Regression Analysis

MANOJ KUMAR - 2048015

30/01/2021

1.Import the dataset data_marketing_budget_mo12 and do the exploratory data analysis .

```
# Importing the shared Marketing Budget Data.

RocketData <- read.csv("data-marketing-budget.csv")
RocketData
```

```
##    Month Spend  Sales
## 1      1  1000   9914
## 2      2  4000  40487
## 3      3  5000  54324
## 4      4  4500  50044
## 5      5  3000  34719
## 6      6  4000  42551
## 7      7  9000  94871
## 8      8 11000 118914
## 9      9 15000 158484
## 10    10 12000 131348
## 11    11  7000  78504
## 12    12  3000  36284
```

```
# Summary
summary(RocketData)
```

```
##      Month           Spend           Sales
##  Min.   : 1.00   Min.   : 1000   Min.   :  9914
##  1st Qu.: 3.75   1st Qu.: 3750   1st Qu.: 39436
##  Median : 6.50   Median : 4750   Median : 52184
##  Mean   : 6.50   Mean   : 6542   Mean   : 70870
##  3rd Qu.: 9.25   3rd Qu.: 9500   3rd Qu.:100882
##  Max.   :12.00   Max.   :15000   Max.   :158484
```

Insight 1

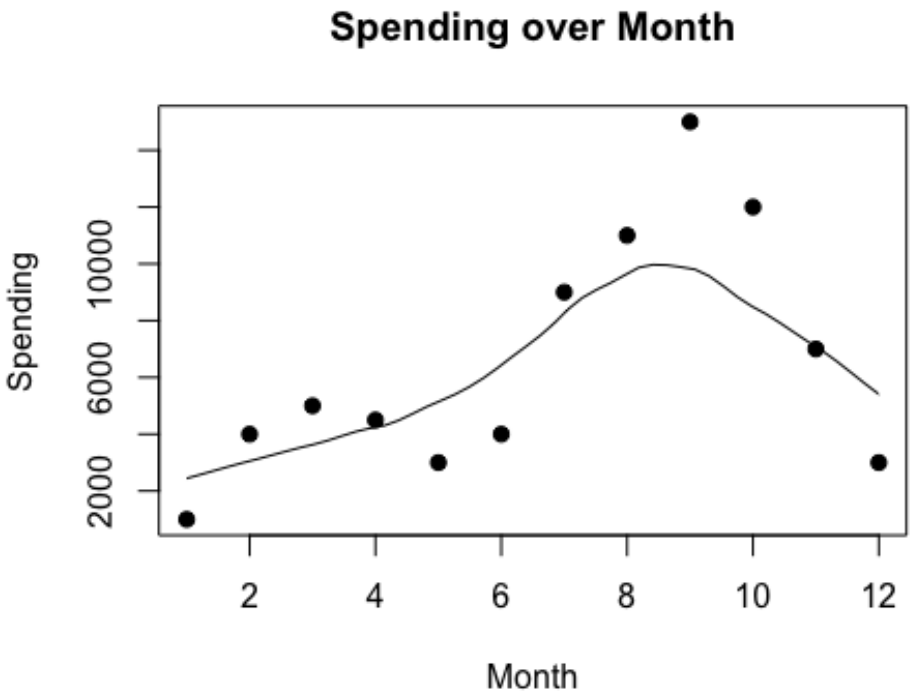
```
    -Marketing Budget is a financial dataset consisting of Spending and Sales records for one complete Fiscal Year.
    -Sales seem to be approximately ten times of Spending by their respective month.
```

Insight 2

```
    -Annual Spending is 15000, and total market Sales turned around 158484.
likely to be ten times profits in the market.
    -September recorded as highest spending and sales.
    -January recorded as lowest spending and sales.
```

