Quiz - Ahmad Hathim bin Ahmad Azman (P153146)

Load Data and statistical summary

```
campaign_data <- read.table(file = "E:/MSc DSc/Sem 1/Business Analytics/Data Part B.csv",sep = ",", hea
head(campaign_data, 5)
```

```
##   campaign_id price units_sold advertising_budget website_visits
## 1           1    25        150               3000          12000
## 2           2    30        200               3500          15000
## 3           3    28        180               3200          13000
## 4           4    26        170               3100          12500
## 5           5    29        160               3000          13500
```

```
str(campaign_data)
```

```
## 'data.frame':    30 obs. of  5 variables:
##  $ campaign_id       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ price             : int  25 30 28 26 29 32 31 35 40 45 ...
##  $ units_sold        : int  150 200 180 170 160 210 175 220 250 230 ...
##  $ advertising_budget: int  3000 3500 3200 3100 3000 4000 3700 4200 5000 4700 ...
##  $ website_visits    : int  12000 15000 13000 12500 13500 16000 14000 17000 20000 19000 ...
```

```
summary(campaign_data)
```

```
##   campaign_id        price         units_sold    advertising_budget
##  Min.   : 1.00   Min.   :25.00   Min.   :150.0   Min.   :3000
##  1st Qu.: 8.25   1st Qu.:33.25   1st Qu.:211.2   1st Qu.:4025
##  Median :15.50   Median :40.50   Median :247.5   Median :4900
##  Mean   :15.50   Mean   :40.43   Mean   :243.5   Mean   :4760
##  3rd Qu.:22.75   3rd Qu.:47.75   3rd Qu.:278.8   3rd Qu.:5575
##  Max.   :30.00   Max.   :55.00   Max.   :330.0   Max.   :6300
##  website_visits
##  Min.   :12000
##  1st Qu.:16125
##  Median :19750
##  Mean   :19450
##  3rd Qu.:22875
##  Max.   :26500
```

Question (A) Linear Regression Model

```r
model <- lm(website_visits ~ price + units_sold + advertising_budget, data = campaign_data)
model
```

```
##
## Call:
## lm(formula = website_visits ~ price + units_sold + advertising_budget,
##     data = campaign_data)
##
## Coefficients:
##        (Intercept)               price          units_sold  advertising_budget
##          -1199.118             139.681              34.025               1.411
```

The coefficients of the model are as follows:

- *price*: 139.681
- $units\_sold$: 34.025
- $advertising\_budget$: 1.411
- *y-intercept* (B_0): -1199.118

$$website\_visits = -1199.118 + 139.681(price) + 34.025(units\_sold) + 1.411(advertising\_budget)$$

This shows that if the price of the product increases by 1 unit, the website visits will increase by 139.681 visits. Similarly, if the units sold increase by 1 unit, the website visits will increase by 34.025 visits. If the advertising budget increases by 1 unit, the website visits will increase by 1.411 visits. Otherwise, if the variables are held constant, the company will lose website visits by -1199.118 visits.

Question (B) Is the model considered a good fit? Justify the answer.

```r
summary(model)
```

