Assignment 2 (Solution)

In order to answer problems in Assignment 2, you need to use the ‘Carseat’ data, which is part of the ‘ISLR’ library. The goal of this assignment is to predict ‘Sales (child car seat sales)’ in 400 locations based on a number of predictors.

1. Which of the predictors are quantitative, and which are qualitative?
   Hint: str() or summary ()

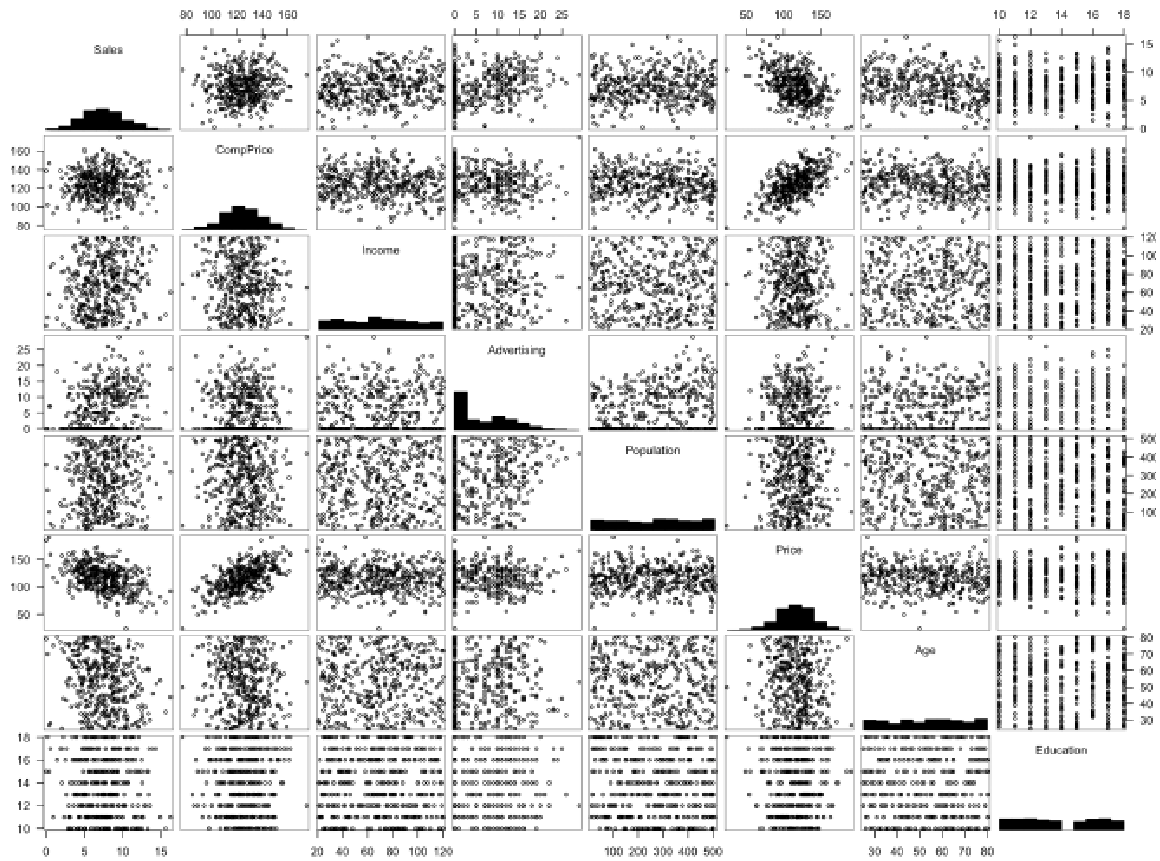
```
> summary(Carseats)
     Sales          CompPrice       Income        Advertising       Population        Price      
 Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000   Min.   : 10.0   Min.   : 24.0  
 1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000   1st Qu.:139.0   1st Qu.:100.0  
 Median : 7.490   Median :125   Median : 69.00   Median : 5.000   Median :272.0   Median :117.0  
 Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635   Mean   :264.8   Mean   :115.8  
 3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000   3rd Qu.:398.5   3rd Qu.:131.0  
 Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000   Max.   :509.0   Max.   :191.0  
  ShelveLoc        Age          Education     Urban       US     
 Bad   : 96   Min.   :25.00   Min.   :10.0   No :118   No :142  
 Good  : 85   1st Qu.:39.75   1st Qu.:12.0   Yes:282   Yes:258  
 Medium:219   Median :54.50   Median :14.0                      
              Mean   :53.32   Mean   :13.9                      
              3rd Qu.:66.00   3rd Qu.:16.0                      
              Max.   :80.00   Max.   :18.0                      
```