2. Use Scatter Plot To Visualise The Relationship.

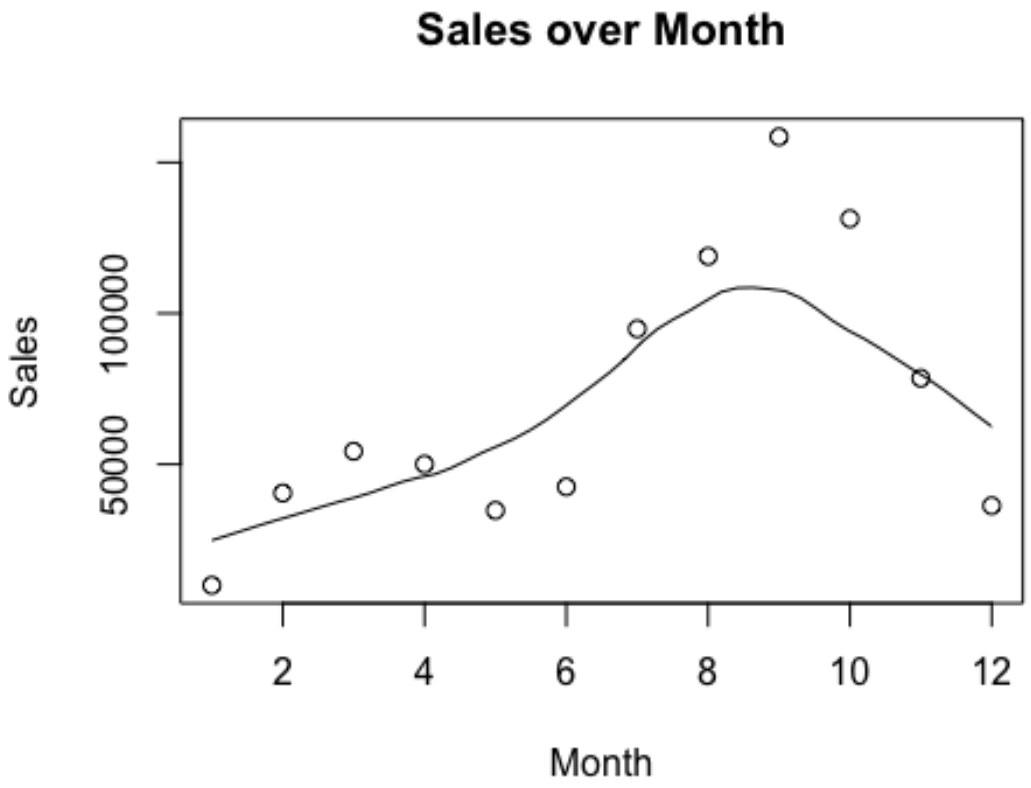
```
scatter.smooth(x = RocketData$Month,
               y = RocketData$Spend,
               xlab = "Month", ylab = "Spending",  # X,Y-Lab
               main = "Spending over Month",       # Title
               col.main = "Black",                 # Title lab color
               cex.lab = "1" ,                     # X,Y-axis Lab color
               pch = 19
)
```



Insight 1

```
    -The distribution is skewed, Spending is Skewed left.
    -January (1) recorded to be the least Spending compare to all other
months.
    -September (9) recorded to be the highest Spending compare to all other
months.
```

```
scatter.smooth(x = RocketData$Month,
               y = RocketData$Sales,
               xlab = "Month", ylab = "Sales",    # X,Y-Lab
               main = "Sales over Month",         # Title
               col.main = "Black",                # Title lab color
               cex.lab = "1"                      # X,Y-axis Lab color
)
```

