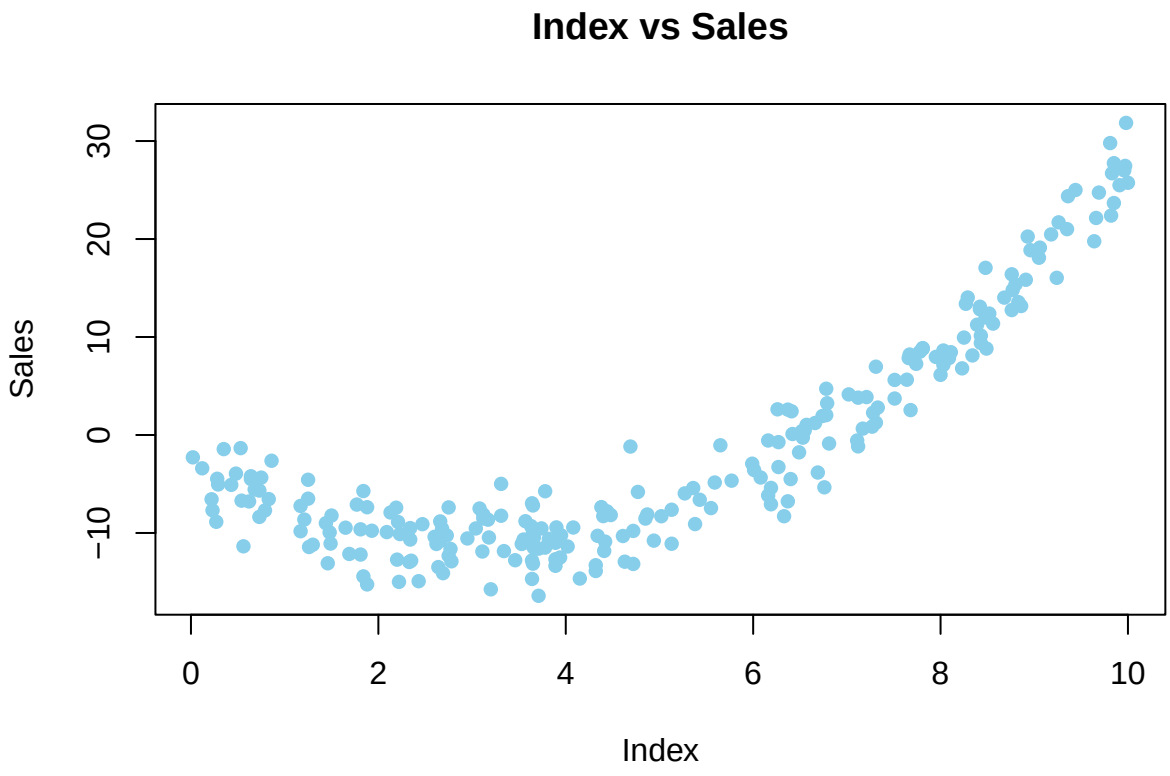
Applied Statistics: Assignment # 01

Question # 01):

Part a):

The scatter plot is created between index and sales and it can be seen that there is no linear relationship between index and sales. The shape is a bit U type and hence it says that there is no linearity among both variables.

