EXERCISE 1:
An online retailer wants to determine whether there is a relationship between price and number of tool sets sold. She tests eleven different prices (11 observations).

Price (X)	Number of tool sets sold
$10.00	1000
$12.00	900
$14.00	800
$16.00	780
$18.00	650
$20.00	600
$25.00	400
$30.00	200
$50.00	100
$60.00	80
$100.00	75

Here is the MS Excel output:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.816177561
R Square	0.666145811
Adjusted R Square	0.629050901
Standard Error	213.3668265
Observations	11

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	817539.5581	817539.5581	17.9578765	0.002181765
Residual	9	409728.6238	45525.40264		
Total	10	1227268.182			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	842.7090571	101.918302	8.268476227	1.69876E-05	612.1536646	1073.264
X Variable 1	**-10.37971726**	**2.449390525**	**-4.237673477**	**0.002181765**	**-15.92062781**	**-4.83881**

This regression is significant; the F-value is 17.9579. If the X-variable explains very little of the Y-variable, you should get an F-value that is 1 or less. In this case, the explained variation (due to regression = explained by the X-variable) is 17.96 times greater than the unexplained (residual) variation. The probability of getting the sample evidence (the X and Y input data) if the X and Y are unrelated (that is the Ho) is .00218. In other words, it is very unlikely to get this kind of data as a result of chance. We have a significant regression.

The regression equation is:
Sales = 842.71 - 10.38 (price).

In theory, at a price of $0, you will sell 842.71 tool sets. For every dollar you raise price, the number of tool sets sold decreases by 10.38.

The correlation coefficient is - .816. It is a strong negative correlation. Note that Excel does not show that the correlation is negative. However, if the b_1 term is negative, the correlation must be negative.
The coefficient of determination, r^2, is 66.6%; the unexplained variation is 33.4%.

Another way to test the regression for significance is to test the b_1 term (slope term which shows the effect of X on Y). This is done via a t-test. The t-value is -4.238 and this is significant. The probability of getting a b_1 of this magnitude if Ho is true (the null hypothesis for this test is that $B_1 = 0$, i.e., the X variable has no effect on Y) is **0.002181765**. Note that this is the same sig. level we got before for the F-test. Indeed, the two tests give exactly the same results.

EXERCISE 2:

Example: A researcher is interested in knowing whether there is a relationship between years of education and longevity. There is a theory that educated people live longer.

Years of Education	Longevity
9	58
10	60
11	63
12	65
13	73
14	74
15	75
16	75
17	77
18	78
15	75
18	78
20	83
10	66
14	70
16	77
17	81

Here is the MS Excel output:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.951065685
R Square	0.904525937
Adjusted R Square	0.898160999
Standard Error	2.34649641
Observations	17

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	782.4681	782.4681425	142.110732	4.73076E-09
Residual	15	82.59068	5.5060454		
Total	16	865.0588			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**40.76702509**	**2.700381**	**15.0967689**	**1.77275E-10**	**35.01129609**	**46.52275409**
X Variable 1	**2.183512545**	**0.183165**	**11.9210206**	**4.73076E-09**	**1.793105562**	**2.573919528**

Note that there were 17 subjects in the study. The regression is significant; the F-value is 142.11. If the X-variable explains very little of the Y-variable, you should get an F-value that is 1 or less. In this case, the explained variation (due to regression = explained by the X-variable)

is 142.11 times greater than the unexplained (residual) variation. The probability of getting the sample evidence (the X and Y input data) if the X and Y are unrelated (i.e., the Ho) is .00000000473. In other words, it is very unlikely to get this kind of data as a result of chance. We have a significant regression.

The regression equation is:
Longevity = 40.77 + 2.18 (years of education).
In theory, an individual with 0 years of education will only live to the age of 40.77. Every year of education increases one's longevity by approximately 2.18 years.

The correlation coefficient is .95. It is a strong positive correlation; the more education one has, the longer one lives. The coefficient of determination, r^2, is 90.5%; the unexplained variation is 9.5%.

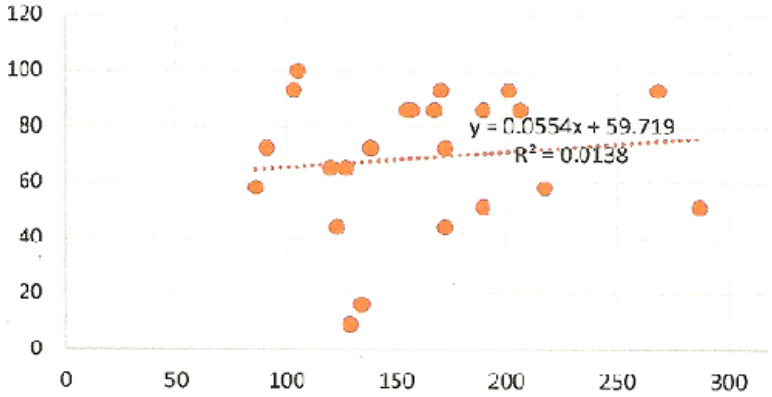
Another way to test the regression for significance is to test the b_1 term (slope term which shows the effect of X on Y). This is done via a t-test. The t-statistic is 11.921 and this is very significant. The probability of getting a b_1 of this magnitude if Ho is true (the null hypothesis for this test is that $B_1 = 0$, i.e., the X variable has no effect on Y) is **4.73076E-09**. Note that this is the same significance level we got before for the F-test. Indeed, the two tests give exactly the same results.

