Anjuman-I-Islam's

Akbar Peerbhoy College of Commerce and Economics M. S. Ali Road, Grant Road (East), Mumbai-8

C E R T I F I C A T E

This is to certify that the work entered in this journal is the work of Mr./Ms._____(Roll No. / Seat No._____) in partial fulfilment for B. Sc. (Data Science) Semester-IV Degree Examination has been found satisfactory in the subject Testing of Hypothesis for the Second Year 2023-24 Year B. Sc. (Data Science) - 2024 in the College Laboratory .

Signature Lecturer-In- Charge Course	Signature External Examiner	Signature Course Coordinator

INDEX

SR NO.	NAME OF EXPERIMENT	DATE	SIGNATURE
1	Perform a Z. Test in excel 1. One sample Z . Test 2. Two sample Z. Test	28/02/2024	
2	Perform a Z.Test in excel	28/02/2024	
3	Perform a T.Test in excel	28/02/2024	
4	Perform a Chi square goodness of fit Test in excel	28/02/2024	
5	Perform a Chi square Test in excel	29/02/2024	
6	Perform a One way Anova in excel	11/03/2024	
7	Perform a One way and Two way Anova in excel	11/03/2024	
8	Perform a One way Anova in excel	11/03/2024	
9	Perform a Wilcoxon Signed Rank Test in Excel.	11/03/2024	
10	Perform a Kruskal-Wallis Test in Excel	11/03/2024	
11	Perform the Friedman Test in Excel	12/03/2024	
12	Perform Multiple Linear Regression in Excel	12/03/2024	
13	How to Perform Simple Linear Regression in Excel	12/03/2024	
14	Perform Polynomial Regression in Excel	12/03/2024	

15	Moving Average in Excel	18/03/2024	
16	Analyze Time Series Data in Excel	18/03/2024	
17	TIME SERIES ANALYSIS AND FORECASTING IN EXCEL	18/03/2024	

Testing of Hyoothesis Practical Journal

Practical No : 1 Perform a Z.Test in excel

- One sample Z TEST

If we have given one dataset, we use the Z TEST function, which falls under the statistical functions category. This Z TEST function in Excel gives the one-tailed probability value of a test.

Z TEST function:

This function gives you the probability that the supplied hypothesized sample mean is greater than the mean of the supplied data values.

Z TEST Function is very simple and easy to use.

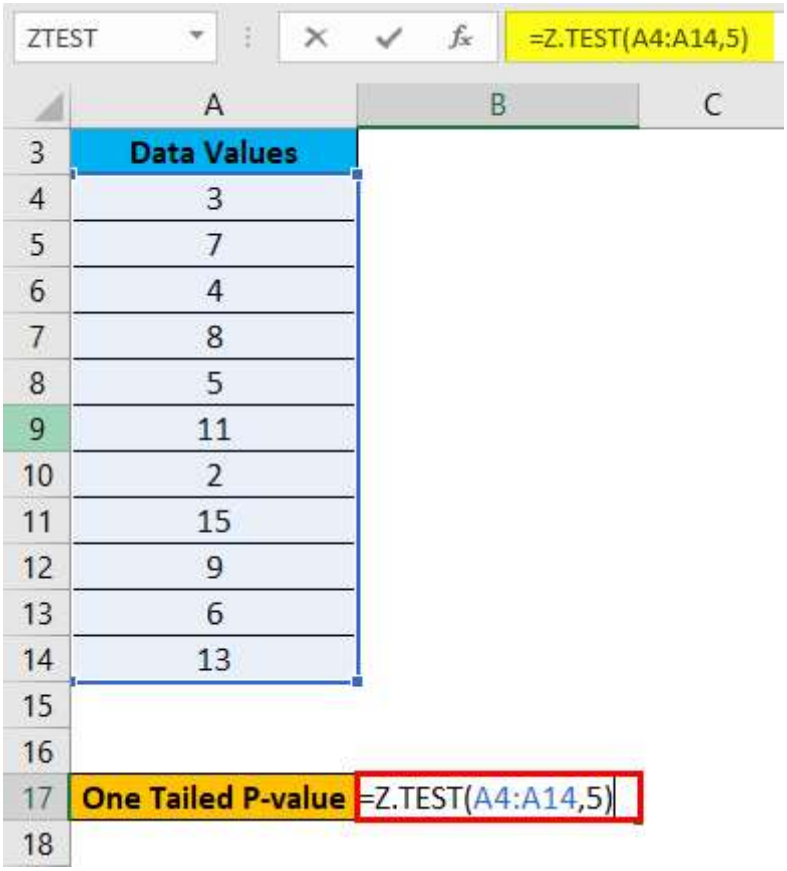
Working of Z TEST Function in Excel with Examples