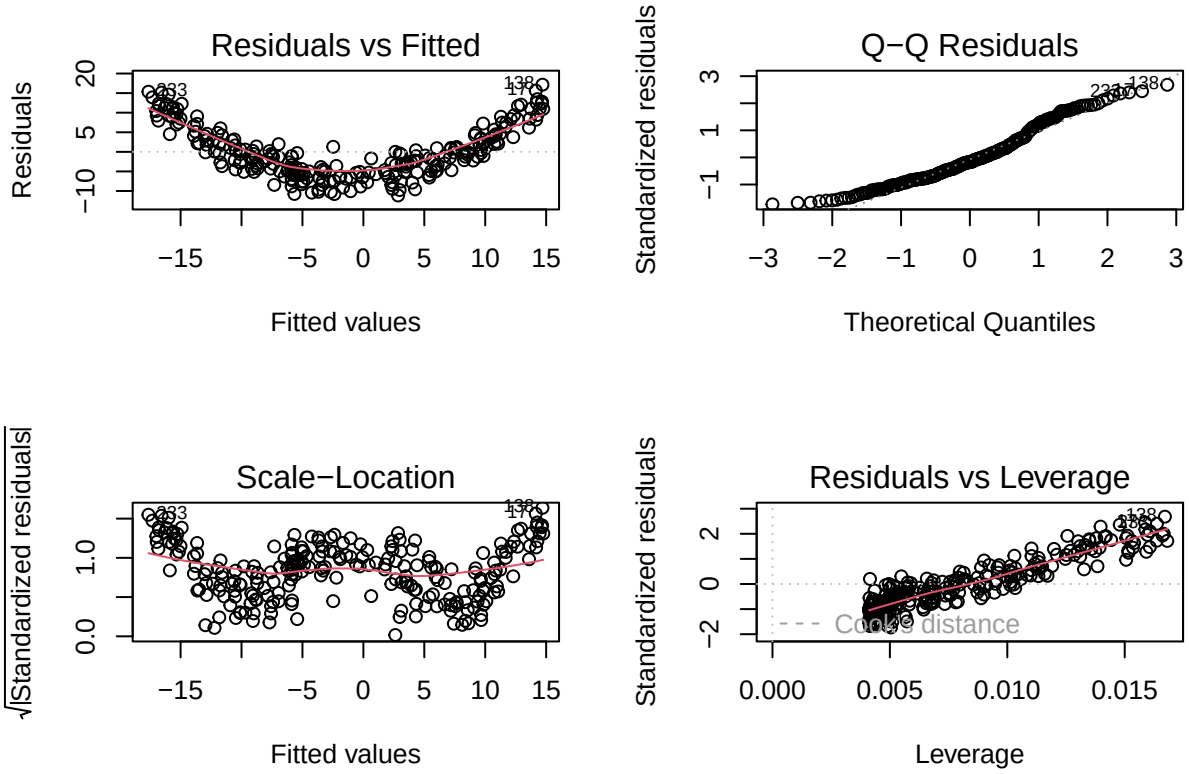
```
library(dplyr)

data <- read.csv("sales.csv")
plot(data$Index, data$Sales,
     main = "Index vs Sales",
     xlab = "Index", ylab = "Sales",
     pch = 16, col = "sky blue")
```



Part b):

The linear model is fitted to the data and the diagnostic plots are attached below. It can be seen that there is a U shape in the residual vs fitted plot which shows that the model violates the assumption of linearity in this case. The assumption of normality is however met as there are no deviating points from the 45 degree line.

```
par(mfrow=c(2,2))
M1 <- lm(Sales ~ Index,
         data = data)
plot(M1)
```



Part c):

The summary of the order 2 polynomial is attached below. The predictors are all significant. The R square of the model is 95.13% which means that the model is able to explain 95.13% variability. The overall model is statistically significant with p value below alpha = 0.05,

```
# Fit a quadratic model (order 2)
M2 <- lm(Sales ~ poly(Index, 2, raw = TRUE), data = data)
summary(M2)
```

```
##
## Call:
```

```
## lm(formula = Sales ~ poly(Index, 2, raw = TRUE), data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.755 -1.967  0.037  1.749  7.827
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -3.50608    0.50308  -6.969 3.06e-11 ***
## poly(Index, 2, raw = TRUE)1 -4.96591    0.23046 -21.548  < 2e-16 ***
## poly(Index, 2, raw = TRUE)2  0.80875    0.02201  36.744  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.511 on 240 degrees of freedom
## Multiple R-squared:  0.9513, Adjusted R-squared:  0.9509
## F-statistic:  2343 on 2 and 240 DF,  p-value: < 2.2e-16
```

The summary of the order 3 polynomial is attached below. The predictors are all significant. The R square of the model is 95.13% which means that the model is able to explain 95.13% variability. The overall model is statistically significant with p value below alpha = 0.05. The performance of both models is approximately same when it comes to goodness of fit.

```
# Fit a cubic model (order 3)
M3 <- lm(Sales ~ poly(Index, 3, raw = TRUE), data = data)
summary(M3)
```

```
##
## Call:
## lm(formula = Sales ~ poly(Index, 3, raw = TRUE), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7850 -1.9384  0.0545  1.7424  7.8321
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -3.421148   0.668122  -5.121 6.27e-07 ***
## poly(Index, 3, raw = TRUE)1 -5.062632   0.550206  -9.201  < 2e-16 ***
## poly(Index, 3, raw = TRUE)2  0.832770   0.125982   6.610 2.48e-10 ***
## poly(Index, 3, raw = TRUE)3 -0.001599   0.008255  -0.194    0.847
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.516 on 239 degrees of freedom
## Multiple R-squared:  0.9513, Adjusted R-squared:  0.9507
## F-statistic:  1556 on 3 and 239 DF,  p-value: < 2.2e-16
```