   Quantitative: Sales, CompPrice, Income, Advertising, Population, Price, Age, Education
   Qualitative: ShelveLoc (3-levels), Urban (2-levels), and US (2-levels)

2. Using Quantitative variables, describe the distributions in terms of shape, symmetry, and potential outlier. Do you think it is required to transform some variable(s)? If so, transform the variable(s) and justify your answer (Since the Advertising includes 1 missing value, please delete the Advertising variable when you compute correlation).
   Hint: gpair(), cor()

   Sort the quantitative variables, then the new data set, Carseats1, contains 8 quantitative variables.

```
     Sales          CompPrice       Income        Advertising       Population        Price      
 Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000   Min.   : 10.0   Min.   : 24.0  
 1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000   1st Qu.:139.0   1st Qu.:100.0  
 Median : 7.490   Median :125   Median : 69.00   Median : 5.000   Median :272.0   Median :117.0  
 Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635   Mean   :264.8   Mean   :115.8  
 3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000   3rd Qu.:398.5   3rd Qu.:131.0  
 Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000   Max.   :509.0   Max.   :191.0  
      Age          Education   
 Min.   :25.00   Min.   :10.0  
 1st Qu.:39.75   1st Qu.:12.0  
 Median :54.50   Median :14.0  
 Mean   :53.32   Mean   :13.9  
 3rd Qu.:66.00   3rd Qu.:16.0  
 Max.   :80.00   Max.   :18.0  
```



Based on the scatterplot, Advertising is highly skewed to the right. For the skewed distribution, the log-transformation can be considered to achive the more accurate results. After dropping Advertisement, the correlation coefficients are computed.

```
> cor(Carseats1[,-4])
                Sales   CompPrice       Income   Population       Price          Age    Education
Sales       1.00000000  0.06407873  0.151950979  0.050470984 -0.44495073 -0.231815440 -0.051955242
CompPrice   0.06407873  1.00000000 -0.080653423 -0.094706516  0.58484777 -0.100238817  0.025197050
Income      0.15195098 -0.08065342  1.000000000 -0.007876994 -0.05669820 -0.004670094 -0.056855422
Population  0.05047098 -0.09470652 -0.007876994  1.000000000 -0.01214362 -0.042663355 -0.106378231
Price      -0.44495073  0.58484777 -0.056698202 -0.012143620  1.00000000 -0.102176839  0.011746599
Age        -0.23181544 -0.10023882 -0.004670094 -0.042663355 -0.10217684  1.000000000  0.006488032
Education  -0.05195524  0.02519705 -0.056855422 -0.106378231  0.01174660  0.006488032  1.000000000
```

	Distribution			Association with Price	Transformation
	# of peaks	Symmetry	Outlier		
Sales	1	Yes	No	Moderate negative	No
CompPrice	1	Yes	No	Moderate positive	No
Income	2	No	No	Moderate negative	No (consider 2 groups)
Population	0 (Uniform)	No	No	Weak negative	No (consider 2 groups)
Price	1	Yes	No	Perfectly linear	No
Age	0 (Uniform)	No	No	Weak negative	No
Education	2	No	No	Weak positive	No (consider 2 groups)

Output of the 4 separate regression models for question 3-5.