Example #1

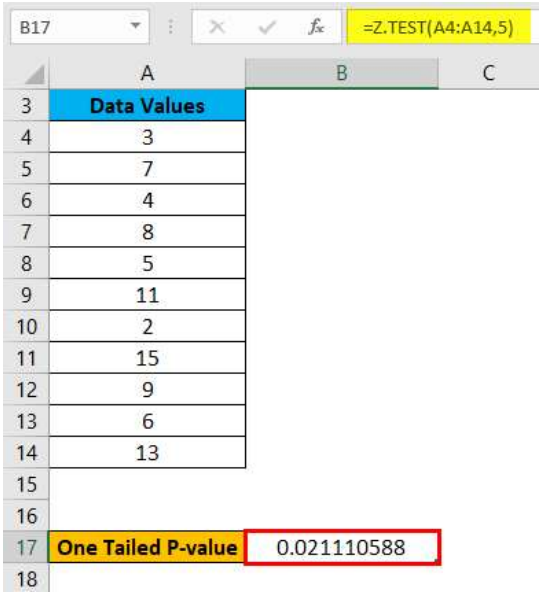
We have given the below set of values:

	A	B
3	**Data Values**	
4	3	
5	7	
6	4	
7	8	
8	5	
9	11	
10	2	
11	15	
12	9	
13	6	
14	13	
15		

To calculate the one-tailed probability value of a Z Test for the above data, let's assume the hypothesized **population mean** is 5. Now we will use the Z TEST formula as shown below:

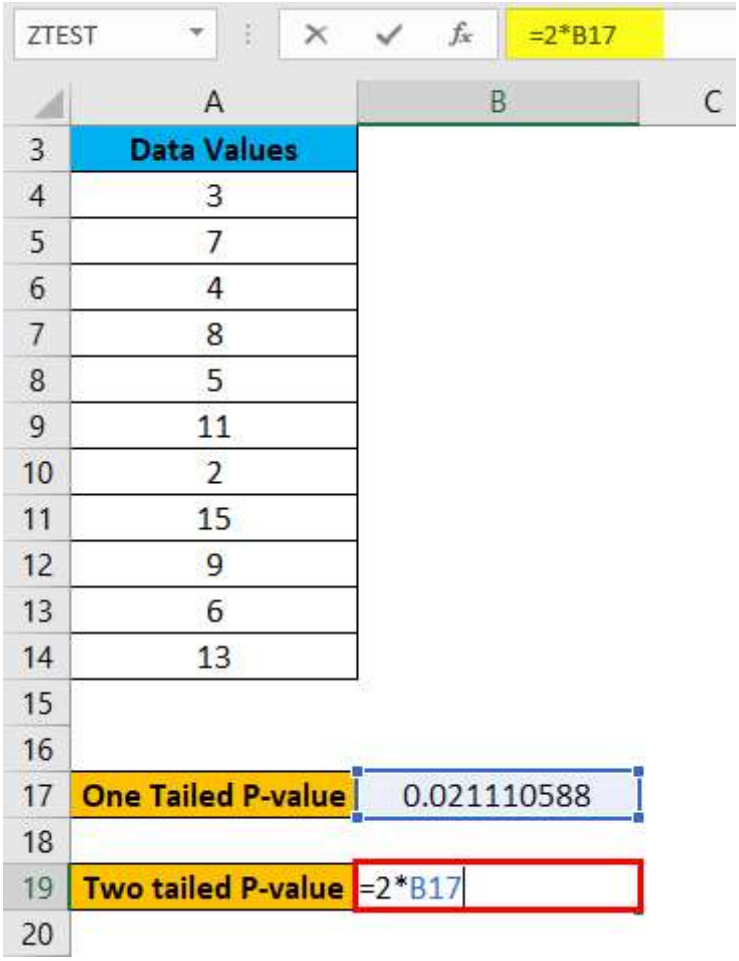


The result is given below:

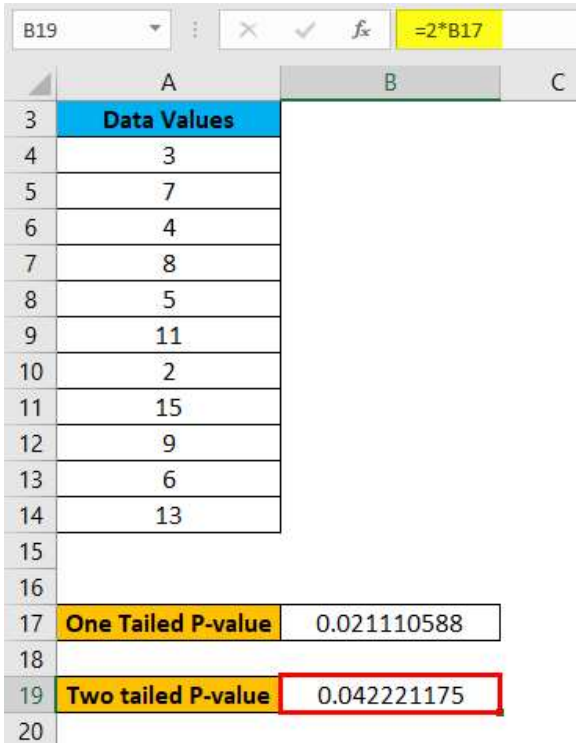


Using the above result, we can also calculate the two-tailed probability of a Z TEST.

The formula below calculates the two-tailed **P-value** of a Z TEST for the given hypothesized population, which is 5.



The result is given below:



Two Sample Z Test:

While using the Z Test, we test a null hypothesis that states that the two population's mean is equal.