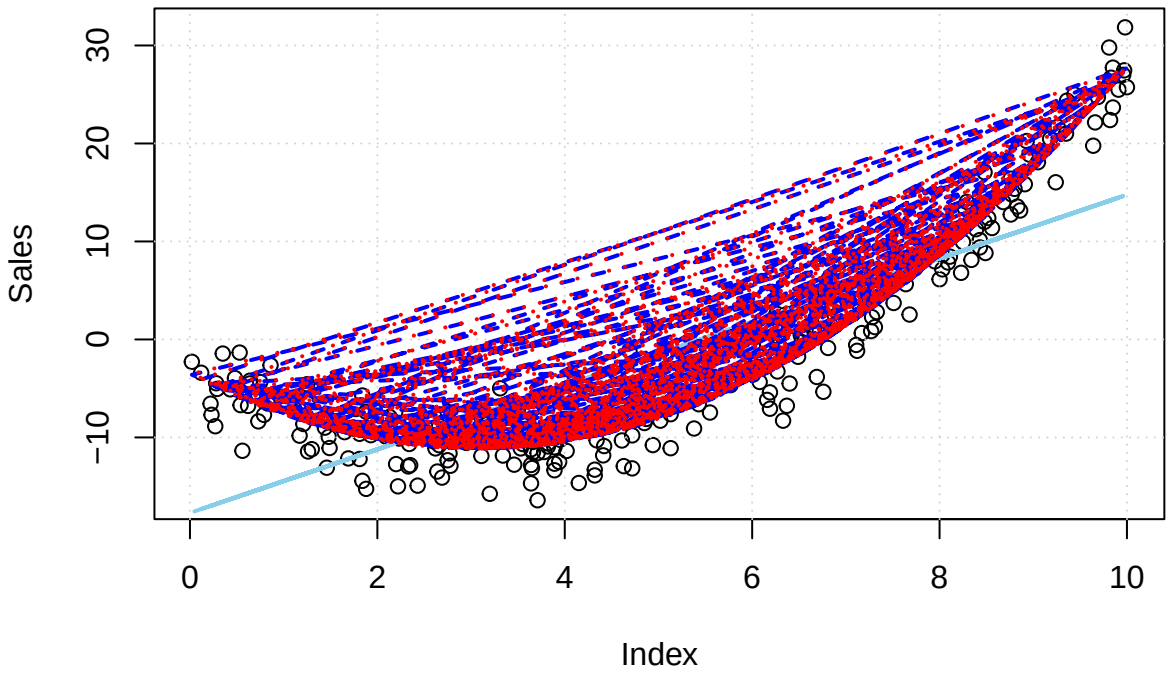
Part d):

The scatter plot is created with three lines one for each of the model. It can be seen that the red line which corresponds to the M3 model better explains the model and covers most of the point and shape. Hence, the better fitting model is M3.

```r
# Plot the data
plot(data$Index, data$Sales, xlab = "Index",
     col = "black", ylab = "Sales",
     ain = "Polynomial Regression Models")
```

```
## Warning in plot.window(...): "ain" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ain" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ain" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ain" is not a
## graphical parameter

## Warning in box(...): "ain" is not a graphical parameter

## Warning in title(...): "ain" is not a graphical parameter
```

```r
grid()

# Add predicted lines from models M2 and M3
lines(data$Index, predict(M1), col = "sky blue", lwd = 2, lty = 2)
lines(data$Index, predict(M2), col = "blue", lwd = 2, lty = 2)  # Quadratic model (M2)
lines(data$Index, predict(M3), col = "red", lwd = 2, lty = 3)   # Cubic model (M3)
```



Part e):

