MATH1324 Assignment 3

Vinay Nagamangala Rame Gowda

June 2, 2019

- Import the File onto R Environment
- Data Exploration
    - Understand the distribution of data
- Running the ANOVA model — Analysis of Variance Table
- Diagnostic Check to check if the assumptions have been satisfied
- Interpretations
- Shapiro-Wilk Test for Normality

Import the File onto R Environment

Data Exploration

```
summarizeColumns(ad_data)
```

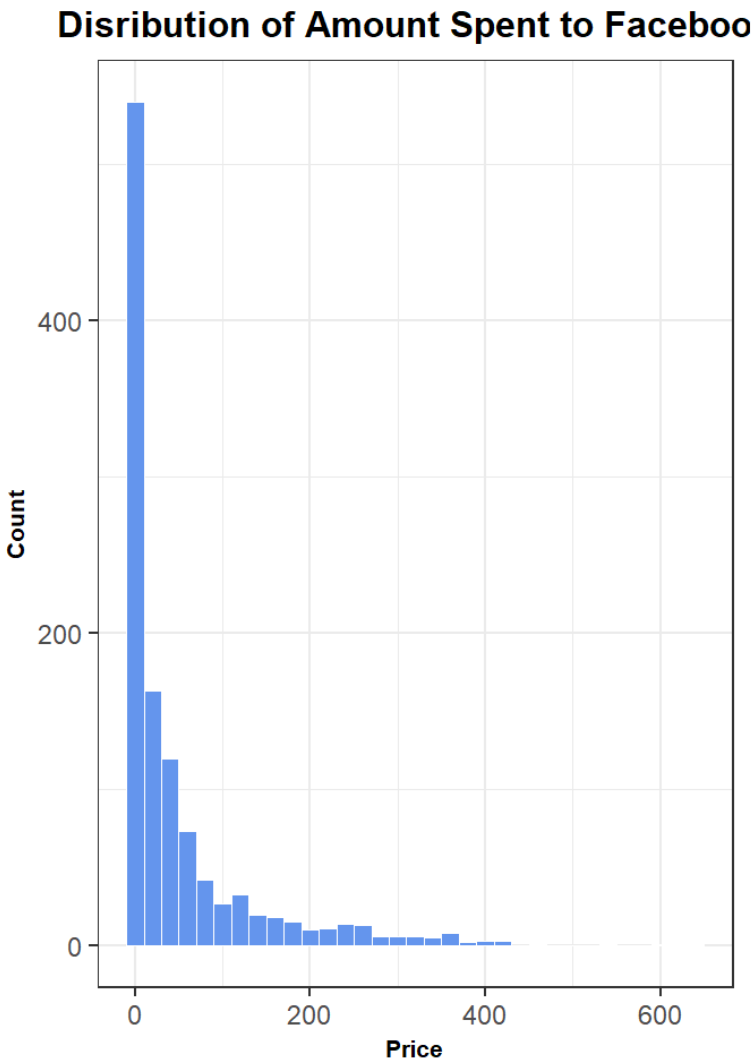
```
##                   name    type na         mean         disp     median
## 1                ad_id integer  0 9.872611e+05 1.939928e+05 1121185.00
## 2      xyz_campaign_id integer  0 1.067382e+03 1.216294e+02    1178.00
## 3       fb_campaign_id integer  0 1.337840e+05 2.050031e+04  144549.00
## 4                  age  factor  0           NA 6.272966e-01         NA
## 5               gender  factor  0           NA 4.820647e-01         NA
## 6             interest integer  0 3.276640e+01 2.695213e+01      25.00
## 7          Impressions integer  0 1.867321e+05 3.127622e+05   51509.00
## 8               Clicks integer  0 3.339020e+01 5.689244e+01       8.00
## 9                Spent numeric  0 5.136066e+01 8.690842e+01      12.37
## 10    Total_Conversion integer  0 2.855643e+00 4.483593e+00       1.00
## 11 Approved_Conversion integer  0 9.440070e-01 1.737708e+00       1.00
##              mad    min        max nlevs
## 1  252016.79580 708746 1314415.00     0
## 2       0.00000    916    1178.00     0
## 3   31018.95720 103916  179982.00     0
## 4            NA    210     426.00     4
## 5            NA    551     592.00     2
## 6      10.37820      2     114.00     0
## 7   74063.28300     87 3052003.00     0
## 8      11.86080      0     421.00     0
## 9      18.33976      0     639.95     0
## 10      0.00000      0      60.00     0
## 11      1.48260      0      21.00     0
```