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

Where H1 is called an alternative hypothesis, the mean of the two populations is not equal.

Let's take an example to understand the usage of two sample Z tests.

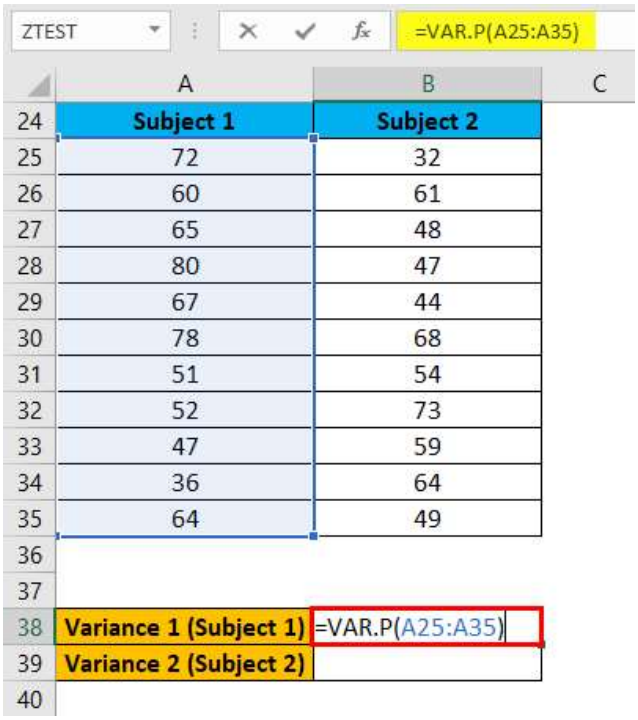
Example #2

Let's take the example of student's marks in two different subjects.

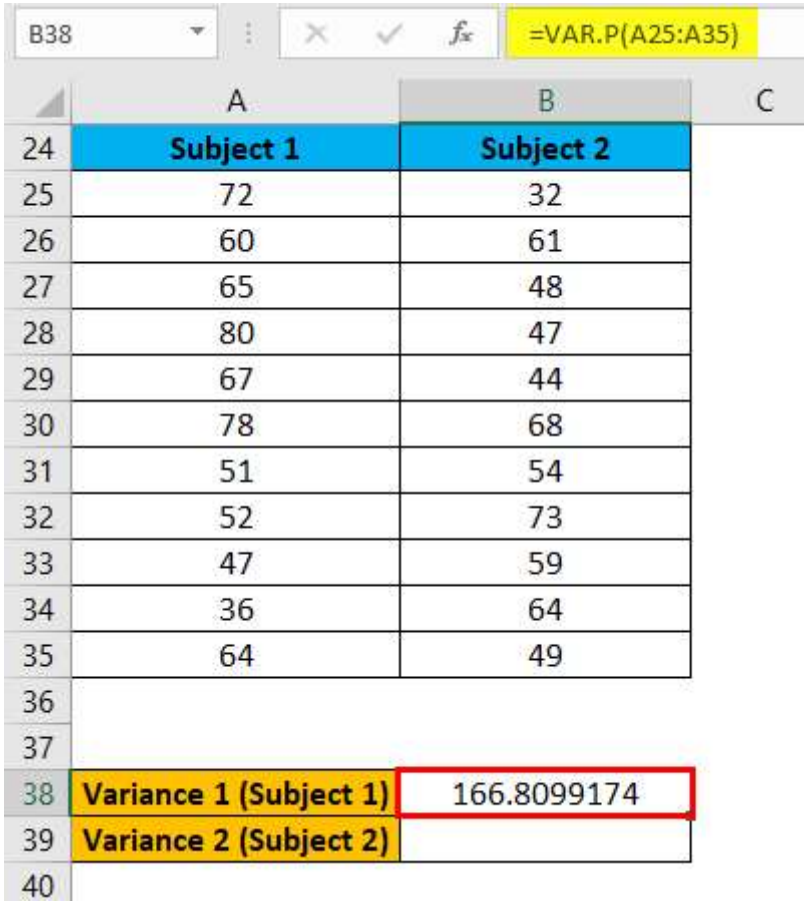
	A	B	C
24	**Subject 1**	**Subject 2**	
25	72	32	
26	60	61	
27	65	48	
28	80	47	
29	67	44	
30	78	68	
31	51	54	
32	52	73	
33	47	59	
34	36	64	
35	64	49	
36			

Now we need to **calculate the variance** of both subjects, so we will use the below formula for this:

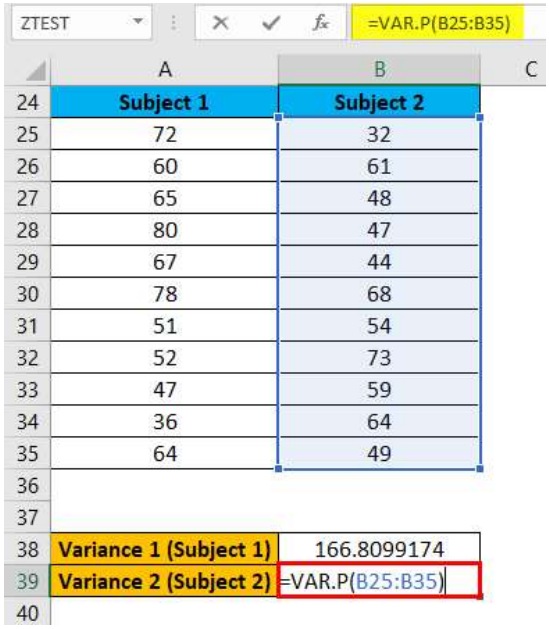
The above formula applies for Variance 1 (Subject 1) like below:



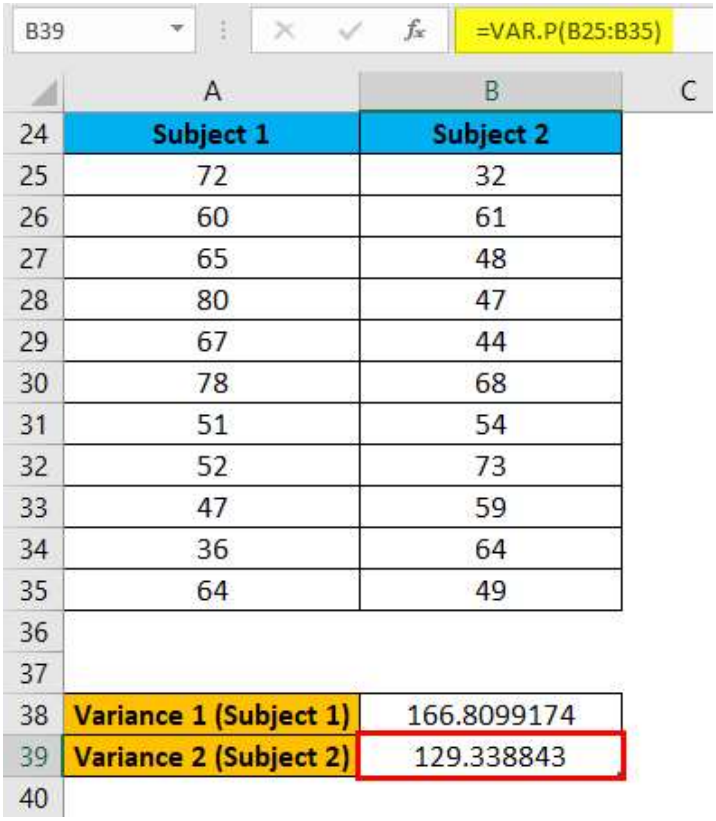
The result is given below:



The above same formula applies for Variance 2 (Subject 2) like below:

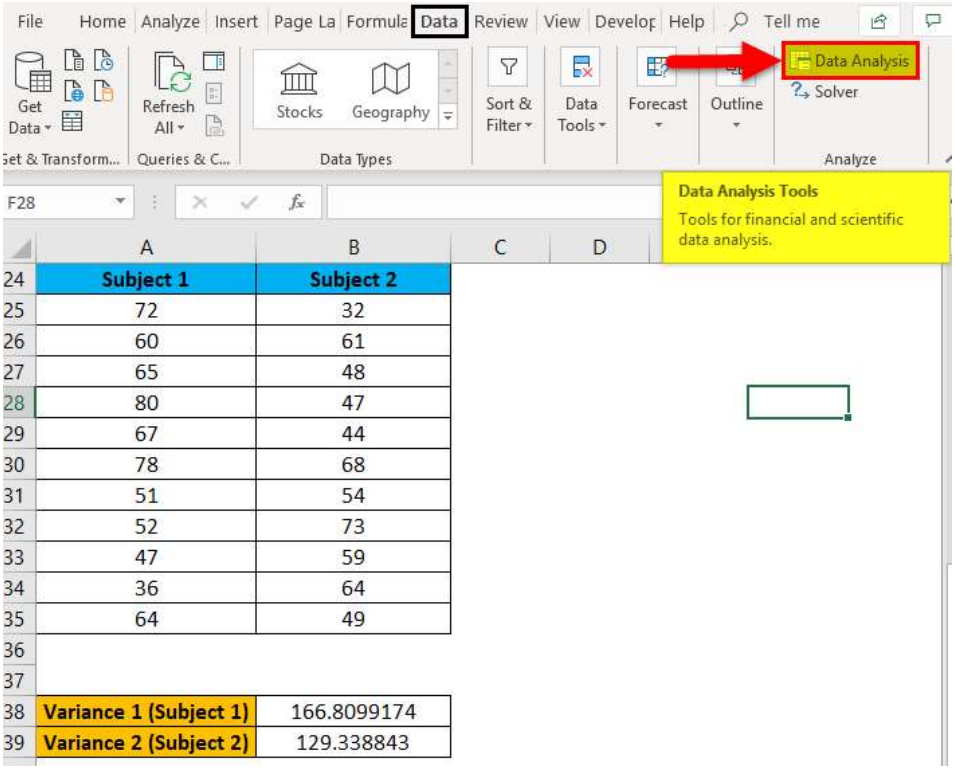


The result is given below:



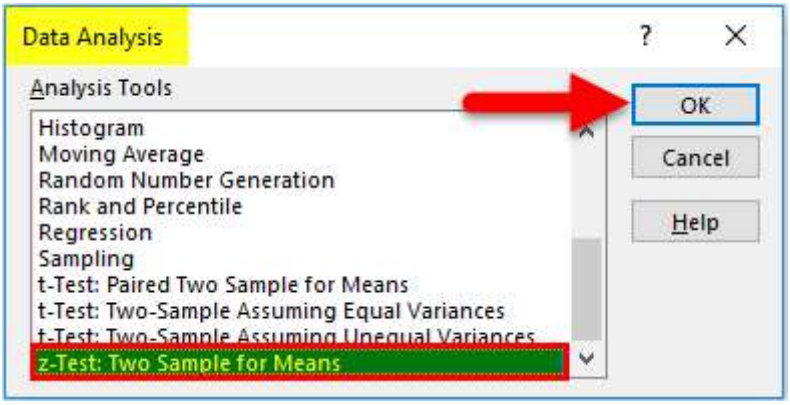
	A	B	C
24	**Subject 1**	**Subject 2**	
25	72	32	
26	60	61	
27	65	48	
28	80	47	
29	67	44	
30	78	68	
31	51	54	
32	52	73	
33	47	59	
34	36	64	
35	64	49	
36			
37			
38	**Variance 1 (Subject 1)**	166.8099174	
39	**Variance 2 (Subject 2)**	129.338843	
40			

- Now, Go to the Data Analysis tab in the extreme upper right corner under the DATA tab as shown below screenshot:

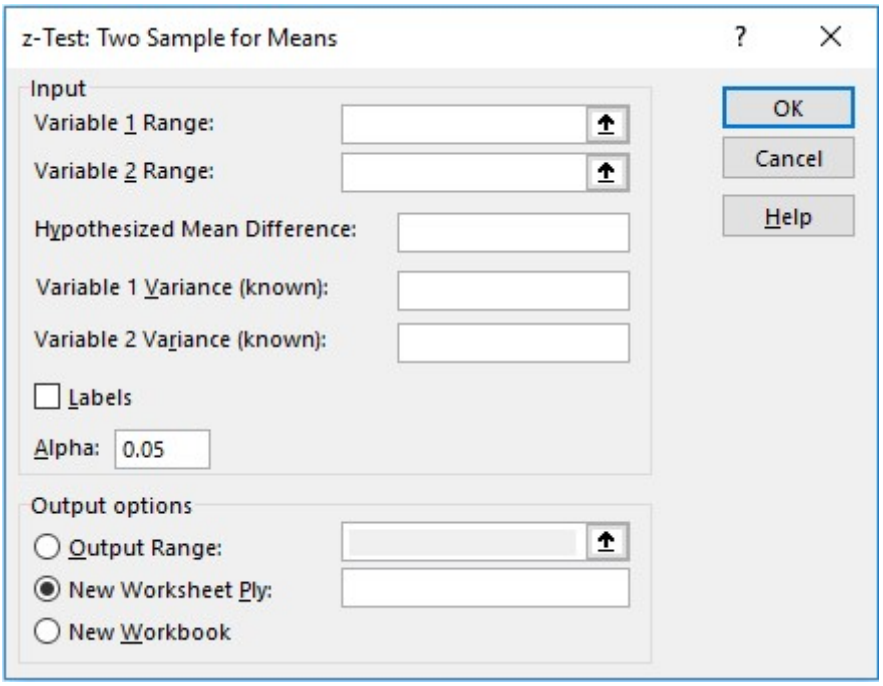


	A	B	C	D
24	**Subject 1**	**Subject 2**		
25	72	32		
26	60	61		
27	65	48		
28	80	47		
29	67	44		
30	78	68		
31	51	54		
32	52	73		
33	47	59		
34	36	64		
35	64	49		
36				
37				
38	**Variance 1 (Subject 1)**	166.8099174		
39	**Variance 2 (Subject 2)**	129.338843		

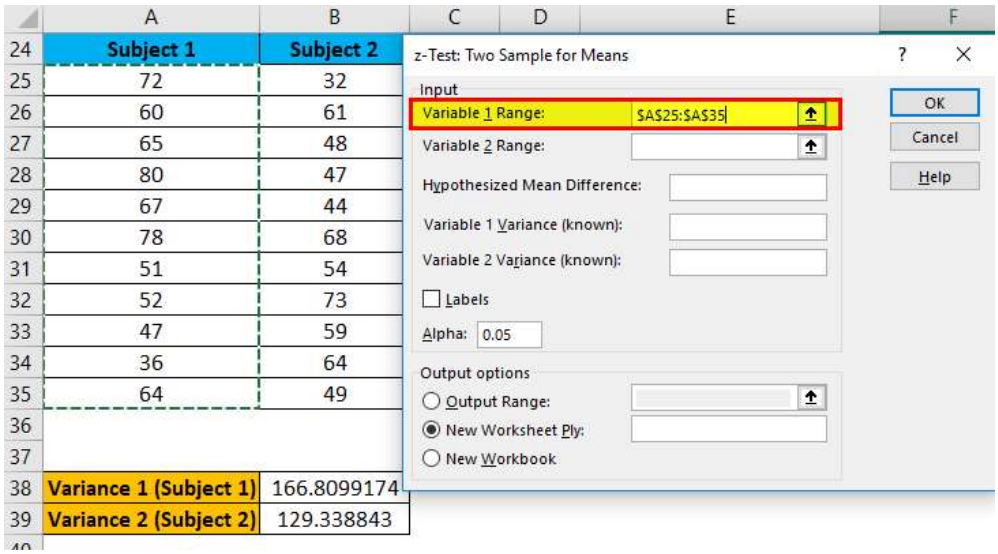
- It will open a dialog box with **Data Analysis** options.
- Click on **z-Test: Two-Sample for Means** and click on **OK,** as shown below.