The term "Index" is significant ($p < 0.001$), indicating that the linear relationship between Sales and Index is significant. The quadratic term "poly(Index, 2, raw = TRUE)" is highly significant ($p < 0.001$), suggesting that the quadratic relationship between Sales and Index significantly improves the model fit compared to the linear model (M1). The cubic term "poly(Index, 3, raw = TRUE)" is also highly significant ($p < 0.001$), indicating that adding the cubic term significantly enhances the model fit compared to both the linear (M1) and quadratic (M2) models.

```
anova(M1)
```

```
## Analysis of Variance Table
##
## Response: Sales
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Index       1  21036 21036.0  505.59 < 2.2e-16 ***
## Residuals 241  10027    41.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(M2)
```

```
## Analysis of Variance Table
##
## Response: Sales
##                              Df  Sum Sq Mean Sq F value    Pr(>F)
## poly(Index, 2, raw = TRUE)    2 29549.7 14774.9    2343 < 2.2e-16 ***
## Residuals                   240  1513.4     6.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(M3)
```

```
## Analysis of Variance Table
##
## Response: Sales
##                              Df  Sum Sq Mean Sq F value    Pr(>F)
## poly(Index, 3, raw = TRUE)    3 29550.0  9850.0  1555.8 < 2.2e-16 ***
## Residuals                   239  1513.2     6.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part f):

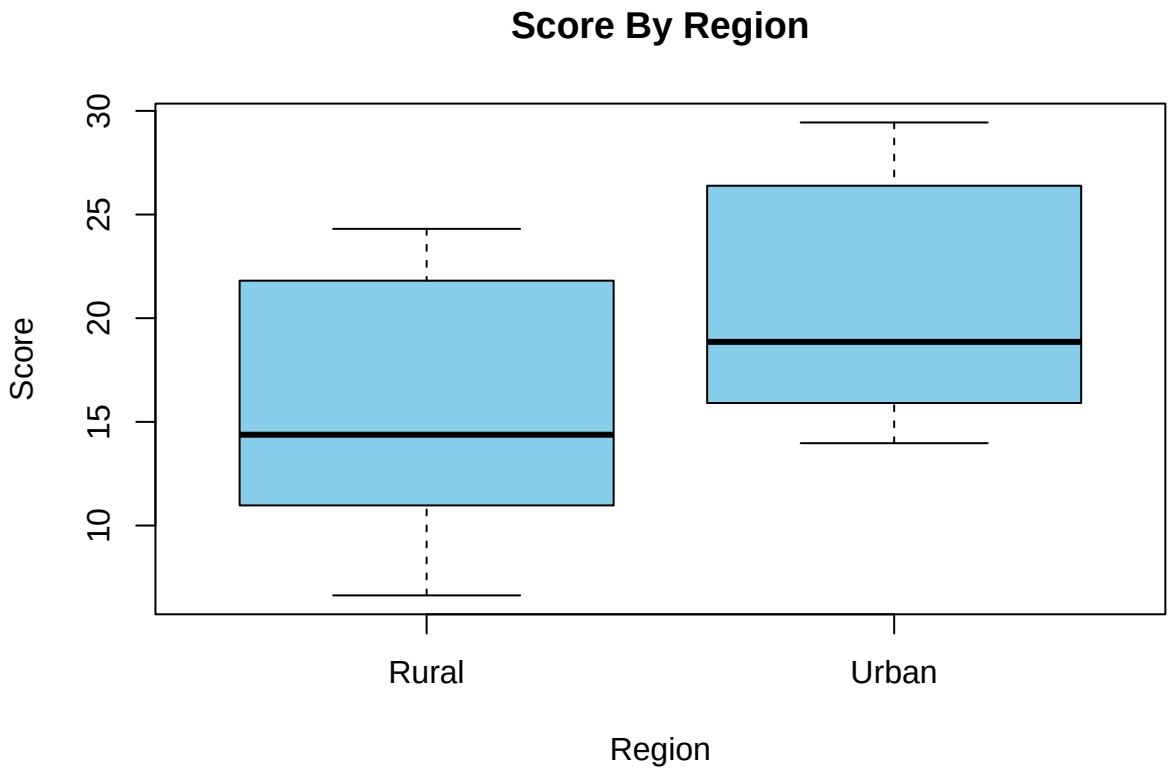
It is observed that each successive term (linear, quadratic, and cubic) significantly contributes to improving the model fit, suggesting that the relationship between Sales and Index is best captured by the cubic model (M3) among the three. Similarly, the R square is also high for the M3 model and the plot shows that the M3 model better explains the trend and shape of the data compared to M1 and M2 models.

Question # 02):

Part a):

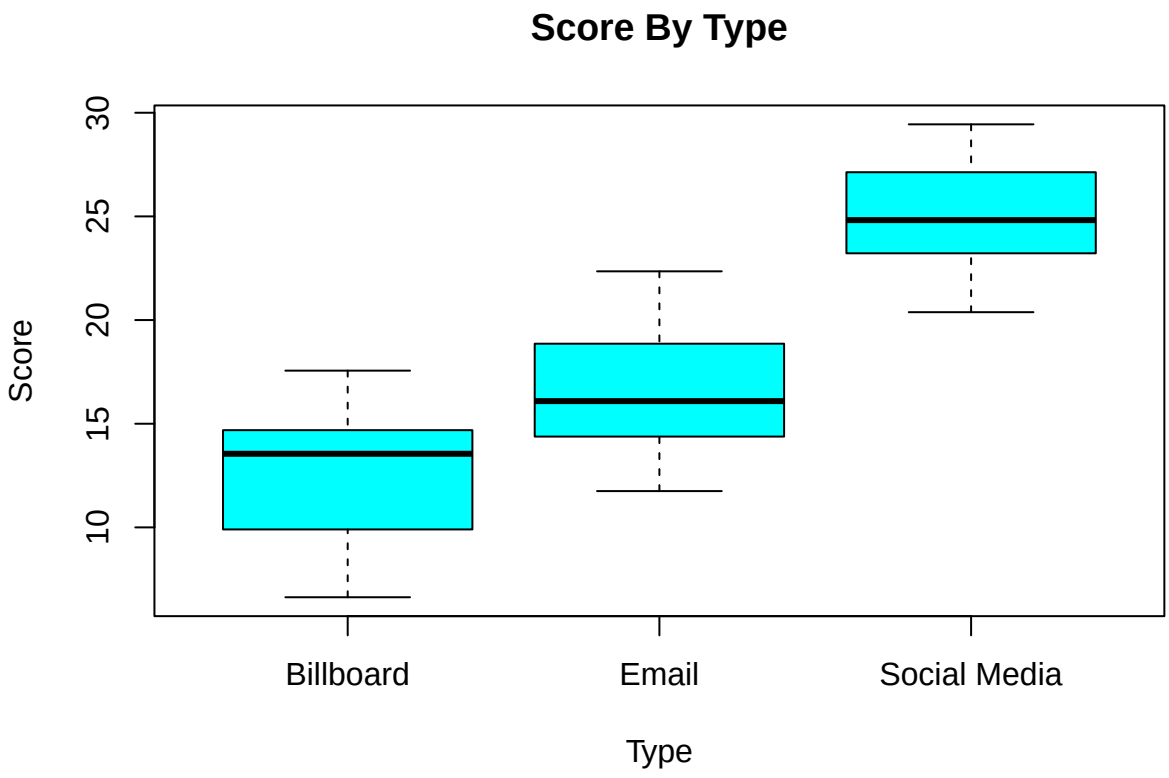
A boxplot is created to show the distribution of increase in engagement score by region and is attached below. It can be seen that the median increase in engagement score is higher in Urban region compared to Rural region.

