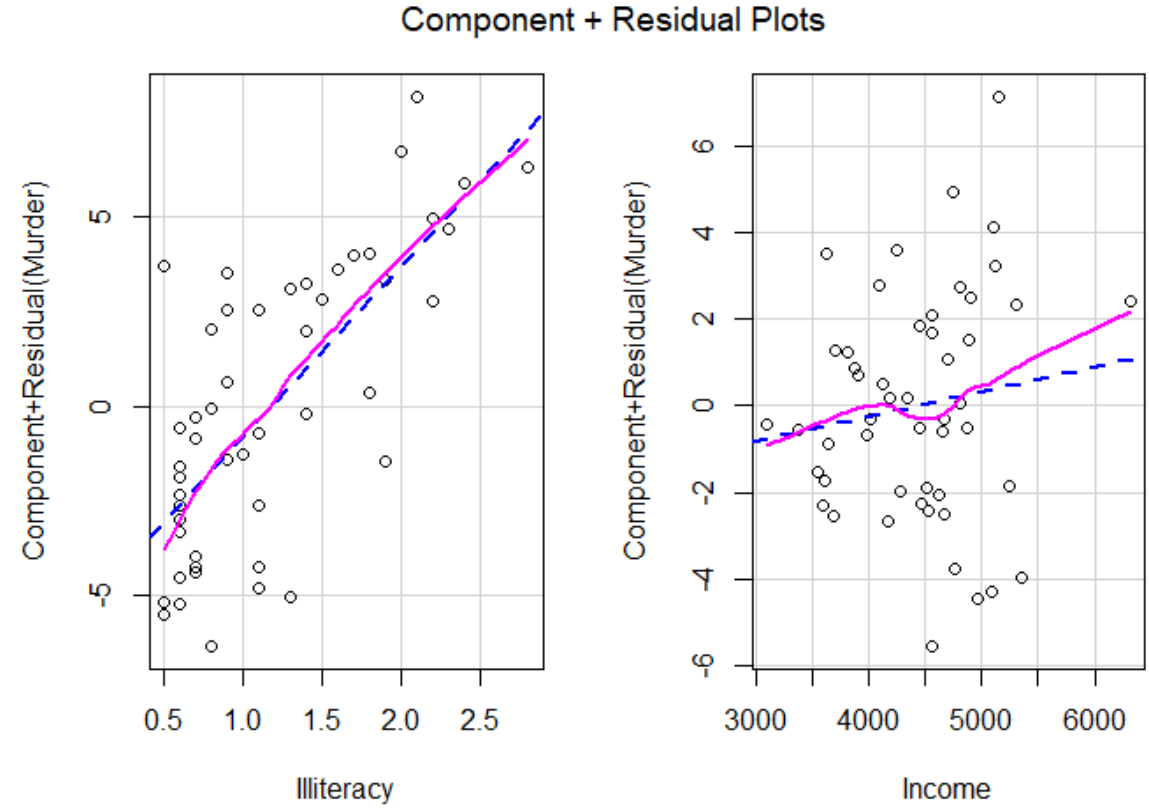
normality assumption is quite ok given that most of the points are on the dotted line
of the normal Q-Q plot.

```
#### constant of variance (Homoscedasticity) has also been met given that there is no
##### clear pattern between residuals Vs fitted plot.
```

Input

```
## LINEARITY-------------------------------------------------
install.packages('car')
library(car)
crPlots(fit)
```

Output




```
## linearity assumptions met
## since the broken lines are quite straight
```

Input

```
###### test statistic, slope, pvalue, intercept and R squared
summary(fit)
```

Output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4409926  3.5034602  -0.126    0.900
Illiteracy   4.5099882  0.6934465   6.504 4.63e-08 ***
Income       0.0005731  0.0006879   0.833    0.409
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.661 on 47 degrees of freedom
Multiple R-squared:  0.5015,   Adjusted R-squared:  0.4803
F-statistic: 23.64 on 2 and 47 DF,  p-value: 7.841e-08
```

Input

```
##### confidence interval------------------------------
confint(fit)
```

Output

```
                    2.5 %     97.5 %
(Intercept) -7.4890453235 6.60706020
Illiteracy   3.1149537003 5.90502263
Income      -0.0008106927 0.00195696


confint(fit)
#### since both the intercept and income confidence interval contain zero
## it means they are insignificant
### and only illiteracy is significant
```

Conclusion

```
##### conclusion
##illiteracy---- ## holding all other factors constant it is expected a unit increase in
### illiteracy level will cause an increase in murder by 4.51 according to the model

##### income---- ## holding all other factors constant it is expected a dollar increase in
### income will cause an increase in murder by 0.00057 according to the model

#### only variable that explains the model is income but intercept and illiteracy are all
### insignificant
```

Input

```
qf(0.05, 1, 47, lower.tail = F)
```

Output

```
qf(0.05, 1, 47, lower.tail = F)
[1] 4.0471
```

overall, the F-statistic when compared to F-critical
tells us that the entire
model is significant
since the F test statistic is greater than critical value of F