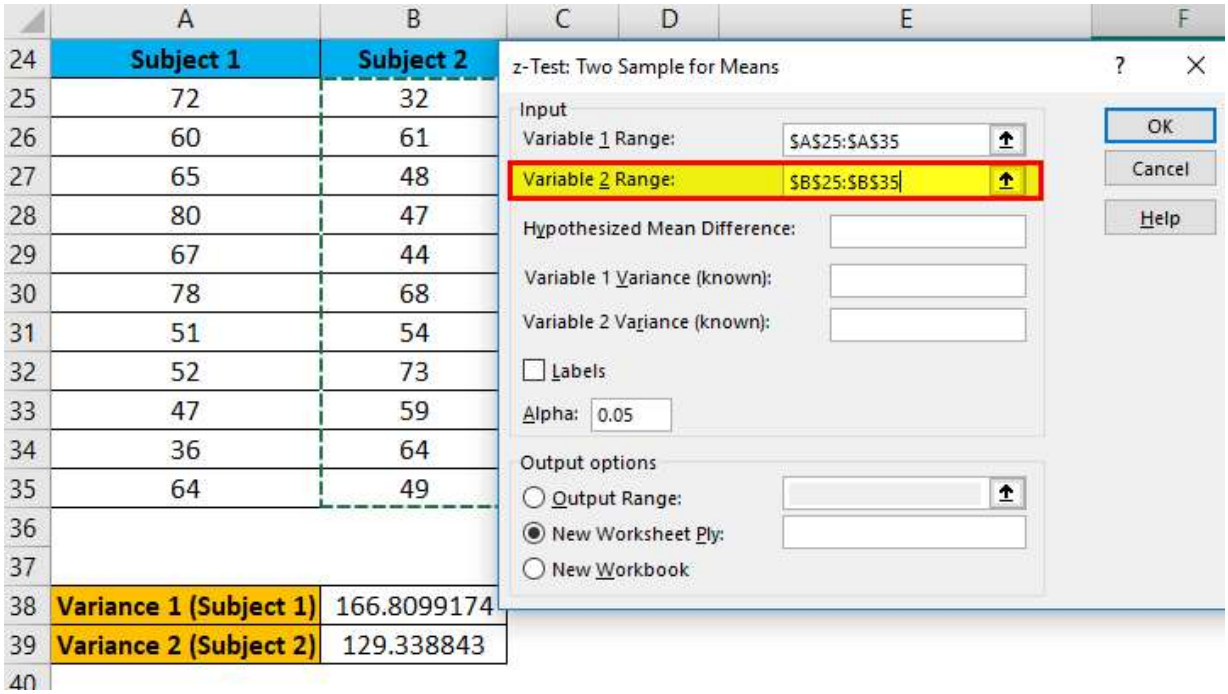
- It will open a dialog box for **Z-test,** as shown below.



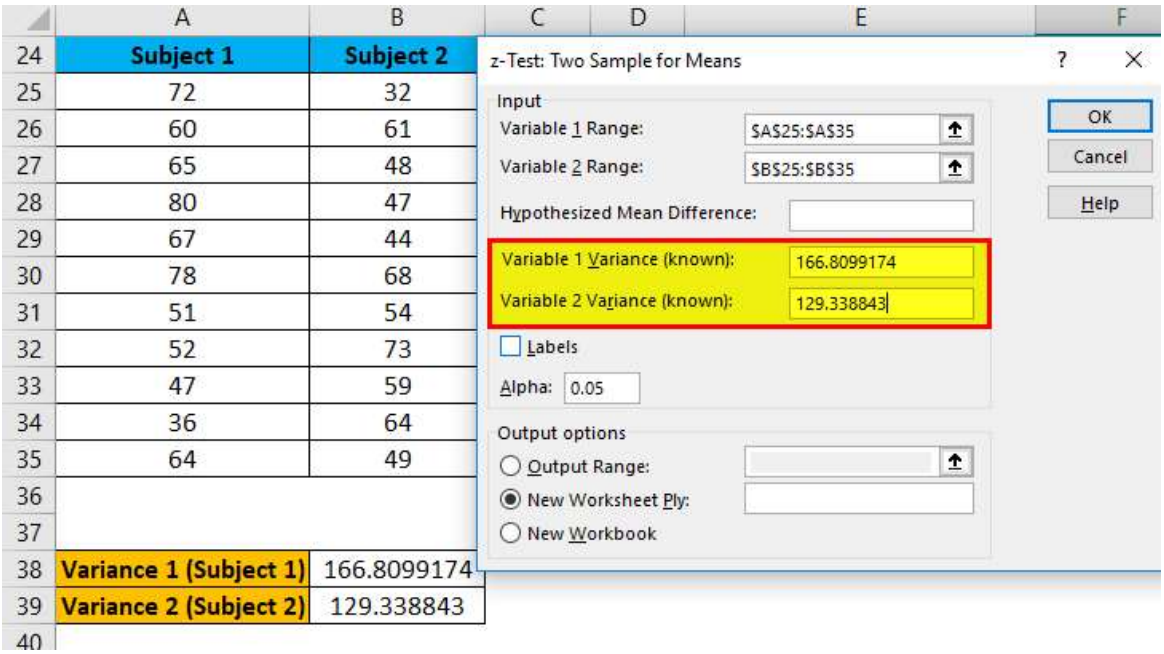
- Now in the Variable 1 range box, select subject 1 range from A25:A35



