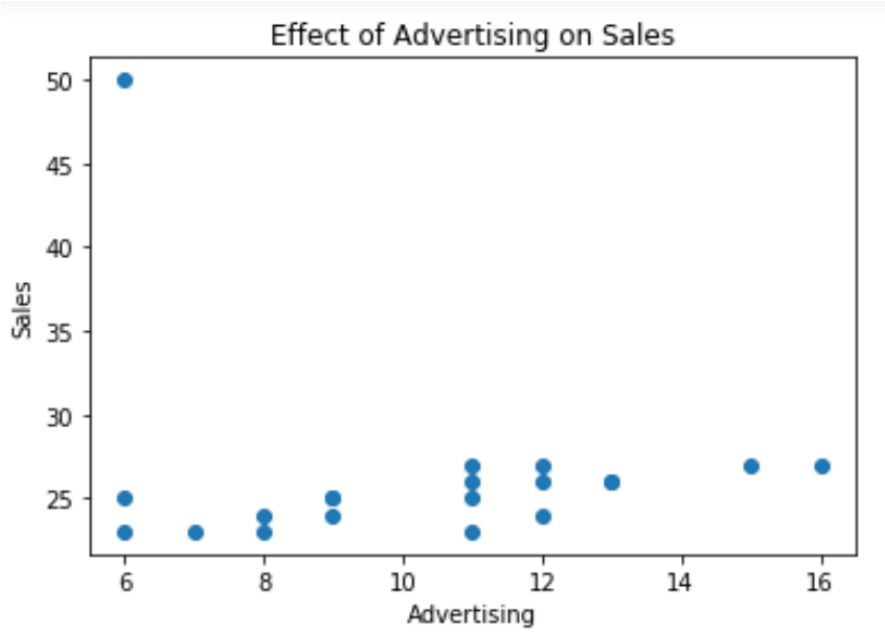
Test Exercise

a)



In this case, I would say that the regression line might not be the best estimator for the trend of the effect of advertising on sales due to the outlier. This could even make the sign of the estimator of beta change.

b)

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.027
Model:	OLS	**Adj. R-squared:**	-0.027
Method:	Least Squares	**F-statistic:**	0.5002
Date:	Sun, 24 May 2020	**Prob (F-statistic):**	0.488
Time:	14:30:18	**Log-Likelihood:**	-62.608
No. Observations:	20	**AIC:**	129.2
Df Residuals:	18	**BIC:**	131.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	29.6269	4.882	6.069	0.000	19.371	39.883
x1	-0.3246	0.459	-0.707	0.488	-1.289	0.640

Omnibus:	40.109	**Durbin-Watson:**	1.994
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	121.776
Skew:	3.178	**Prob(JB):**	3.60e-27
Kurtosis:	13.283	**Cond. No.**	40.1

a = 29.629

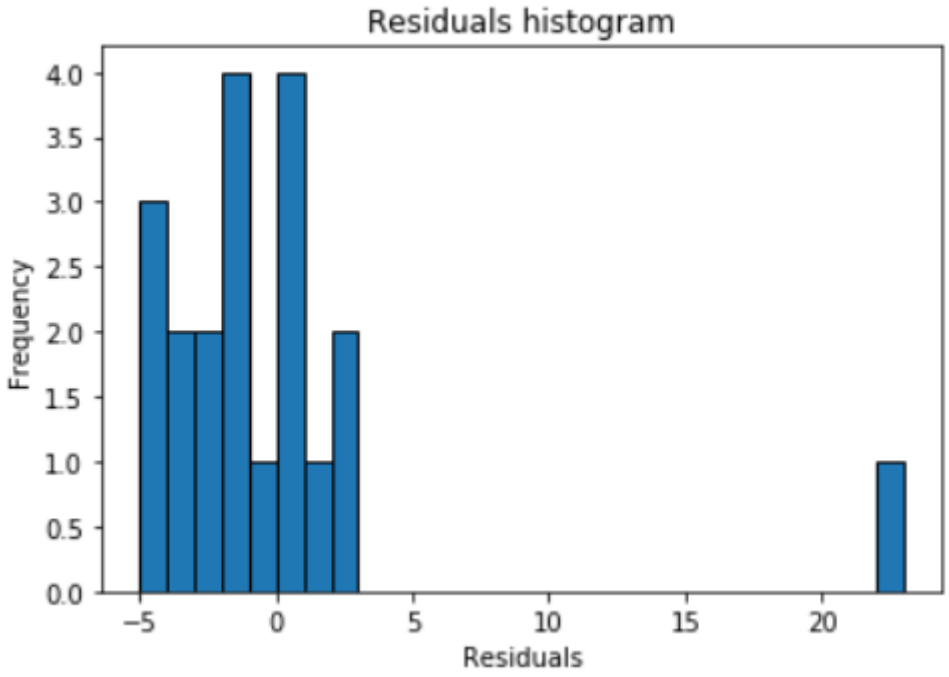
b = -0.3246

Confidence interval for b with 95% of confidence:

[-0.3246-(2.1009*0.459), -0.3246+(2.1009*0.459)] = [-1.2889,0.6397]

As 0 is inside the confidence interval, b is not significantly different from 0 with 95% of confidence.

c)



The shape of the histogram doesn´t look like a bell mostly because of the outlier, meaning the distribution is not similar to the normal distribution. Apart from this, skewness is 3.178, which is much higher than 0, the standard value for normal distribution and the kurtosis is 13.283 and is also far away from the expected value for normal distribution, 3. Therefore, the distribution of the sample violates the assumption that the errors have a similar distribution to the normal distribution.

d) The R-squared is really low, just 2.7% of the errors of the variations of the sales are explained by the model. I would say that the best way to fix this problem is to remove the outlier as we can notice that the distribution of the rest of the data can adjust to the conditions required by the simple regression model more easily.

e)

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.515
Model:	OLS	**Adj. R-squared:**	0.487
Method:	Least Squares	**F-statistic:**	18.08
Date:	Mon, 25 May 2020	**Prob (F-statistic):**	0.000538
Time:	18:50:53	**Log-Likelihood:**	-26.897
No. Observations:	19	**AIC:**	57.79
Df Residuals:	17	**BIC:**	59.68
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	21.1250	0.955	22.124	0.000	19.110	23.140
x1	0.3750	0.088	4.252	0.001	0.189	0.561

Omnibus:	0.597	**Durbin-Watson:**	1.749
Prob(Omnibus):	0.742	**Jarque-Bera (JB):**	0.204
Skew:	-0.252	**Prob(JB):**	0.903
Kurtosis:	2.933	**Cond. No.**	43.1

a = 21.1250

b = 0.3750

Confidence interval for b with 95% of confidence:

[0.375- (2.1098*0.088), 0.375+(2.1098*0.088)] = [0.189,0.561]

As 0 is not inside the confidence interval, b is significantly different from 0 with 95% of confidence.

f) Comparing both results, we can conclude that the second model worked much better since the R-squared is higher, 51.5% compared to 2.7%, what means that more than half of the variations of sales are explained by the model while in the first model, a minimal amount of the changes of sales are captured by the model. Moreover, the figures of skewness and kurtosis of the second model are closer to those that are expected for a normal distribution, so the second model doesn´t violate the assumption that the errors are distributed similarly to the normal distribution. I learnt from this exercise that it is very important to select the right data to run regressions because including incorrect data can lead to misinterpretations.