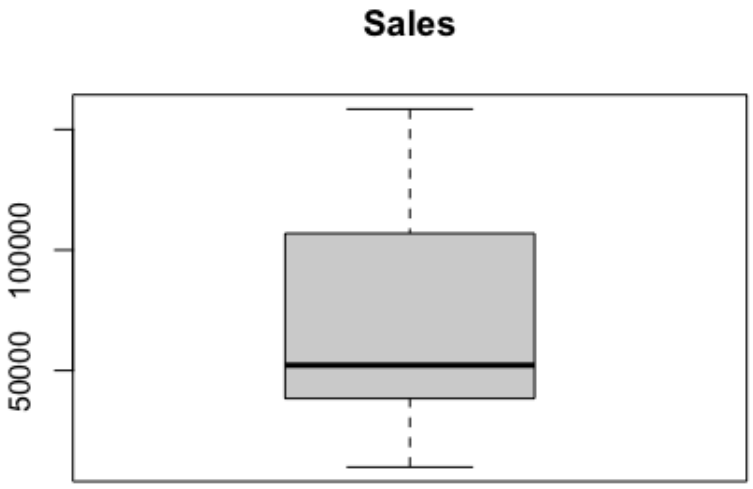


Insight 2

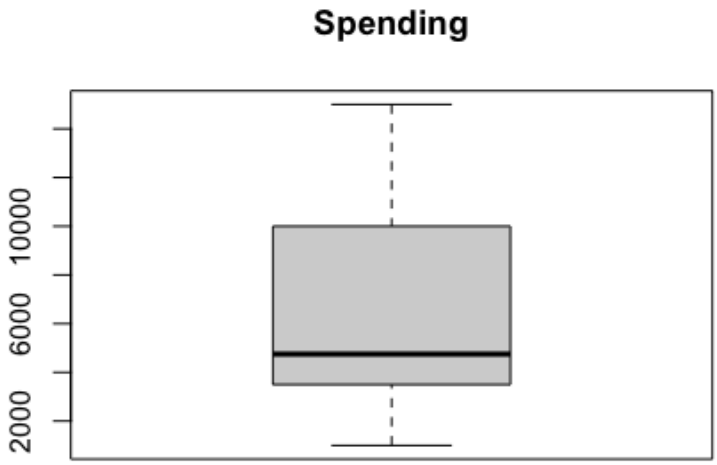
```
    -The distribution is skewed, Sales is Skewed left.
    -July, August, and November recorded to be the average in Sales compare
to all other months.
```

3.Using BoxPlot To Check For Outliers.

```
boxplot(RocketData$Sales, main = "Sales", border = "black" )
```




```
boxplot.default(RocketData$Spend, main = "Spending", border = "black")
```



Insight

```
    -boxplots are useful to detect potential outliers.
    -Clear no outliers to seem to be present in both Sales and Spendings.
```

4. Using Density Plot To Check If Response Variable Is Close To Normal.

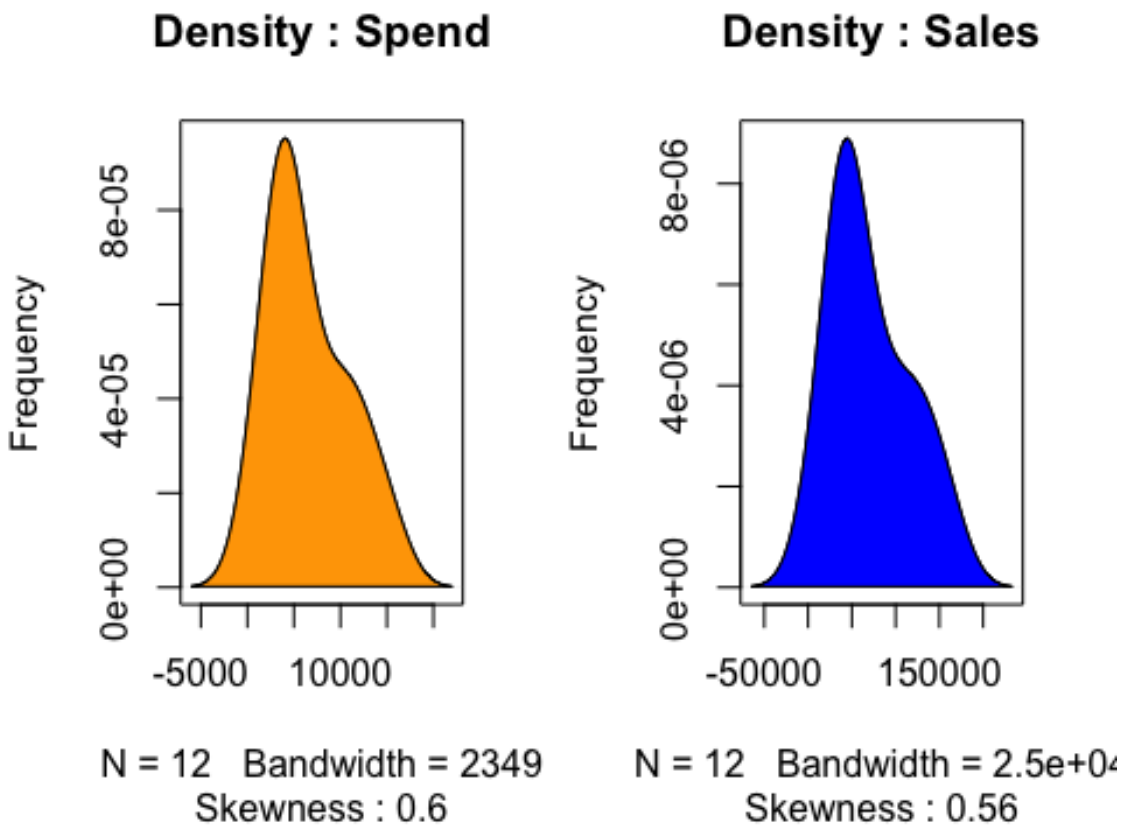
```
# A density plot shows the distribution of a numeric variable.

library(e1071)       # For skewness function
par(mfrow = c(1,2))  # Dividing graph area in 2 columns

# Density plot for 'Spend'
plot(density(RocketData$Spend),
     main = "Density : Spend",
     ylab = "Frequency",
     sub=paste("Skewness :",
               round(e1071::skewness(RocketData$Spend),
                     2
                     )
               )
     )
polygon(density(RocketData$Spend), col = "orange")

# Density plot for 'Sales'
plot(density(RocketData$Sales),
     main = "Density : Sales",
     ylab = "Frequency",
     sub=paste("Skewness :",
               round(e1071::skewness(RocketData$Sales),
                     2
                     )
               )
     )

polygon(density(RocketData$Sales), col = "blue")
```