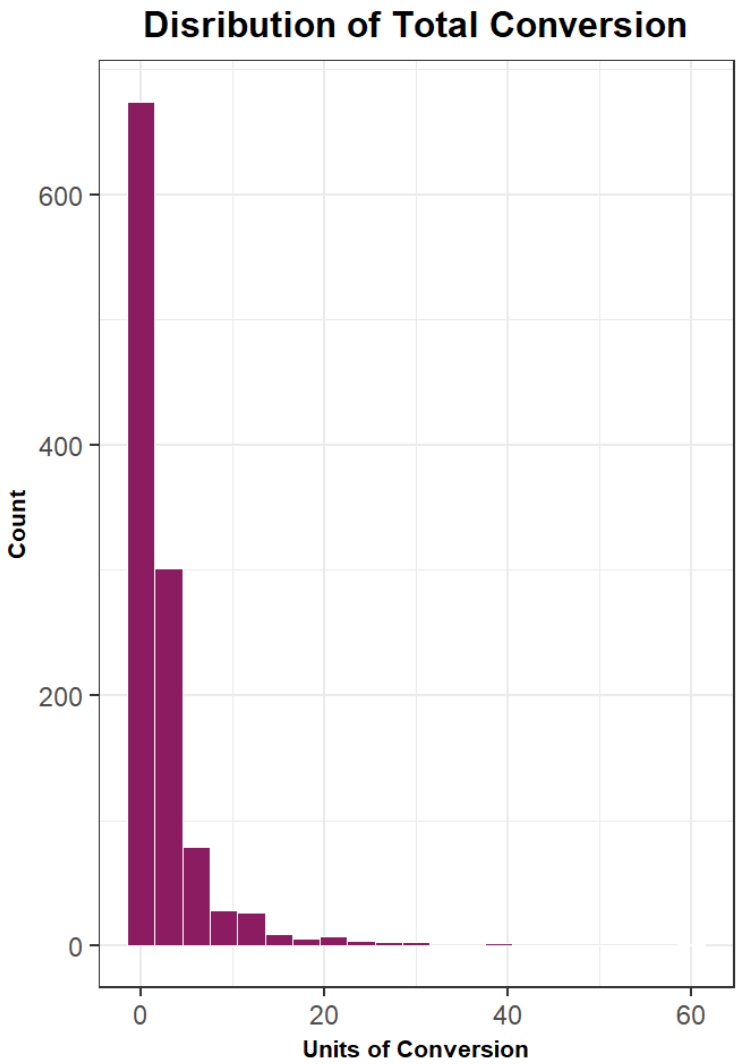
Summarising the data, there is a need to remove or subset only necessary columns.

```
ad_data_cleaned<-ad_data[,c(9:10)]
```

Understand the distribution of data

```
black.bold <- element_text(face = "bold", color = "black", size = 10)

# parameters to plot distribution
n = nrow(ad_data_cleaned)
mean = mean(ad_data_cleaned$Spent)
sd = sd(ad_data_cleaned$Spent)
binwidth = 20 # passed to geom_histogram and stat_function

spent<-ggplot(ad_data_cleaned,aes(Spent,mean=mean,sd=sd,binwidth=binwidth,n=n))+
  geom_histogram(binwidth = binwidth,size = 0.1,colour = "white", fill = "cornflowerblue")+
theme_bw()+
  labs(title="Disribution of Amount Spent to Facebook",
            x ="Price",y="Count")+
  theme(title = element_text(face = "bold", color = "black", size = 10),
        axis.title.x = element_text(face="bold", colour = "black", size = 8),
        axis.title.y= element_text(face="bold", colour = "black", size = 8),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle =element_text(hjust = 0.5))


n1 = nrow(ad_data_cleaned)
mean1 = mean(ad_data_cleaned$Total_Conversion)
sd1 = sd(ad_data_cleaned$Total_Conversion)
binwidth1 = 3 # passed to geom_histogram and stat_function

total_con<-ggplot(ad_data_cleaned,aes(Total_Conversion,mean=mean1,sd=sd1,binwidth=binwidth
1,n=n1))+
  geom_histogram(binwidth = binwidth1,size = 0.1,colour = "white", fill = "maroon4")+theme_
bw()+
  labs(title="Disribution of Total Conversion",
            x ="Units of Conversion",y="Count")+
  theme(title = element_text(face = "bold", color = "black", size = 10),
        axis.title.x = element_text(face="bold", colour = "black", size = 8),
        axis.title.y= element_text(face="bold", colour = "black", size = 8),
plot.title = element_text(hjust = 0.5),
        plot.subtitle =element_text(hjust = 0.5))



grid.arrange(spent, total_con, nrow = 1)
```