```
data <- read.csv("campaign.csv")

boxplot(data$Score ~ data$Region,
        main = "Score By Region",
        xlab = "Region", ylab = "Score", col = "sky blue")
```



Similarly, the distribution of percentage increase in engagement score is plotted based on type of marketing campaign used and the highest median percentage increase in engagement score is observed for Social Media, followed by Email and Billboards.

```
boxplot(data$Score ~ data$Type,
        main = "Score By Type",
        xlab = "Type", ylab = "Score", col = "cyan")
```



Part b):

The interaction model in this case is:

Score = B0 + (B1 * Region) + (B2 * Type) + B3 * (Region:Type) + e

- B0 represents the intercept.
- B1 represents the effect of Type (Social Media, Email, Billboard) on Score.
- B2 represents the effect of Region (Urban, Rural) on Score.
- B3 represents the interaction effect between Type and Region on Score.
- e represents the error term.

Part c):

The null and alternate hypothesis are:

- Null Hypothesis: There is no significant difference in the mean Score across different Types and Regions.
- Alternative hypothesis : There is a significant difference in the mean Score across different Types and Regions.

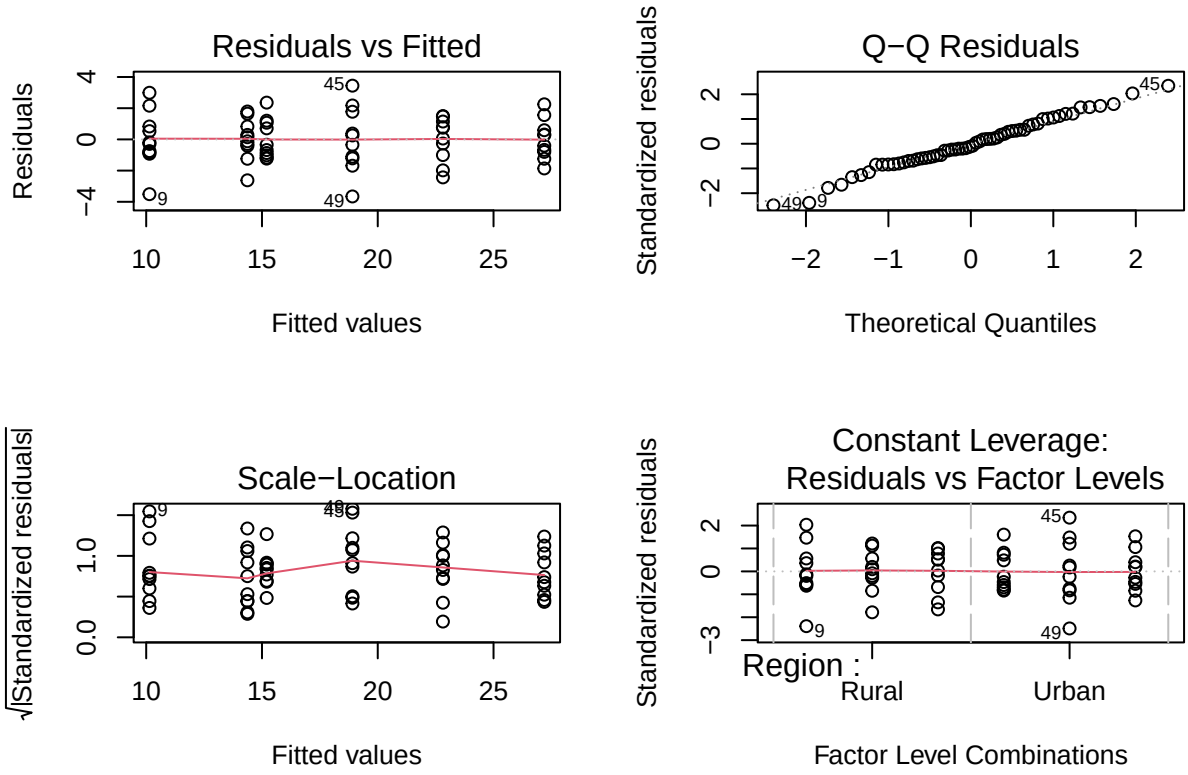
The p-value for Region is less than 0.05, indicating that there is a significant effect of Region on the percentage increase in engagement Score. Similarly, the p-value for Type is also much less than 0.05, indicating that there is a significant effect of Type on the percentage increase in engagement Score. Finally, the interaction term (Region:Type) has a p-value greater than 0.05, indicating that there is no significant interaction effect between Type and Region on the percentage increase in engagement Score. In other words, the impact of Type on Score is not significantly different across different Regions.

```
model <- lm(Score ~ Region + Type + Region * Type,
            data = data)
# ANOVA for interaction model
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Score
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## Region       1  325.45  325.45 135.7281 2.336e-16 ***
## Type         2 1585.09  792.54 330.5242 < 2.2e-16 ***
## Region:Type  2    1.29    0.64   0.2683    0.7657
## Residuals   54  129.48    2.40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The assumptions of the model are validated using the diagnostic plots attached below and it can be seen that the assumption of linearity, normality are met and there are no issues of heteroskedasticity.

```
par(mfrow=c(2,2))
plot(model)
```



Part d):

For main effect of type of campaigns used, the null and alternate hypothesis are attached below:

- Null Hypothesis: There is no significant difference in the mean Score across different Types of campaigns used.
- Alternate Hypothesis: There is a significant difference in the mean Score across different Types of campaigns used.