Insight

5. Check the Correlation Analysis.

```
cor(RocketData$Spend, RocketData$Sales)
```

```
## [1] 0.9988322
```

Insight

```
    -The correlation value is 0.998, which implies Spend and Sales having a
Strong Correlation.
```

6. Build the Linear Regression Model.

```
LinearRegressionModel <- lm(formula = Spend ~ Sales, data = RocketData)
LinearRegressionModel
```

```
##
## Call:
## lm(formula = Spend ~ Sales, data = RocketData)
##
## Coefficients:
## (Intercept)        Sales
##  -114.67027      0.09392
```

```
summary(LinearRegressionModel)

##
## Call:
## lm(formula = Spend ~ Sales, data = RocketData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -293.22 -165.15  -20.82  188.67  312.02
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.147e+02  1.196e+02  -0.959     0.36
## Sales        9.392e-02  1.437e-03  65.378 1.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 217.5 on 10 degrees of freedom
## Multiple R-squared:  0.9977, Adjusted R-squared:  0.9974
## F-statistic:  4274 on 1 and 10 DF,  p-value: 1.707e-14
```

Insight

```
    -We could see that the value of the Intercept of the model is -114.67027
and the Slope of the model is 0.09392. So, we get the complete formula of the
linear model as, Sales = Slope*(Spend) - Intercept
```

Sales = 0.09392*(Spend) - 114.67027

7. Using p-value Check For Statistical Significance.

```
t.test(RocketData$Spend, RocketData$Sales,
       paired = TRUE,
       alternative = "two.sided"
       )

##
##  Paired t-test
##
## data:  RocketData$Spend and RocketData$Sales
## t = -5.3869, df = 11, p-value = 0.000221
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -90612.01 -38045.32
## sample estimates:
## mean of the differences
##               -64328.67
```

8.Capture the summary of the linear model.

```
ModelSummary <-summary(LinearRegressionModel)
ModelSummary
```

```
##
## Call:
## lm(formula = Spend ~ Sales, data = RocketData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -293.22 -165.15  -20.82  188.67  312.02
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.147e+02  1.196e+02  -0.959     0.36
## Sales        9.392e-02  1.437e-03  65.378 1.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 217.5 on 10 degrees of freedom
## Multiple R-squared:  0.9977, Adjusted R-squared:  0.9974
## F-statistic:  4274 on 1 and 10 DF,  p-value: 1.707e-14
```

Insight

-The COEFFICIENTS show the linear regression model's intercept and slope, and it is given by 'Sales 0.09392 * (Spend) - 114.67027'. The estimate of sales, when the spend is 0, is 9.392e-02. The standard error coefficient is 1.437e-03, which gives the approximate variations that the sales variable can have concerning the spending variable. The t-value is 65.378, which is significantly far from 0, which shows that the variables are statistically significant.

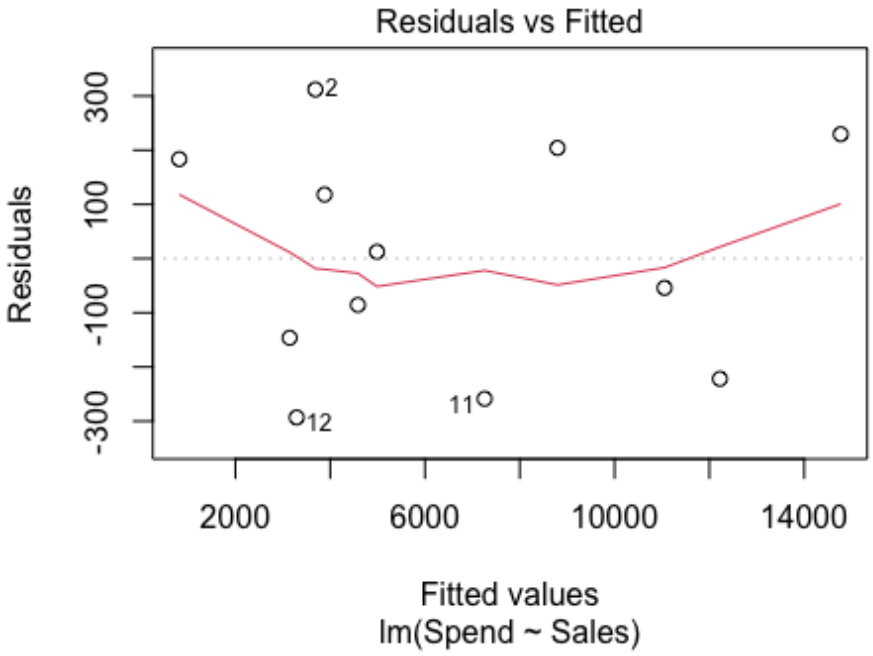
-Residual Standard Error is a measure of the quality of a linear regression fit, and here, it is calculated as 217.5 on 10 degrees of freedom.