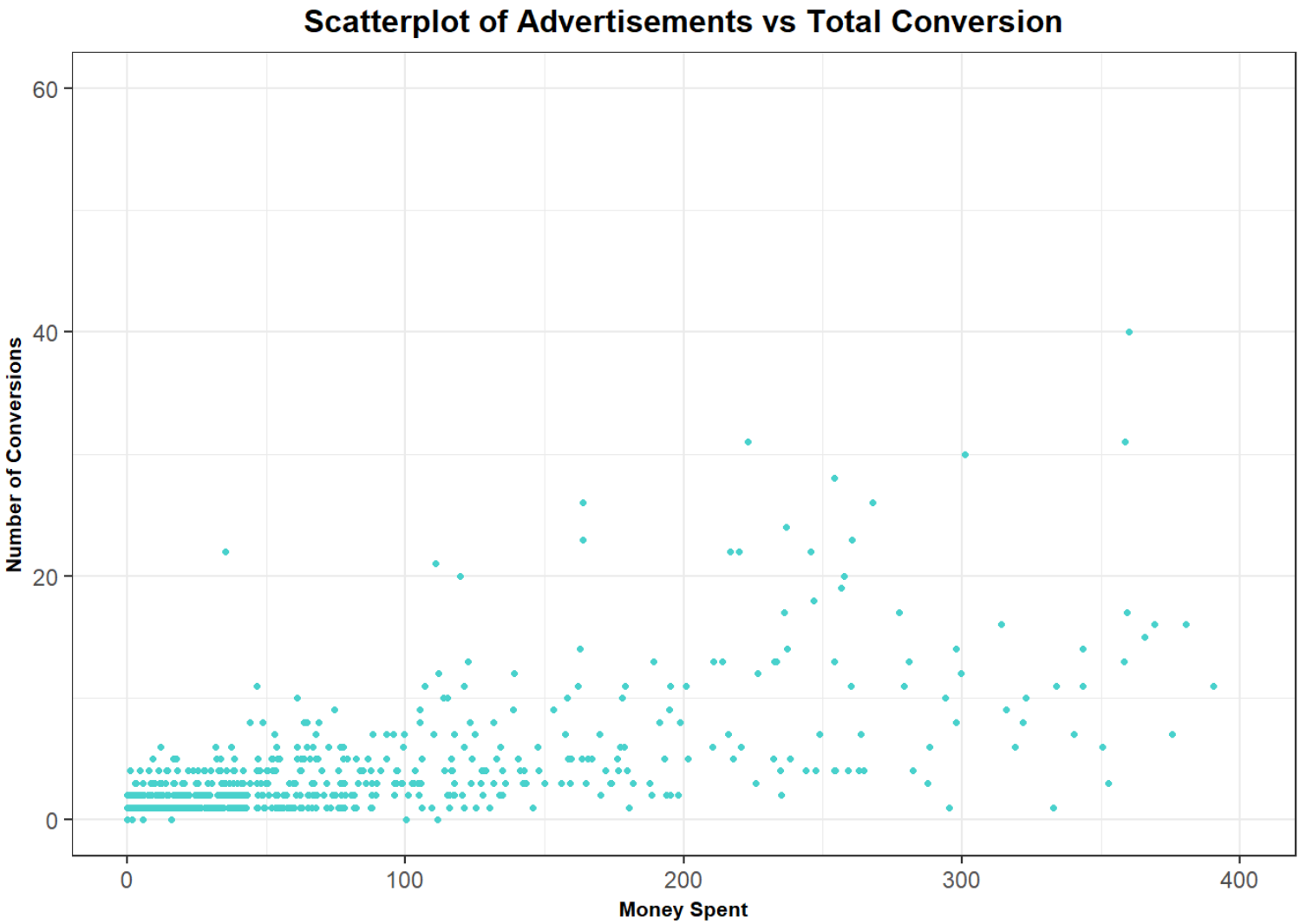




Let us plot a 1-1 relationship between Spent and Total Conversion.

```
spent_totcon<-ggplot(ad_data_cleaned,aes(x=Spent,y=Total_Conversion))+
  geom_point(color="mediumturquoise",size = 0.9)+theme_bw()+
  labs(title="Scatterplot of Advertisements vs Total Conversion",
             x ="Money Spent",y="Number of Conversions")+
  theme(title = element_text(face = "bold", color = "black", size = 10),
        axis.title.x = element_text(face="bold", colour = "black", size = 8),
        axis.title.y= element_text(face="bold", colour = "black", size = 8),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle =element_text(hjust = 0.5)) + xlim(0,400)
spent_totcon
```

```
## Warning: Removed 10 rows containing missing values (geom_point).
```