Model #	Predictor	Coefficient ($\hat{\beta}_1$)	P-value	SE	R^2	F-Stat
1	Income	0.0153	0.0023	0.0153	0.023	9.401***
2	Population	0.0010	0.314	2.824	0.003	1.016
3	Price	-0.0531	<0.000	2.532	0.198	98.25***
4	US (1=Yes;0=No)	1.0439	0.0004	2.783	0.031	12.89***

3. Fit four separate simple regression models to predict ‘Sales’ using ‘Income’, ‘Population’ and ‘Price’ and US. Then, Write out the estimated model in equation form.
   Hint: lm ()

Model #	Predictor	Coefficient ($\hat{\beta}_1$)	~~P-value~~
1	Income	0.0153	For every additional income (in unit), the average sales increases by 0.0153 (in unit).
2	Population	0.0010	For every additional population (in unit), the average sales increases by 0.0010 (in unit).
3	Price	-0.0531	For every additional price(in unit), the average sales decreases by 0.0531 (in unit).
4	US (1=Yes;0=No)	1.0439	The average difference in sales (in unit) between US and non-US is 1.0439. The average sales in US is 1.0439 units more than the average sales in outside of US.

4. Provide an interpretation of each coefficient in the model. Be careful-some of the variables in the model are qualitative!

   See the table above.

5. For which of the predictors can you reject the null hypothesis $H_0: \beta_j = 0$?

4 sets of hypothesis can be tested based on the p-value. Null hypothesis is rejected if the p-value is less than 0.05.

Model #	Predictor	Hypotheses	P-value	Decision
1	Income	$H_0: \beta_j = 0$ $H_a: \beta_j \neq 0$	0.0023	Reject H_0.
2	Population		0.314	Fail to reject H_0.
3	Price		<0.000	Reject H_0.
4	US (1=Yes;0=No)		0.0004	Reject H_0.

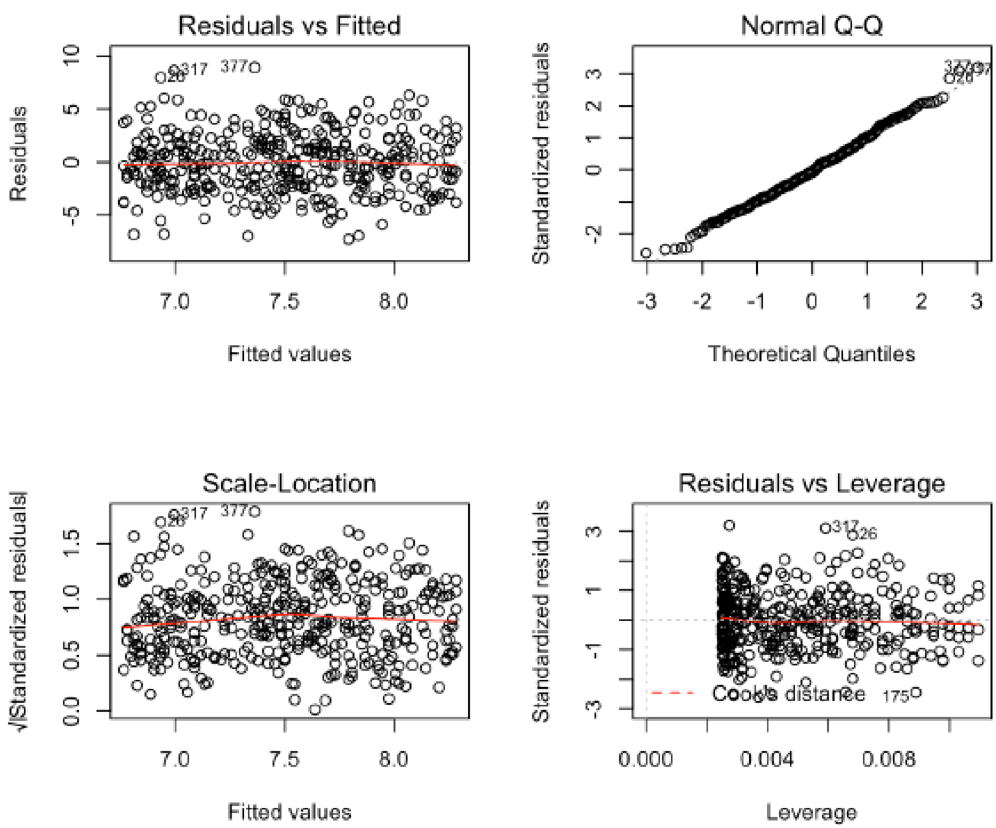
6. Using the models Question 3, obtain 95% confidence intervals for the coefficient(s). Using the confidence intervals, test the null hypothesis $H_0: \beta_j = 0$.
   Hint: confint()