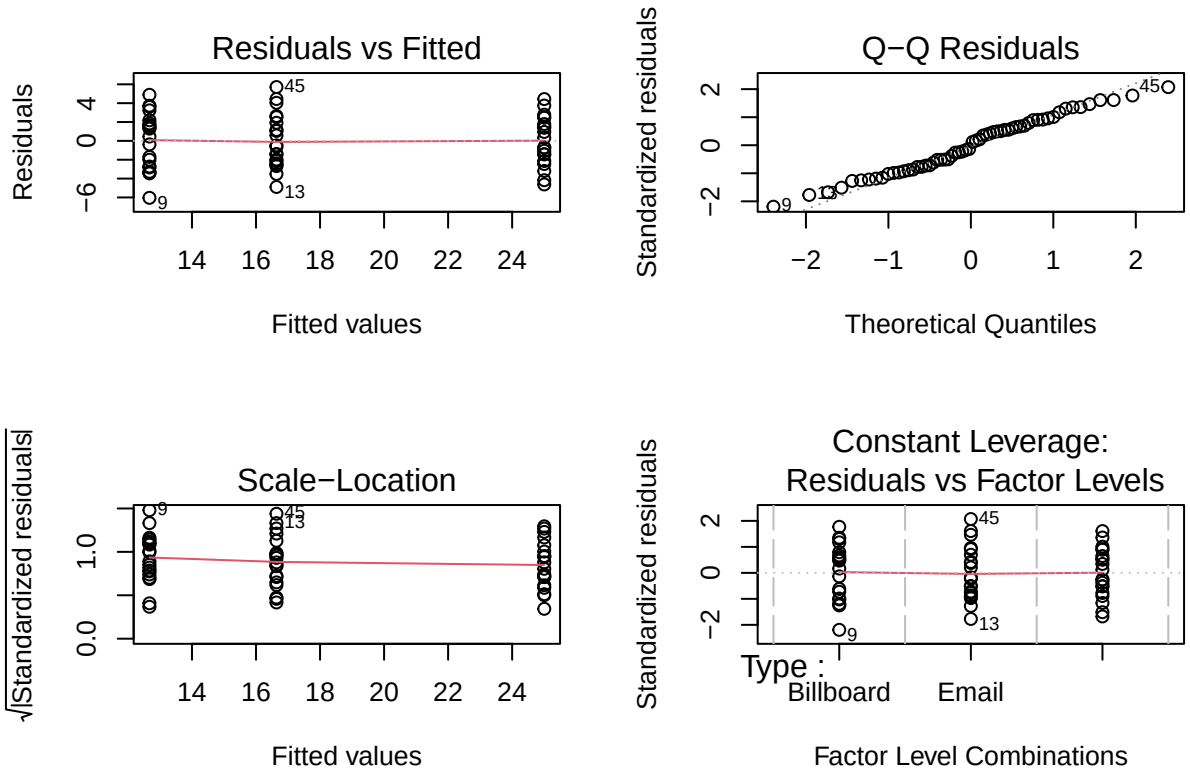
The p-value for Type is less than 0.05, indicating that there is a significant effect of Type of campaigns used on the percentage increase in engagement Score.

```
# ANOVA for Type
type_model <- lm(Score ~ Type, data = data)
anova(type_model)
```

```
## Analysis of Variance Table
##
## Response: Score
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Type       2 1585.09  792.54  99.019 < 2.2e-16 ***
## Residuals 57  456.22    8.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The assumptions of the model are validated using the diagnostic plots attached below and it can be seen that the assumption of linearity, normality are met and there are no issues of heteroskedasticity.