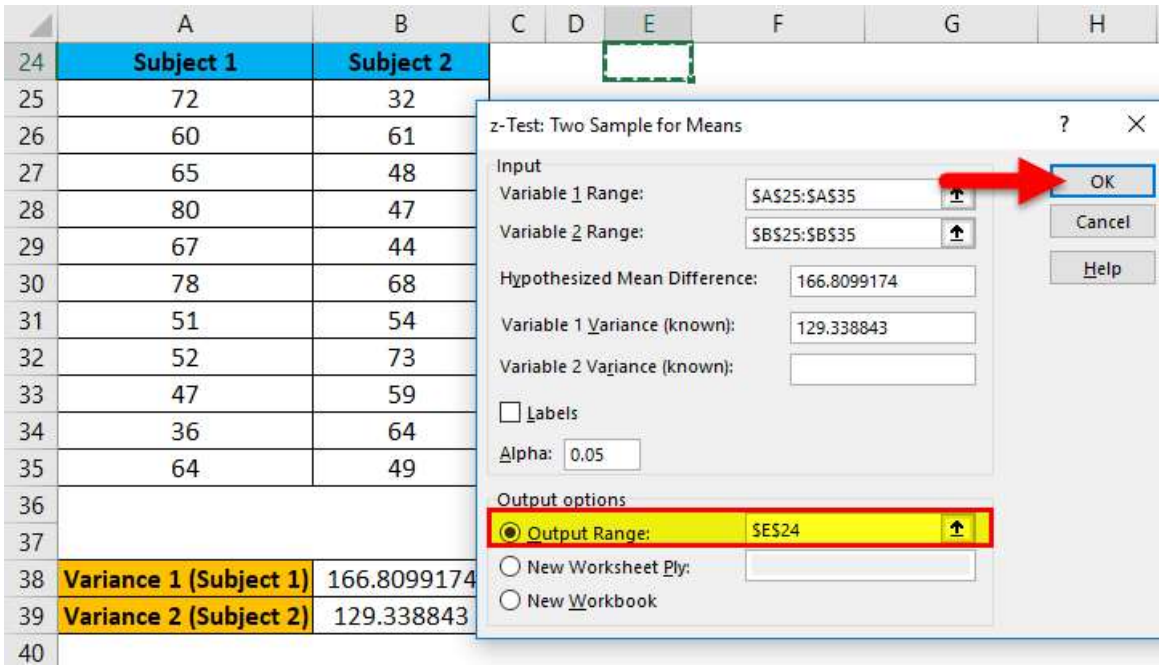
- Similarly, in the Variable 2 range box, select subject 2 range from B25:B35



- Under the Variable 1 variance box, enter cell B38 variance value.
- Under the Variable 2 variance box, enter cell B39 variance value.



- In Output Range, Select the cell where you want to see the result. Here we have passed cell E24 and then clicked on **OK**.



The result is shown below:

	A	B	C	E	F	G
24	**Subject 1**	**Subject 2**		z-Test: Two Sample for Means		
25	72	32				
26	60	61			*Variable 1*	*Variable 2*
27	65	48		Mean	61.09090909	54.45454545
28	80	47		Known Variance	166.8099174	129.338843
29	67	44		Observations	11	11
30	78	68		Hypothesized Mean Difference	0	
31	51	54		z	1.279002985	
32	52	73		P(Z<=z) one-tail	0.100448002	
33	47	59		z Critical one-tail	1.644853627	
34	36	64		P(Z<=z) two-tail	0.200896005	
35	64	49		z Critical two-tail	1.959963985	
36						
37						
38	**Variance 1 (Subject 1)**	166.8099174				
39	**Variance 2 (Subject 2)**	129.338843				
40						

Explanation :

- We can reject the null hypothesis if z < -z Critical two-tail or z stat > z Critical two-tail.
- Here 1.279 > -1.9599 and 1.279 < 1.9599; hence we can't reject the null hypothesis.
- Thus, the means of both populations don't differ significantly.

Practical No : 2 Perform a Z.Test in excel

A **one sample z-test** is used to test whether a population mean is significantly different than some hypothesized value.

A **two sample z-test** is used to test whether two population means are significantly different from each other.