   4 sets of hypothesis can be tested based on the confidence interval. Null hypothesis is rejected if the hypothesized value (=0) is outside of the confidence interval.

Model #	Predictor	Hypotheses	95% CI	Decision
1	Income	$H_0: \beta_j = 0$	(0.0055, 0.02516)	Reject H_0.
2	Population	$H_a: \beta_j \neq 0$	(-0.0009, 0.00285)	Fail to reject H_0.
3	Price		(-0.0636, -0.0426)	Reject H_0.
4	US (1=Yes;0=No)		(0.47219,1.61556)	Reject H_0.

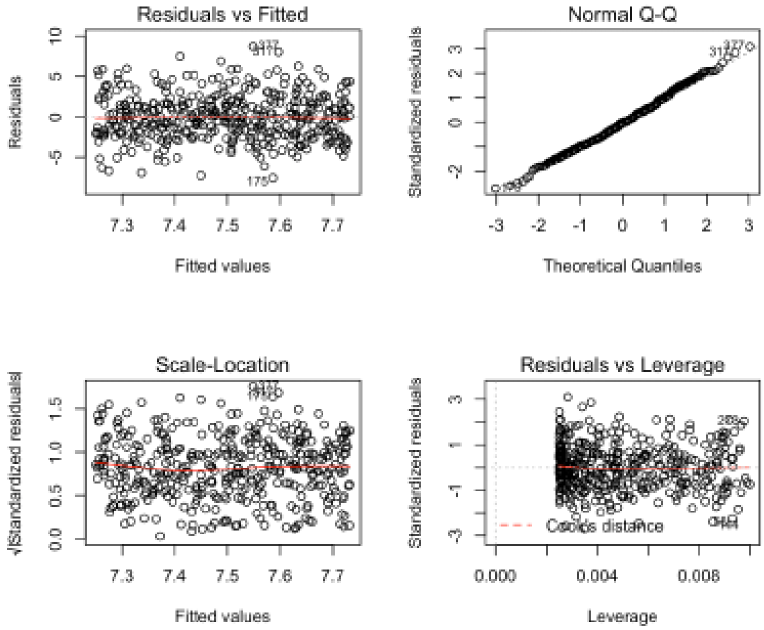
7. Check the assumptions of the models using plot(). Is there evidence of outliers or high leverage observations in the models? If so, please inspect the outliers.
   Hint: plot()

   Model 1:



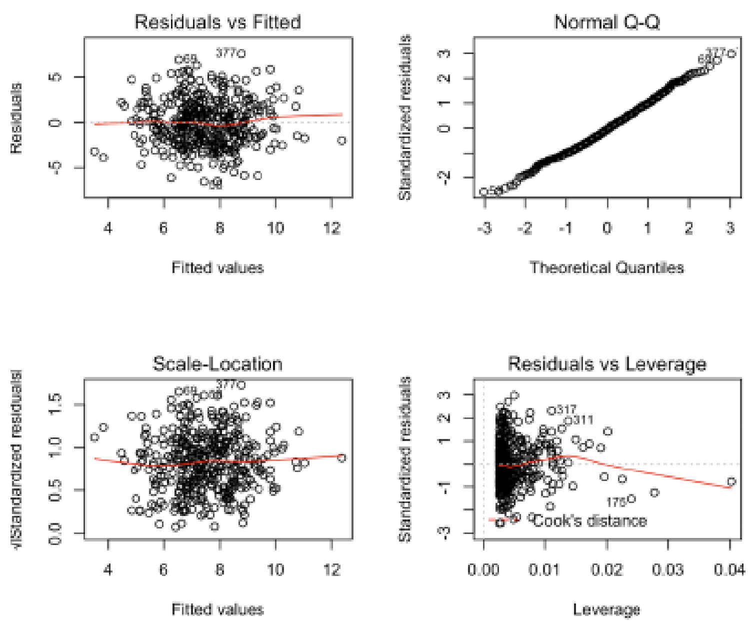
Linearity assumption is met since the normal Q-Q plot shows linear pattern.
Normality assumption is met since there is no pattern on the residual plot.
Independece assumption is met since there is no pattern on the residual plot.
Constant variance assumption is met since there is no pattern on the residual plot.