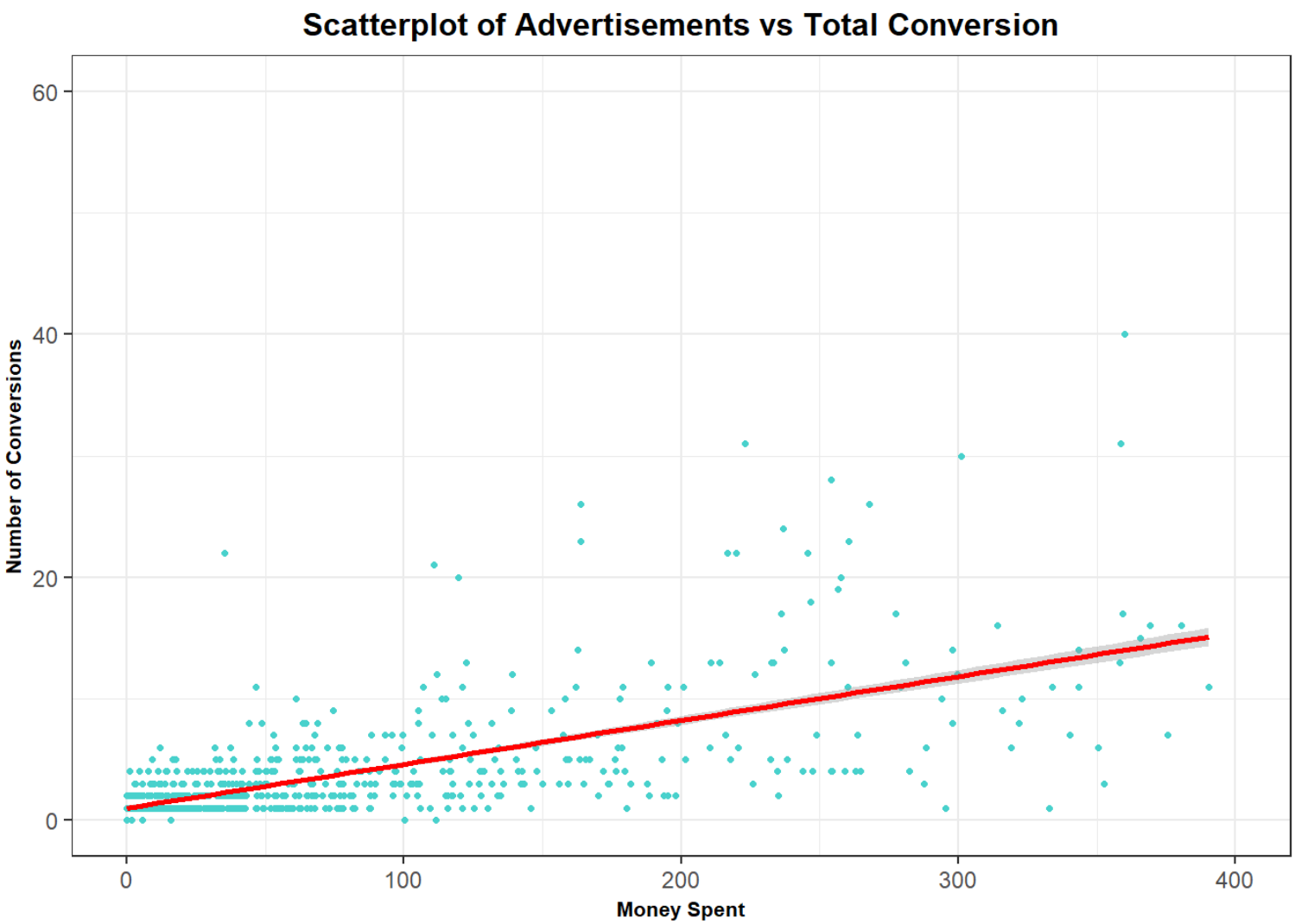


Adding a regression line to the Scatterplot.

```
spent_totcon+stat_smooth(method="lm",
              col="red",
              size=1)
```

```
## Warning: Removed 10 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 10 rows containing missing values (geom_point).
```



From the above visualisations, there seems to be a positive correlation between Money Spent and Total Conversion Proving this using the Pearson's Correlation Coefficient Test.

```
cor(ad_data_cleaned$Spent,ad_data_cleaned$Total_Conversion)
```

```
## [1] 0.7253794
```

From the Pearson's Correlation Coefficient, it is clear that there is a "Strong Positive Correlation" between Money Spent and Total Conversion with a corelation value of 0.72.

Fitting a simple linear regression model to the data to examine the relationship between independent variable x (Money Spent) and the dependent variable (Total Conversion) y

H0 - The Null Hypothesis —- Amount Spent on Ad Campaigns does not increase Conversion to Sales. HA- The Alternate Hypothesis —- Amount Spent on Ad Campaigns increases Conversion to Sales.

```
lm_model<-lm(formula = Total_Conversion~Spent,data = ad_data_cleaned)
summary(lm_model)
```

```
##
## Call:
## lm(formula = Total_Conversion ~ Spent, data = ad_data_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.395  -0.576  -0.007   0.066  35.118
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.933613   0.106096     8.8   <2e-16 ***
## Spent       0.037422   0.001051    35.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.088 on 1141 degrees of freedom
## Multiple R-squared:  0.5262, Adjusted R-squared:  0.5258
## F-statistic:  1267 on 1 and 1141 DF,  p-value: < 2.2e-16
```

From the summary of the above linear regression, the coefficient of predictor spent seems to be statistically significant. The final equation can be written as follows:
Total_Conversion = 0.933613 + (0.037422)Spent + Residuals