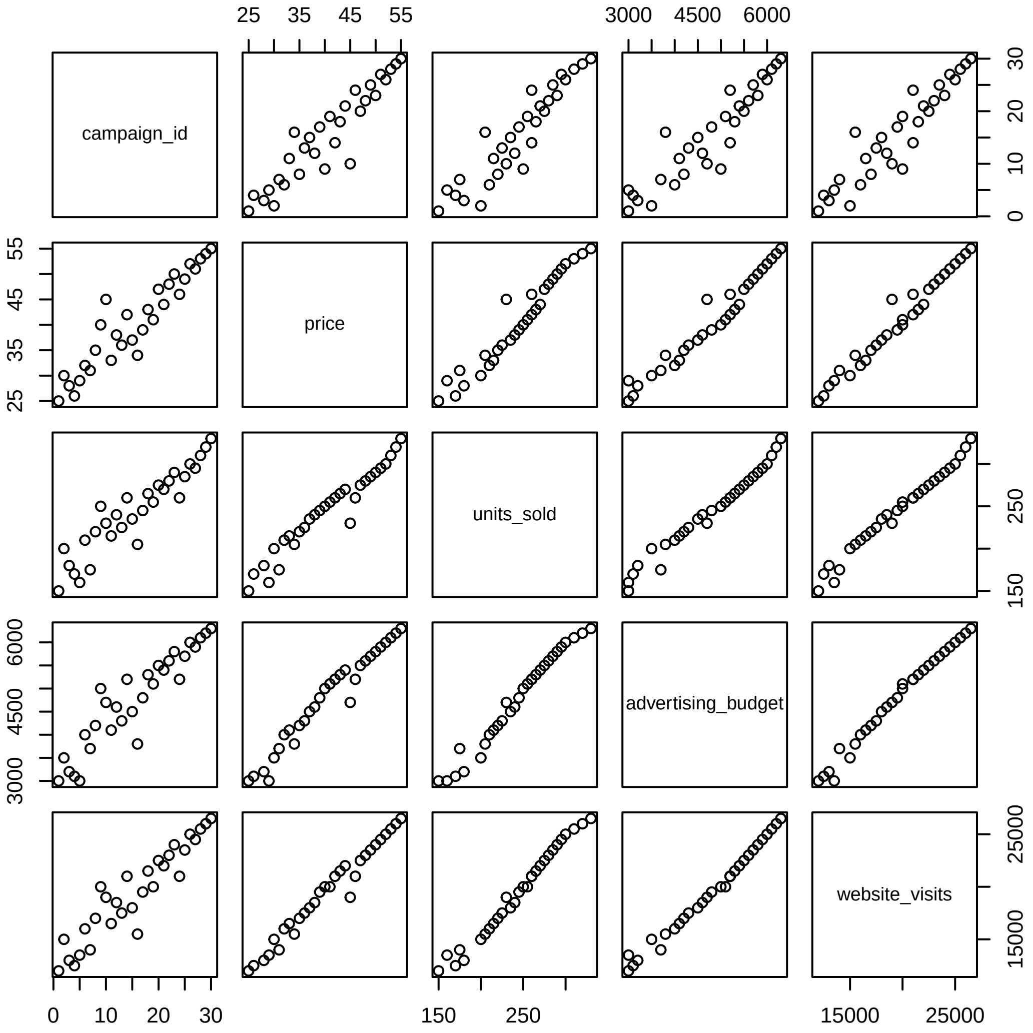
```
##
## Call:
## lm(formula = website_visits ~ price + units_sold + advertising_budget,
##     data = campaign_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -543.88 -224.28   18.56  160.34  971.43
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1199.1183   330.2159  -3.631 0.001213 **
## price                139.6814    33.5722   4.161 0.000307 ***
## units_sold            34.0254     8.3439   4.078 0.000381 ***
## advertising_budget     1.4110     0.4741   2.976 0.006239 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 320.5 on 26 degrees of freedom
## Multiple R-squared:  0.9951, Adjusted R-squared:  0.9945
## F-statistic:  1744 on 3 and 26 DF,  p-value: < 2.2e-16
```

Looking at the summary of the model, the p-value of the multiple linear regression is significant at <2.2e-16, and each variable itself has a significant p-value. Thus all variables; *price*, *units_sold*, and *advertising_budget* are significant in predicting website visits. The adjusted R-squared is also 0.9945, meaning that the model explains 99.45% of the variance in website visits by its independent variables; *price*, *units_sold*, and *advertising_budget*.

Question (C), Check the assumptions of Linear Regression

1) Check for Linearity

```
pairs(campaign_data)
```



From the pair plots above, it can be seen that the relationship between website visits and price, units sold, and advertising budget is linear.