EXERCISE 3:

A researcher wants to determine whether there is a relationship between weight and grade of students on a statistics exam. The data collected are:

Weight	Grade	Weight	Grade	Weight	Grade
91	72	123	44	105	100
155	86	129	9	287	51
157	86	217	58	201	93
86	58	167	86	172	44
120	65	206	86	134	16
268	93	189	51	172	72
170	93	127	65	189	86
138	72	103	93		

See the output from MS Excel below. What is your conclusion? Test at a significance level of (alpha) $\alpha = .05$.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.11734134
R Square	0.01376899
Adjusted R Square	-0.0331944
Standard Error	25.0065518
Observations	23

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	183.3371047	183.337105	0.293185676	0.5938849
Residual	21	13131.88029	625.327633		
Total	22	13315.21739			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	59.719	17.30184	3.451618	0.00239	23.7382	95.70052
X Variable 1	0.0554	0.10239	0.541466	0.59388	-0.1575	0.26836

Answer: There is no (significant) relationship between weight and grade. The F-value is .293 with a p-value of .593. This means that if the null hypothesis is true, and weight is unrelated to grade, there is a probability .593 of getting the sample evidence (or something indicating a stronger relationship). In other words, this is more or less what one expects to see when two variables are unrelated. Note the R^2 value is a paltry .0138, which means that weight only explains 1.38% of the variation in grades. Practically speaking, this is no different from 0. The scatter plot also shows no pattern. Weight does not seem to be related to grade. The 95% confidence interval for the slope term ranges from a negative number (-.1575) to a positive number (.26836). Thus, 0 is the interval. The slope could be 0 and the X-variable would then drop out of the equation. Bottom line: *weight* should not be used to explain or predict *grade*. The regression equation is meaningless.

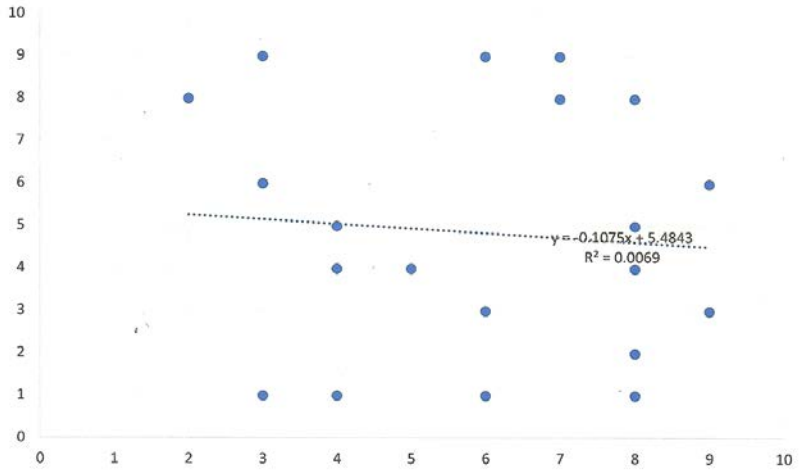
Rule of thumb: When the F-value is 1 or less, it will not be significant.

EXERCISE 4:

A researcher wants to determine whether there is a relationship between hours spent on social media and number of dates. The data:

#hours on social media	3	6	5	9	8	4	6	6	3	2	3	7	8	4	9	8	4	7	8	8
# dates	1	3	4	3	8	1	1	9	6	8	9	8	2	5	6	5	4	9	4	1

See the output from MS Excel below. What is your conclusion? Test at a significance level of (alpha) $\alpha = .05$.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.0831
R Square	0.0069
Adjusted R Square	-0.0483
Standard Error	2.9762
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.1074	1.1074	0.125	0.7278
Residual	18	159.4426	8.8579		
Total	19	160.55			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.4843	1.9135	2.8661	0.0103	1.4642	9.5045
X Variable 1	-0.1075	0.3041	-0.3536	0.7278	-0.7464	0.5313

Answer: There is no relationship between hours on social media and number of dates. The F-value is .125 with a p-value of .7278. This means that if the null hypothesis is true and hours spent on social media is unrelated to number of dates, there is almost a 73% chance of getting the sample evidence. In other words, this is essentially what one expects to see when two variables

are unrelated. Note the R^2 value is a paltry .69% (less than one percent). Hours on social media explains less than 1% (.69%) of the variation in number of dates. Practically speaking, this is no different from 0. The scatter plot also shows no pattern. Hours spent on social media does not seem to be related to number of dates. The 95% confidence interval for the slope term ranges from a negative number (-.7464) to a positive number (.5313). Thus, 0 is the interval. The slope could be 0 and the X variable would then drop out of the equation. Bottom line: *number of hours on social media* should not be used to explain or predict *number of dates*. The regression equation is meaningless.

In theory, if X and Y are totally unrelated, the F-value should be 0 and the significance of F should be 1. This means that the sample evidence totally supports that X and Y are not related. In the real world, however, you do not see F-values of 0 (which also means that r and R^2 are 0).