Running the ANOVA model — Analysis of Variance Table

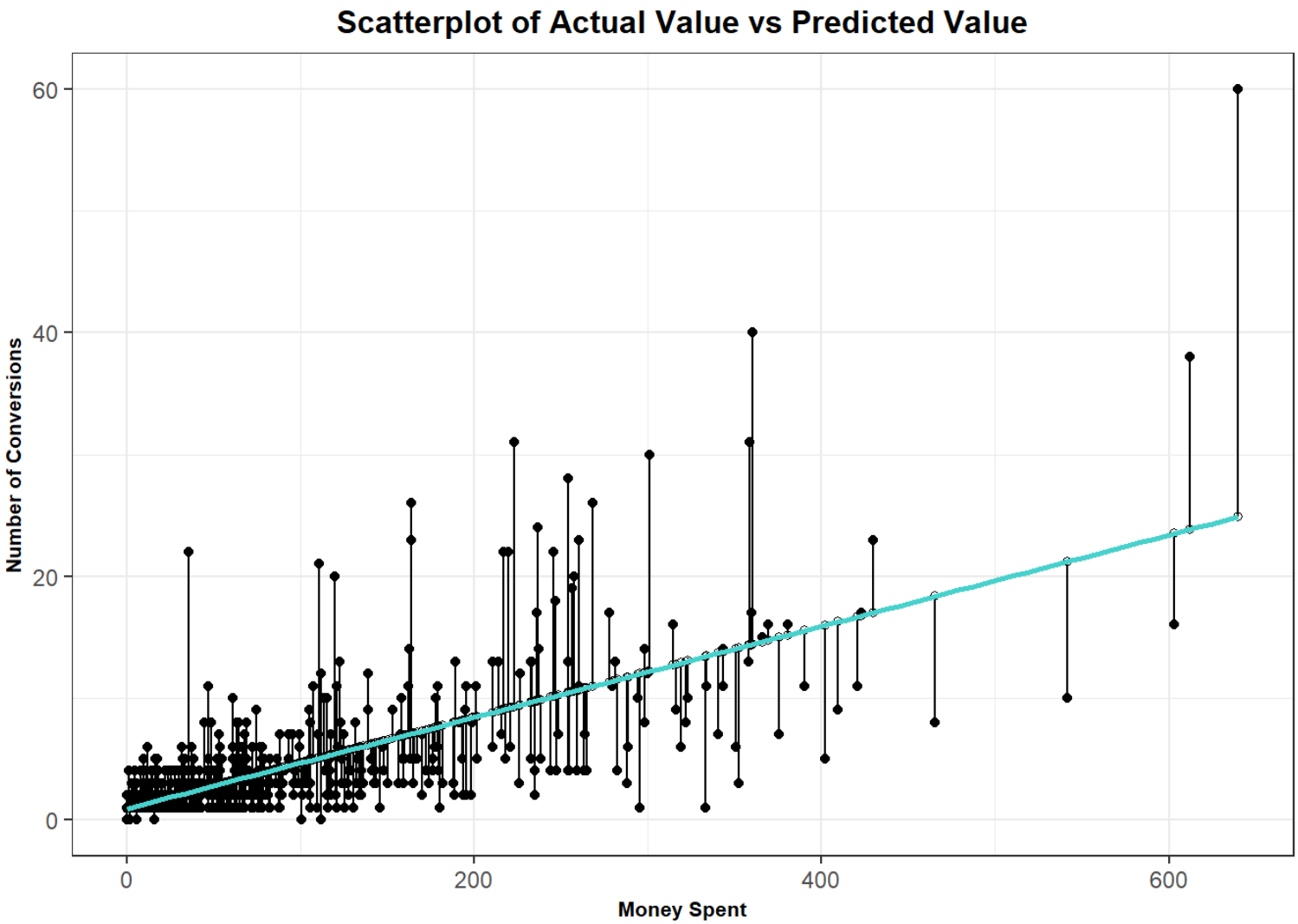
```
anova(lm_model)
```

```
## Analysis of Variance Table
##
## Response: Total_Conversion
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Spent        1  12080 12079.5  1267.1 < 2.2e-16 ***
## Residuals 1141  10878     9.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

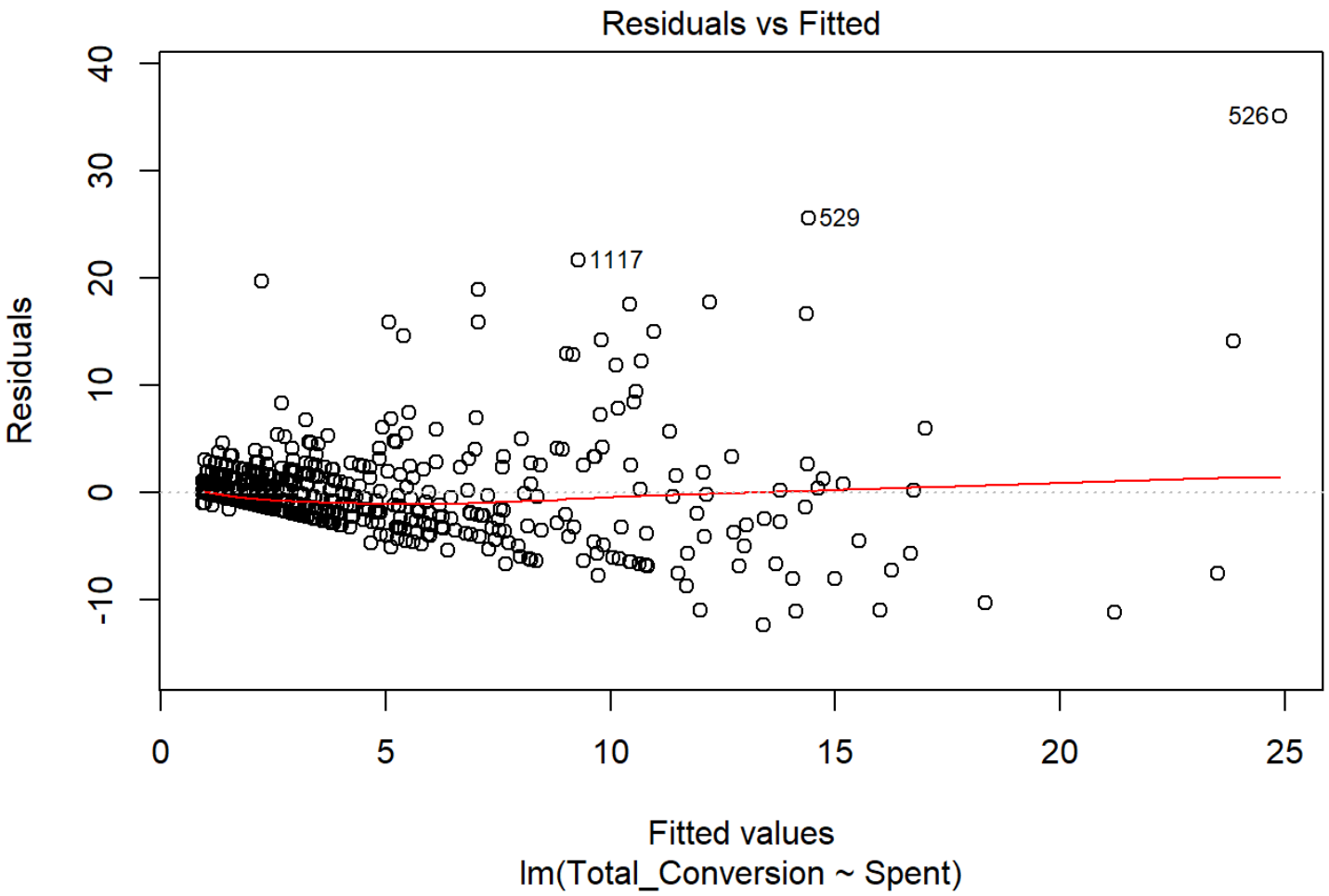
From the ANOVA conducted, the model seems to be adequate (because of the large F value).