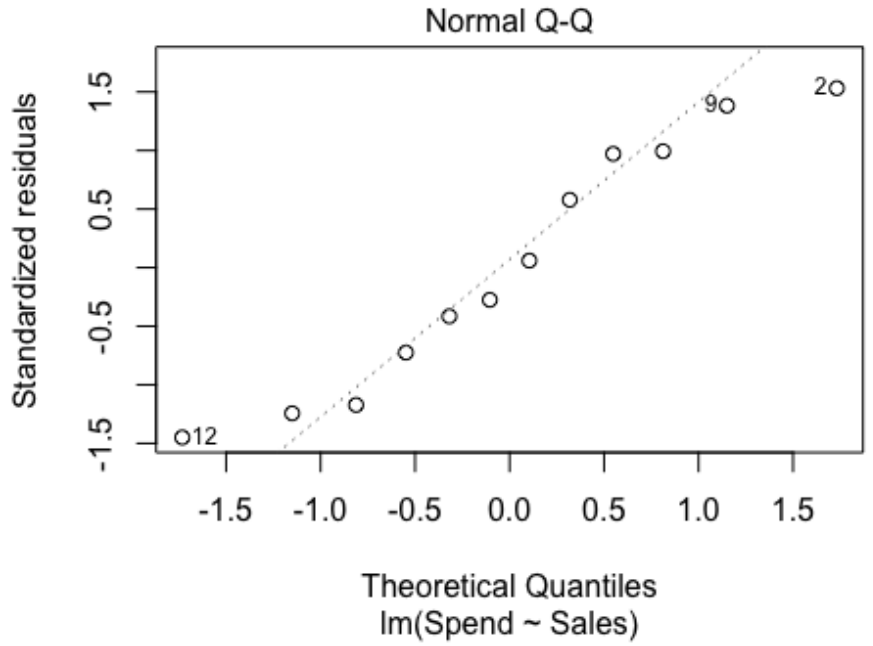
-The R-squared (R^2) statistic provides a measure of how well the model fits the actual data.R^2 is a measure of the linear relationship between our predictor variable and our response/target variable. The value obtained is 0.9977, which shows a 99.77% fitting of the variable to the linear model.

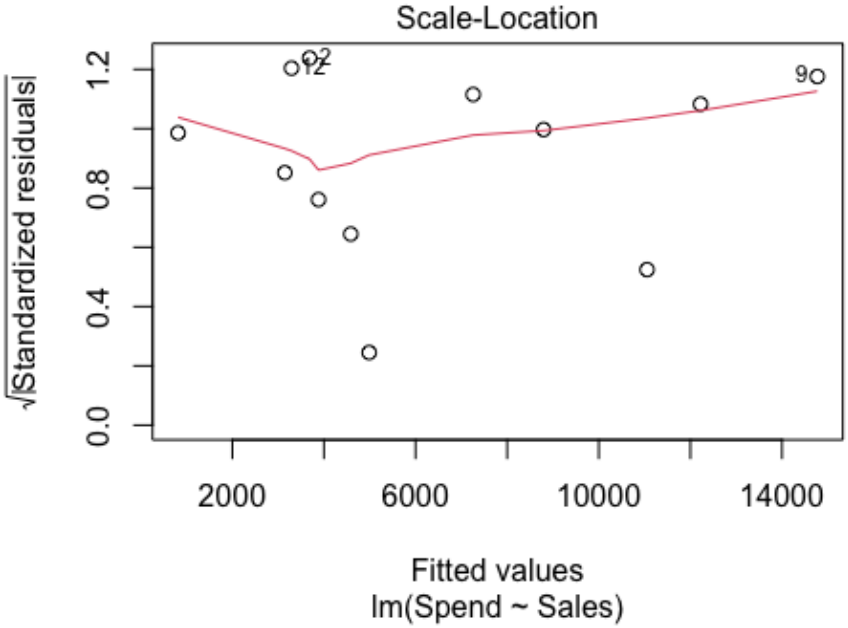
-The F value of 4274 on 1 is relatively larger than 1; hence, we see a good relationship between sales and spend variables.