Model 2:



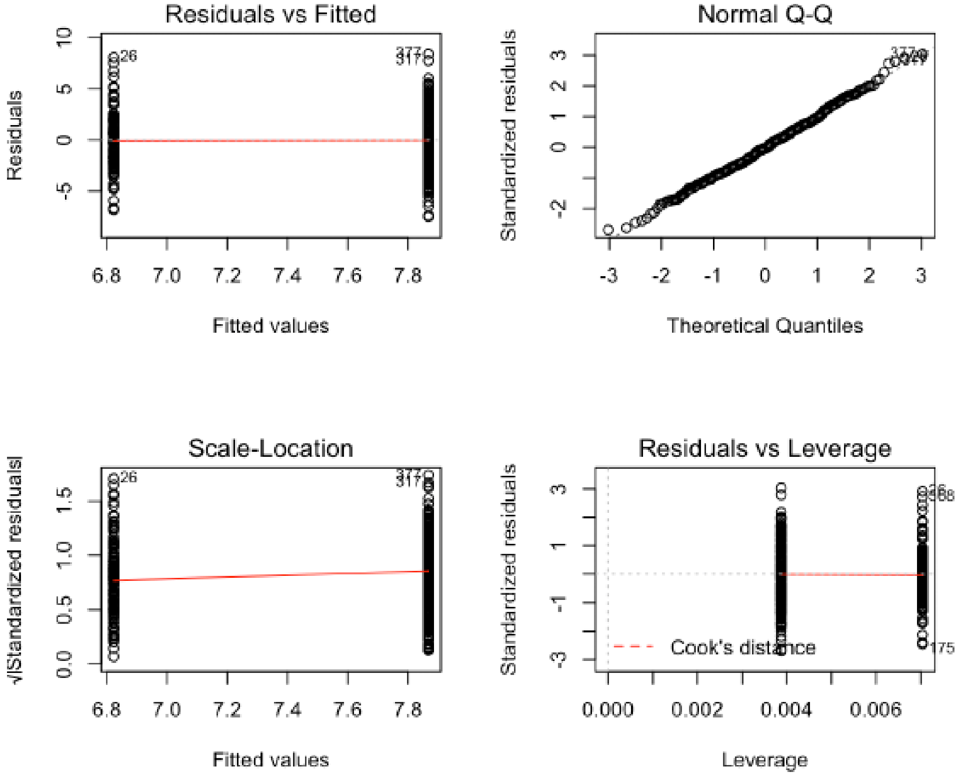
Linearity assumption is met since the normal Q-Q plot shows linear pattern.
Normality assumption is met since there is no pattern on the residual plot.
Independece assumption is met since there is no pattern on the residual plot.
Constant variance assumption is met since there is no pattern on the residual plot.

Model 3:



Linearity assumption is met since the normal Q-Q plot shows linear pattern.
Normality assumption is met since there is no pattern on the residual plot.
Independece assumption is met since there is no pattern on the residual plot.
Constant variance assumption is met since there is no pattern on the residual plot.

Model 4:



Linearity assumption is met since the normal Q-Q plot shows linear pattern.
It is hard to test other 3 assumptions such as Normality assumption, Independecee assumption, and Constant variance assumption because the predictor in model 4 is a categorical variable and there are two lines on the residual plot.