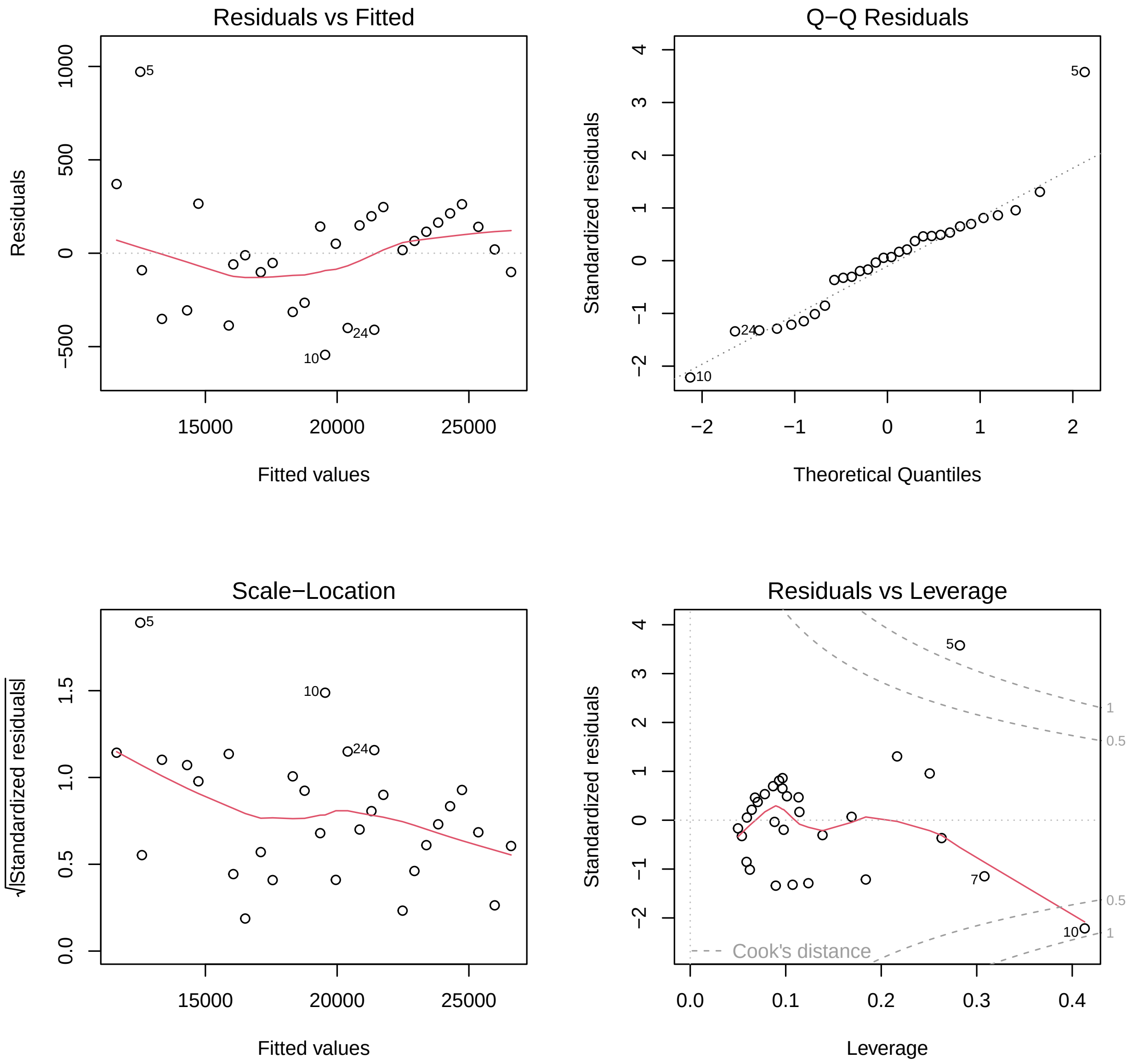
2) Check for Independence

It is often difficult to check for independence between variables. It if often understood during the process of data collection and identifying variables. Thus, it will be on the assumption that the data is independent.

3) Check for Normality and Equal Variance

To test for Normality and Equal Variance, there are 3 plots that can be used; histogram, qq-plot, and residuals vs fitted plot.

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



```r
par(mfrow=c(1,1))
```

From the 3 plots above it can be seen that;

1. The **QQ-Plot** shows that the residuals are relatively normally distributed
2. The **residuals vs fitted plot** shows that the residuals are randomly scattered around the 0 line, but there is a slight irregularity, which indicates that the variance of the residuals is not constant.
3. The **Residuals vs Leverage** plot shows that there is presence of outlier in the data that may affect the model.

Question (D) Does any of the variables require transformation? Justify the answer.

The variables that require transformation for this model is the *websites_visits*. This is because the equal variance does not appear homogenous, thus to adjust for equal variance, the *website_visits* variable should be transformed.