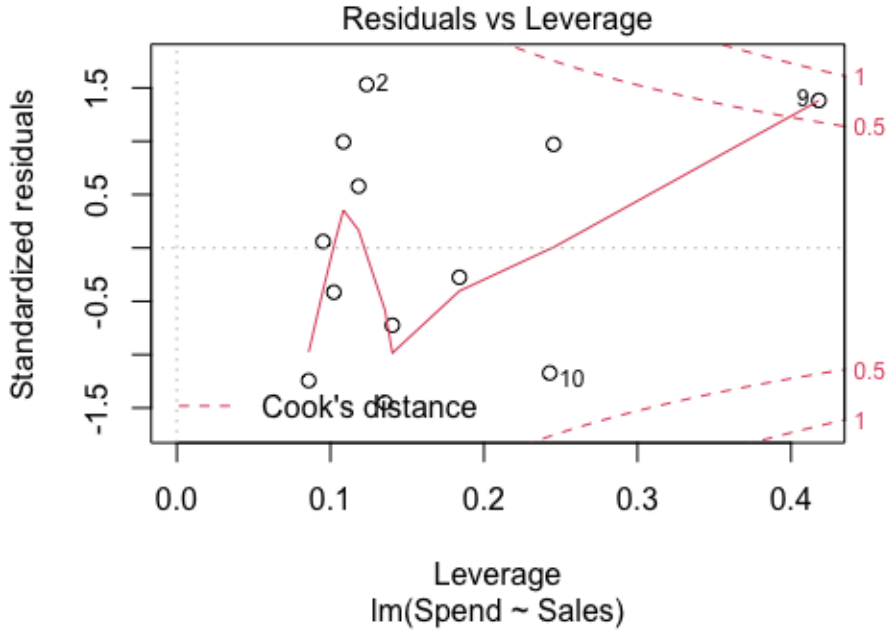
9. Also perform the Linear Diagnostics for the given data set.(Hint: plot(lmmodel))

```
plot(LinearRegressionModel)
```


Residuals vs Fitted
Residuals
300
100
-100
-300
2
12
11
2000
6000
10000
14000
Fitted values
lm(Spend ~ Sales)


Normal Q-Q
Standardized residuals
1.5
0.5
-0.5
-1.5
12
9
2
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
lm(Spend ~ Sales)


Scale-Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
2000
6000
10000
14000
Fitted values
lm(Spend ~ Sales)


Residuals vs Leverage
Standardized residuals
1.5
0.5
-0.5
-1.5
0.0
0.1
0.2
0.3
0.4
Cook's distance
1
0.5
0.5
1
Leverage
lm(Spend ~ Sales)

9. Create the training and test data (70:30).

```
set.seed(100)
rows = sample(nrow(RocketData))

# Randomly order data:
data = RocketData[rows, ]

# Identify row to split on: split
split = round(nrow(data) * .70)

# Create train
train = data[1:split,]
train
```

```
##    Month Spend  Sales
## 10    10 12000 131348
## 7      7  9000  94871
## 6      6  4000  42551
## 3      3  5000  54324
## 1      1  1000   9914
## 2      2  4000  40487
## 12    12  3000  36284
## 4      4  4500  50044
```

```
# Create test
test = data[(split + 1): nrow(data),]
test
```

```
##    Month Spend  Sales
## 9      9 15000 158484
## 11    11  7000  78504
## 5      5  3000  34719
## 8      8 11000 118914
```

Insight

```
    -We are splitting the dataset for training and testing purposes by 70:30
weightage.
```

10. Fit the model on training data and predict sales on test data.

```
Model = lm(formula = Spend ~ Sales, data = train )
Model

##
## Call:
## lm(formula = Spend ~ Sales, data = train)
##
## Coefficients:
## (Intercept)        Sales
##    14.55154      0.09217

Predict = predict(Model, test)
```