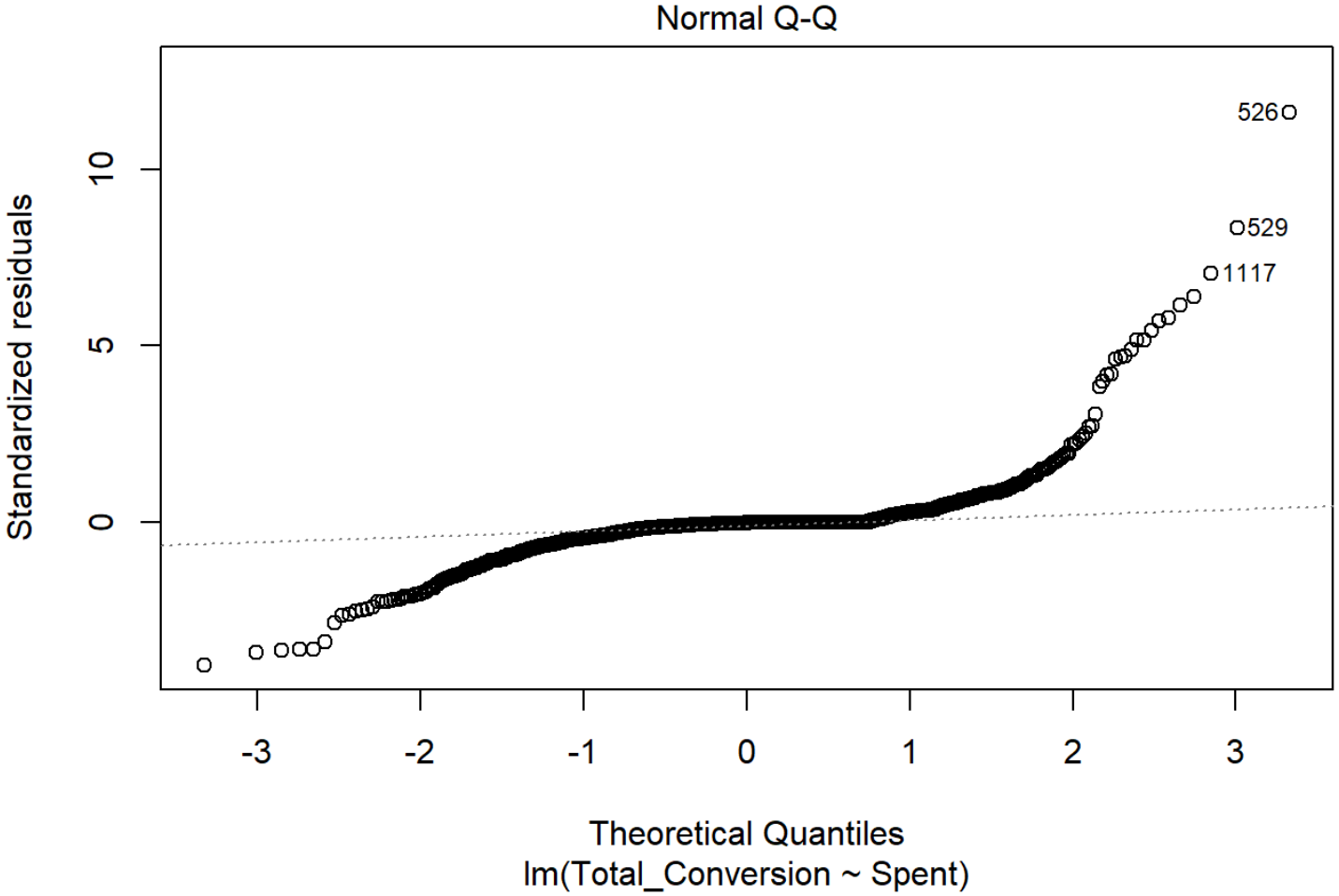
Diagnostic Check to check if the assumptions have been satisfied

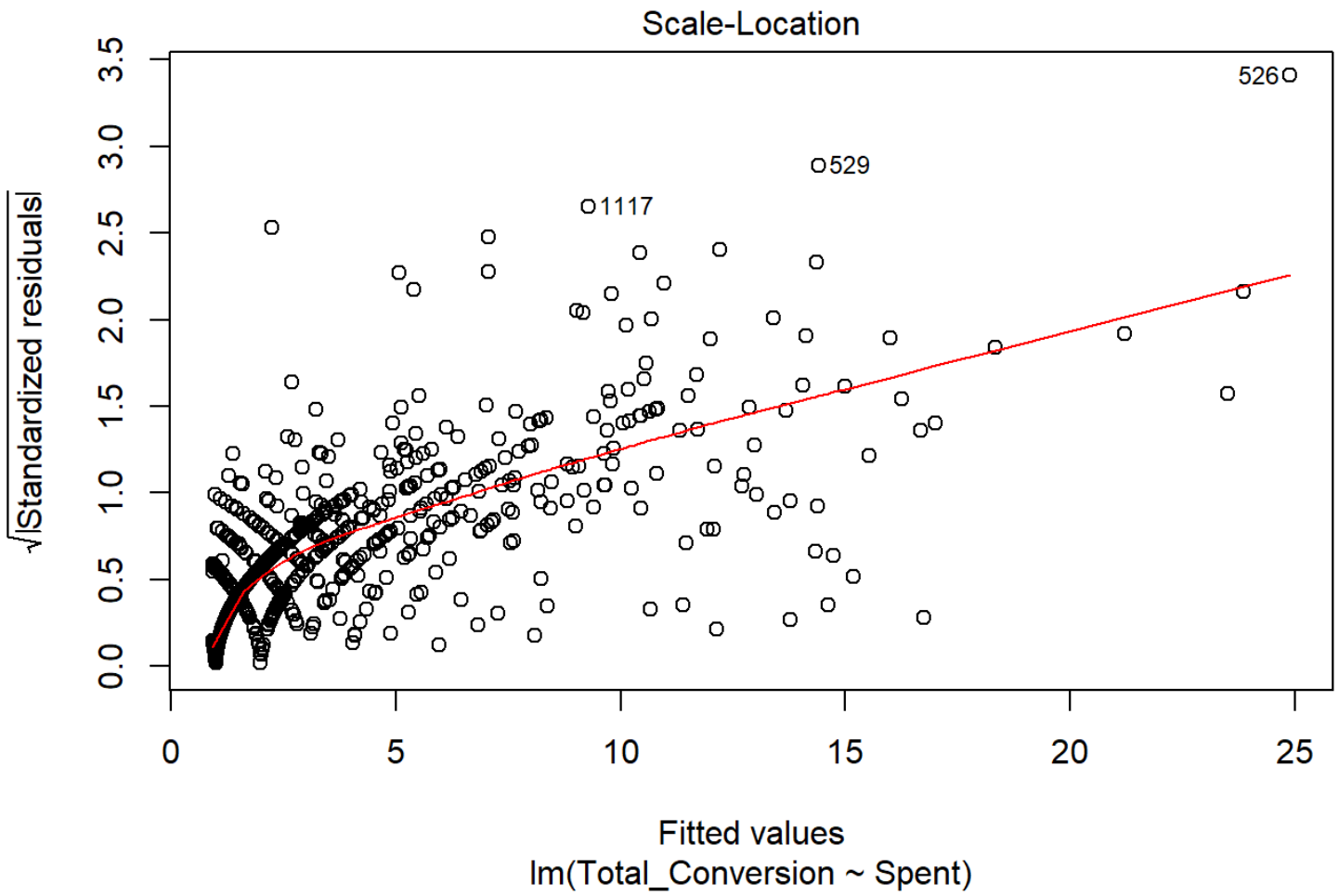
```
d<-ad_data_cleaned
d$pred<-predict(lm_model)
d$residuals<-residuals(lm_model)
ggplot(d,aes(x=Spent,y=Total_Conversion))+
  geom_segment(aes(xend=Spent,yend=pred))+
  geom_point()+
  geom_point(aes(y=pred),shape=1)+theme_bw()+
  labs(title="Scatterplot of Actual Value vs Predicted Value",