```
par(mfrow=c(2,2))
plot(type_model)
```



For main effect of region, the null and alternate hypothesis are attached below:

- Null Hypothesis: There is no significant difference in the mean Score across different regions.
- Alternate Hypothesis: There is a significant difference in the mean Score across different regions.

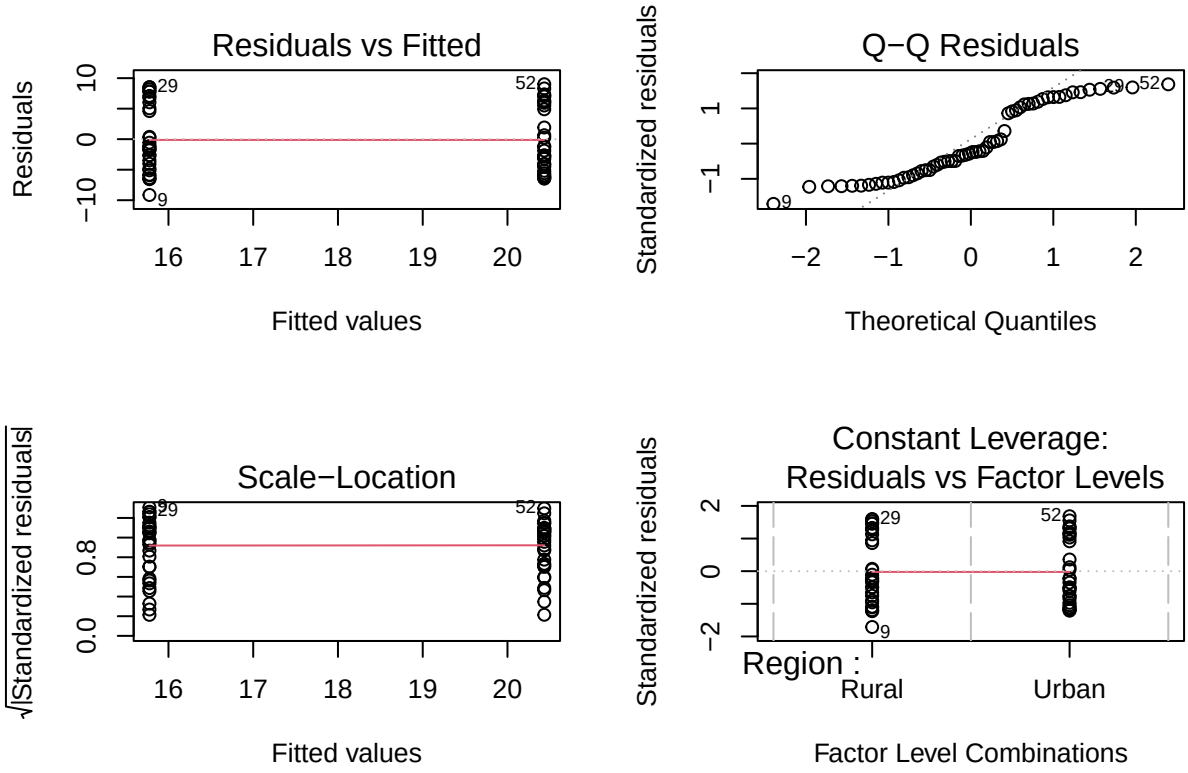
The p-value for Region is less than 0.05, indicating that there is a significant effect of Region on the percentage increase in engagement Score.

```
# ANOVA for Region
region_model <- lm(Score ~ Region, data = data)
anova(region_model)
```

```
## Analysis of Variance Table
##
## Response: Score
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## Region     1  325.45  325.45  11.001 0.001575 **
## Residuals 58 1715.86   29.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The assumptions of the model are validated using the diagnostic plots attached below and it can be seen that the assumption of linearity is met but the assumption of normality is violated as deviation of points is observed and there are no issues of heteroskedasticity.

```
par(mfrow=c(2,2))
plot(region_model)
```



Part e):

The design balance is checked and it is pretty balanced and hence we can move forward to performing pair wise comparison.

```
table(data$Region, data$Type)
```

```
##
##         Billboard Email Social Media
##   Rural        10    10           10
##   Urban        10    10           10
```

The mean percentage increase in engagement Score for Email campaigns is significantly different from Billboard campaigns, with Email campaigns having a higher mean Score by approximately 3.97%. The mean percentage increase in engagement Score for Social Media campaigns is significantly different from Billboard campaigns, with Social Media campaigns having a higher mean Score by approximately 12.33%. Finally,

the mean percentage increase in engagement Score for Social Media campaigns is significantly different from Email campaigns, with Social Media campaigns having a higher mean Score by approximately 8.36%.

The mean percentage increase in engagement Score for Urban regions is significantly different from Rural regions, with Urban regions having a higher mean Score by approximately 4.66%. This suggests that marketing campaigns deployed in Urban regions tend to have a higher impact on customer engagement scores compared to those deployed in Rural regions.

```
# TukeyHSD for Type
tukey_type <- TukeyHSD(aov(Score ~ Type, data = data))
(tukey_type)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Score ~ Type, data = data)
##
## $Type
##                             diff       lwr       upr     p adj
## Email-Billboard         3.9675  1.814604  6.120396 0.0001247
## Social Media-Billboard 12.3315 10.178604 14.484396 0.0000000
## Social Media-Email      8.3640  6.211104 10.516896 0.0000000
```

```
# TukeyHSD for Region
tukey_region <- TukeyHSD(aov(Score ~ Region, data = data))
(tukey_region)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Score ~ Region, data = data)
##
## $Region
##              diff     lwr     upr     p adj
## Urban-Rural 4.658 1.84685 7.46915 0.0015755
```