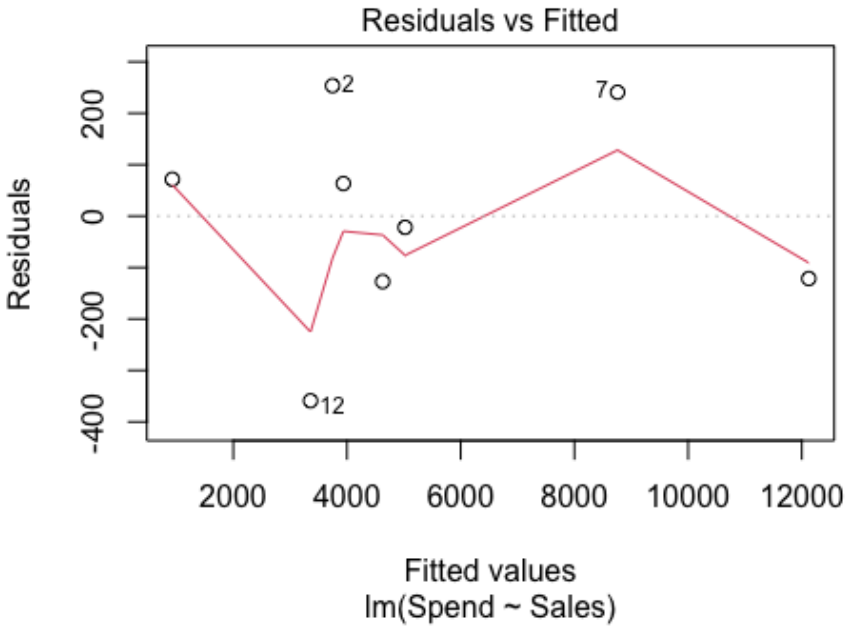
Insight

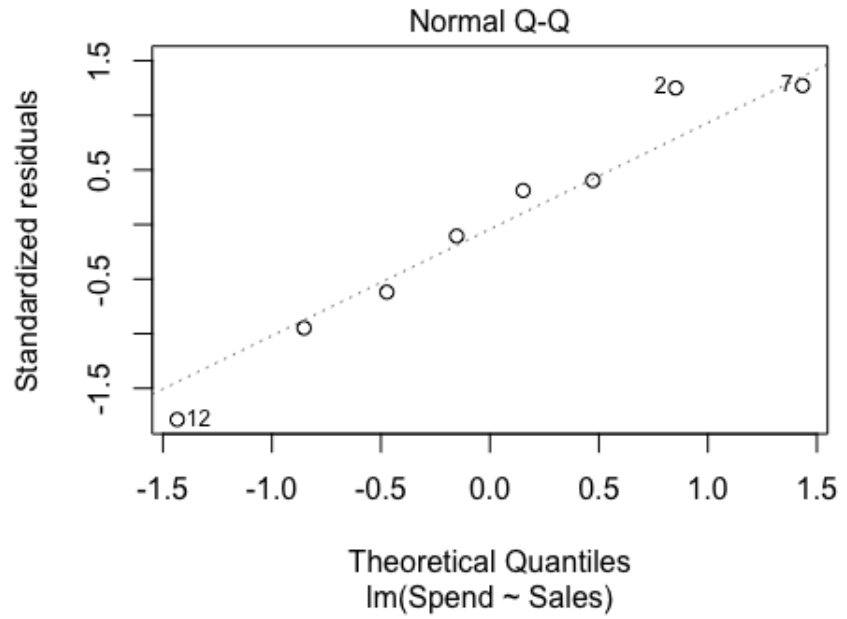
```
    -We could see that the value of the Intercept of the model is 14.55154
and the Slope of the model is 0.09217. So, we get the complete formula of the
linear model as, Sales = Slope * (Spend) - Intercept
```