            x ="Money Spent",y="Number of Conversions")+
  theme(title = element_text(face = "bold", color = "black", size = 10),
        axis.title.x = element_text(face="bold", colour = "black", size = 8),
        axis.title.y= element_text(face="bold", colour = "black", size = 8),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle =element_text(hjust = 0.5))+
  geom_smooth(method = "lm", se = FALSE, color = "mediumturquoise")
```

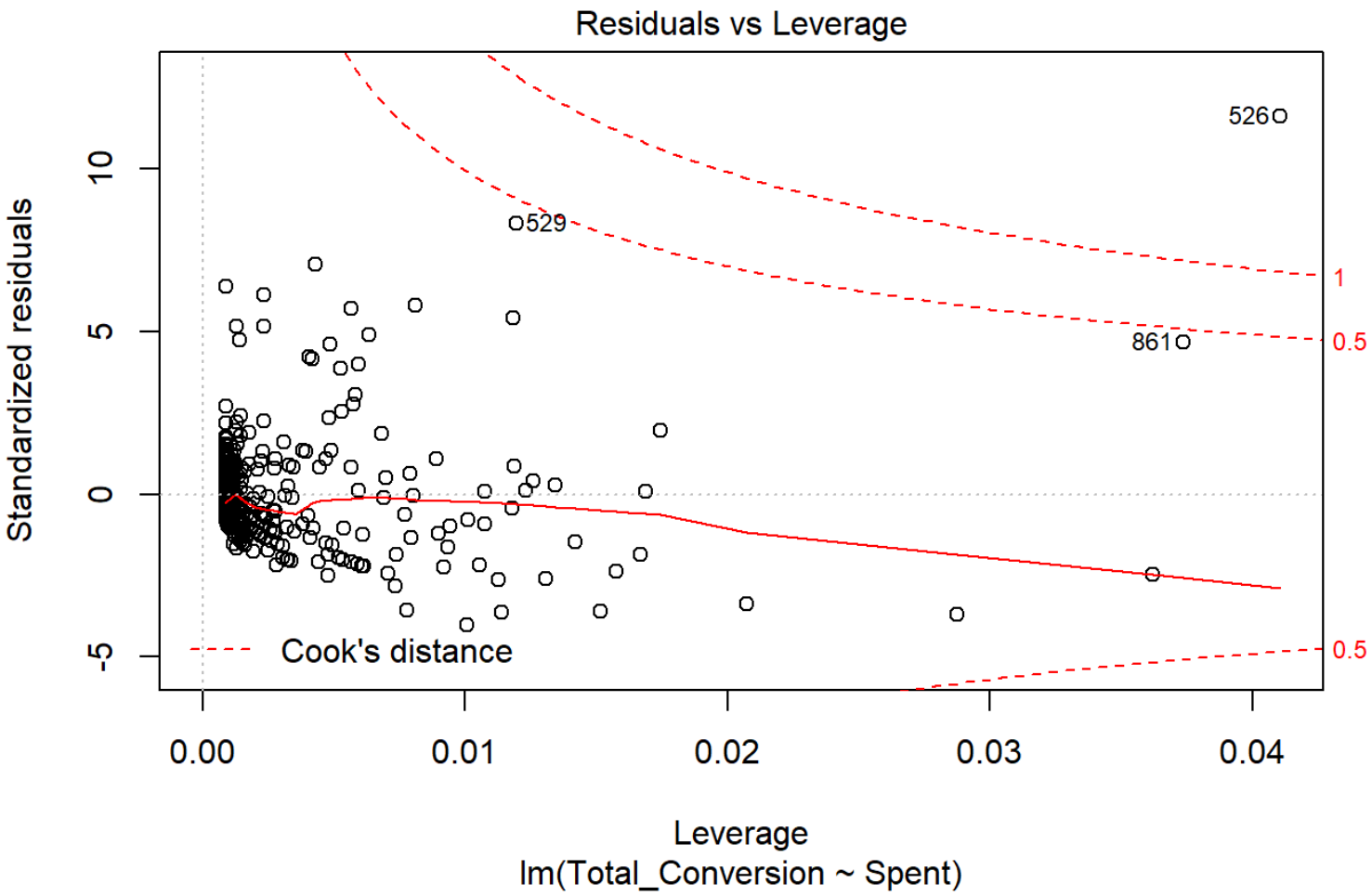