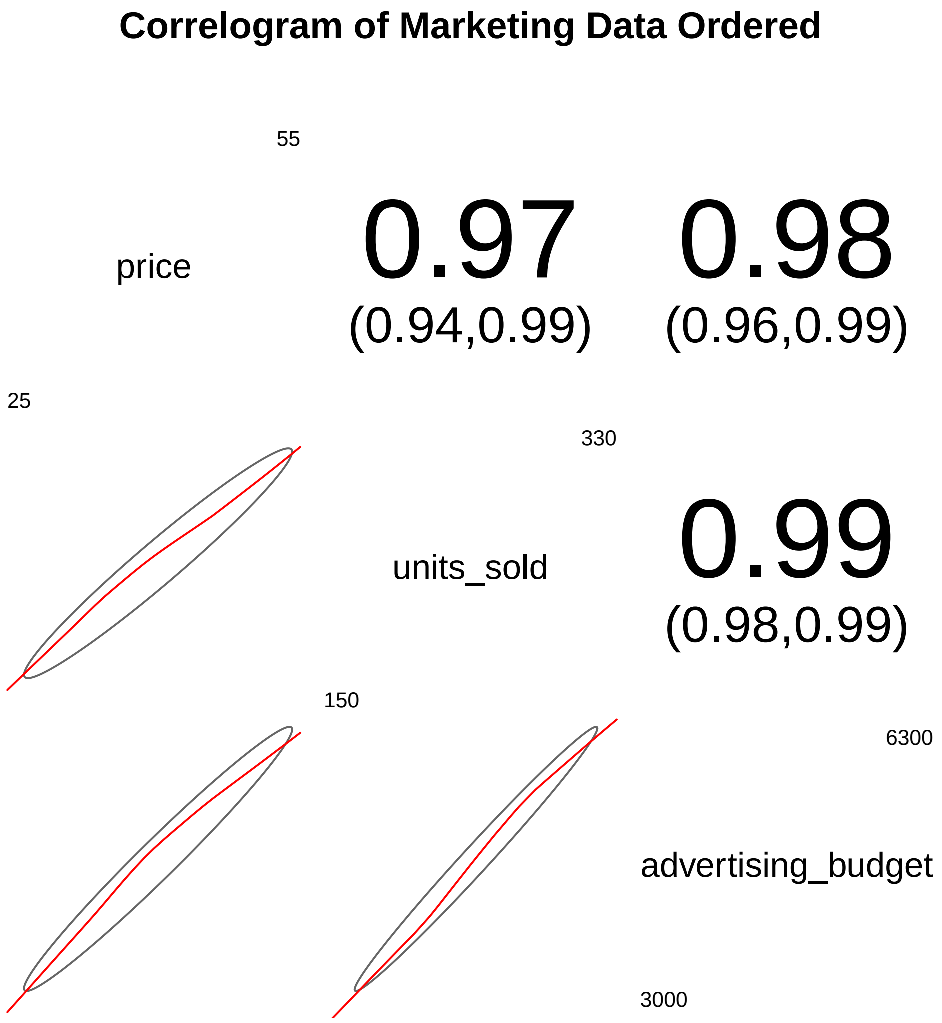
Question (E), perform multicollienarity test

```
corr.test(campaign_data[2:4])
```

```
## Call:corr.test(x = campaign_data[2:4])
## Correlation matrix
##                    price units_sold advertising_budget
## price               1.00       0.97               0.98
## units_sold          0.97       1.00               0.99
## advertising_budget  0.98       0.99               1.00
## Sample Size
## [1] 30
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                    price units_sold advertising_budget
## price                  0          0                  0
## units_sold             0          0                  0
## advertising_budget     0          0                  0
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

```
corrgram(campaign_data[2:4],
         main = 'Correlogram of Marketing Data Ordered',
         order=FALSE,
         lower.panel=panel.ellipse,
         upper.panel=panel.conf,
         text.panel=panel.txt,
         diag.panel = panel.minmax)
```

```
## Warning in par(usr): argument 1 does not name a graphical parameter
## Warning in par(usr): argument 1 does not name a graphical parameter
## Warning in par(usr): argument 1 does not name a graphical parameter
```



From the correlation matrix above, it can be seen that the correlation between the independent variables is high, thus there is evidence of multicollinearity in the model. Thus, to reduce multicollinearity, the variables should be transformed. There are multiple ways to transform data to avoid multicollinearity, such as using dimension reduction such as Principle Component Analysis (PCA), or to remove one of highly correlated variables. Another method of reducing multicollinearity is to use transformation such as log transformation of the independent variables.