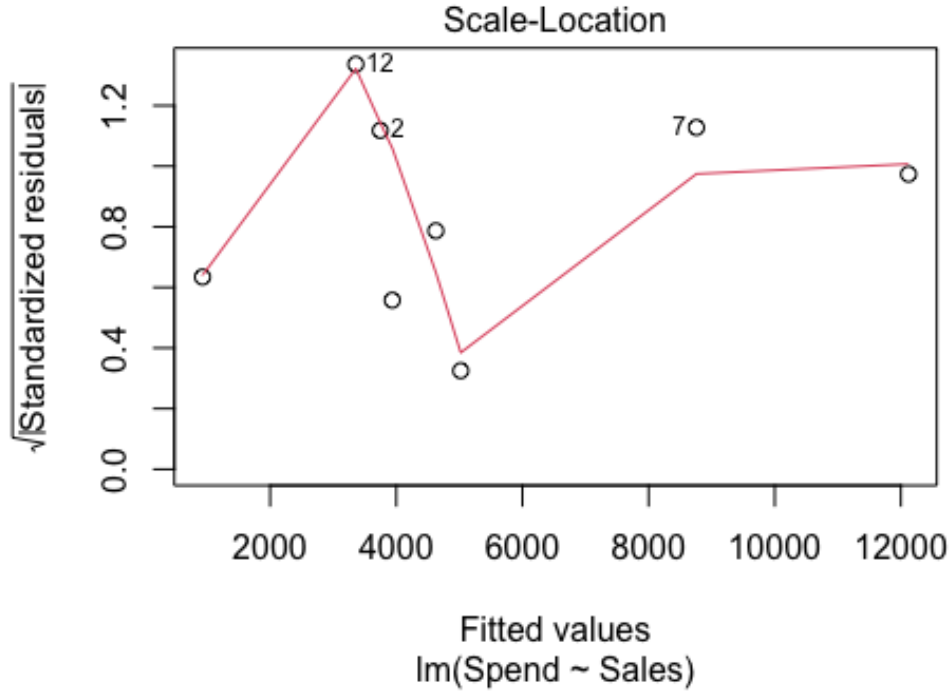
Sales = 0.09217 * (Spend) - 14.55154

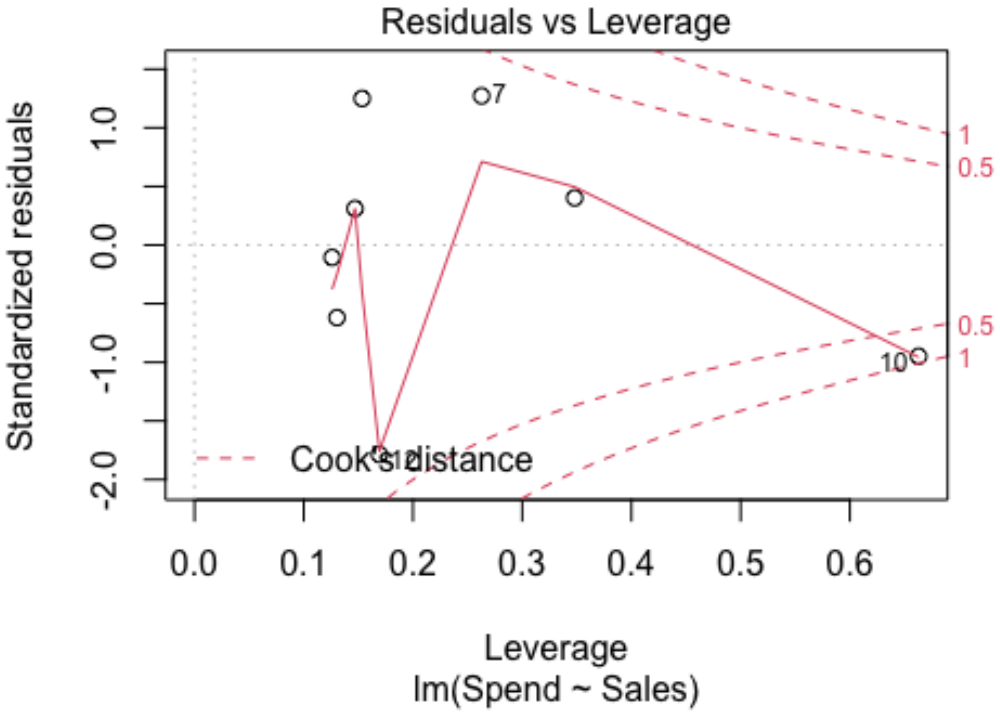
11.Review the diagnostic measures.

```
plot(Model)
```


Residuals vs Fitted
Residuals
200
0
-200
-400
2
7
12
2000
4000
6000
8000
10000
12000
Fitted values
lm(Spend ~ Sales)


Normal Q-Q
Standardized residuals
1.5
0.5
-0.5
-1.5
2
7
12
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
lm(Spend ~ Sales)


Scale-Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
12
2
7
2000
4000
6000
8000
10000
12000
Fitted values
lm(Spend ~ Sales)


Residuals vs Leverage
Standardized residuals
-2.0
-1.0
0.0
1.0
7
1
0.5
0.5
1
10
Cook's distance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Leverage
lm(Spend ~ Sales)