```
plot(lm_model)
```


Residuals vs Fitted
Residuals
40
30
20
10
0
-10
526
529
1117
0
5
10
15
20
25
Fitted values
lm(Total_Conversion ~ Spent)


Normal Q-Q
Standardized residuals
10
5
0
526
529
1117
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Total_Conversion ~ Spent)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
526
529
1117
0
5
10
15
20
25
Fitted values
lm(Total_Conversion ~ Spent)



Interpretations

1. The assumption of homoscedasticity or constant variance seems to be followed
2. There seems to be a violation here, the data does not seem to be normally distributed, however, we will run the Shapiro-Wilk to confirm this
3. From the Cook's distance plot, we see that there are a few points on the red bands, while 1 point that is beyond the red band.

Shapiro-Wilk Test for Normality

```
r1<-residuals(lm_model)
shapiro.test(r1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  r1
## W = 0.63244, p-value < 2.2e-16
```

The results of the Shapiro-Wilk are, we reject the null hypothesis. The residuals are not distributed **"The Assumption of Normality has been violated"**

Thus the Forecasts, confidence intervals and scientific insights given in the above regression model might be inefficient, or it could be biased and misleading.

With the given data, we fail to reject the null hypothesis that there is an increase in Conversion of Sales with an Increase in Amount Spent on Ad Campaigns.