The following examples show how to perform each type of test in Excel.

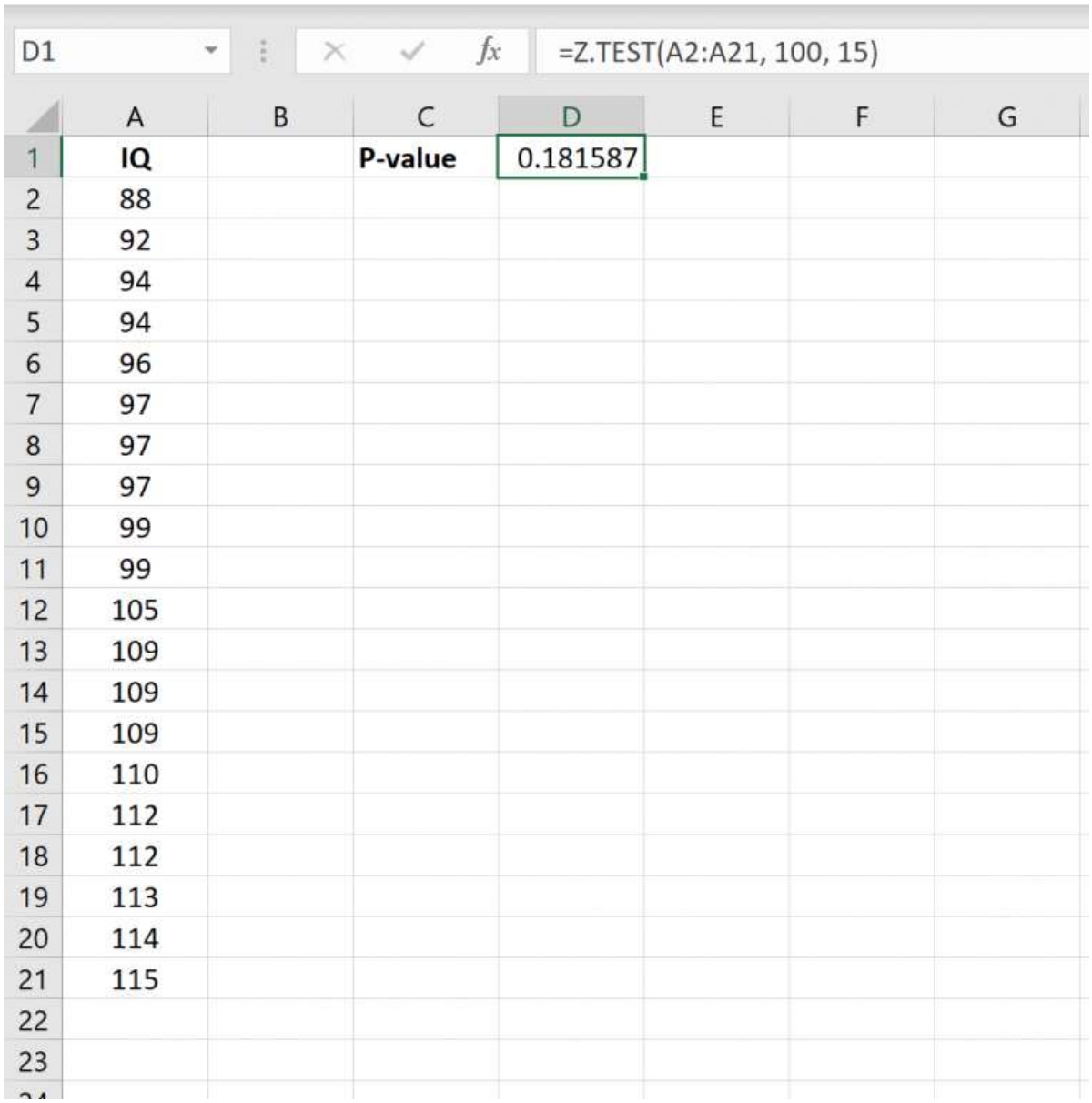
Example 1: One Sample Z-Test in Excel

Suppose the IQ in a population is normally distributed with a mean of $\mu = 100$ and standard deviation of $\sigma = 15$.

A scientist wants to know if a new medication affects IQ levels, so she recruits 20 patients to use it for one month and records their IQ levels at the end of the month.

We can use the following formula in Excel to perform a one sample z-test to determine if the new medication causes a significant difference in IQ levels:

The following screenshot shows how to use this formula in practice:



	A	B	C	D	E	F	G
1	**IQ**		**P-value**	0.181587			
2	88						
3	92						
4	94						
5	94						
6	96						
7	97						
8	97						
9	97						
10	99						
11	99						
12	105						
13	109						
14	109						
15	109						
16	110						
17	112						
18	112						
19	113						
20	114						
21	115						
22							
23							

The one-tailed p-value is **0.181587**. Since we're performing a two-tailed test, we can multiply this value by 2 to get p = **0.363174**.

Since this p-value is not less than .05, we do not have sufficient evidence to reject the null hypothesis.

Thus, we conclude that the new medication does not significantly affect IQ level.

Example 2: Two Sample Z-Test in Excel

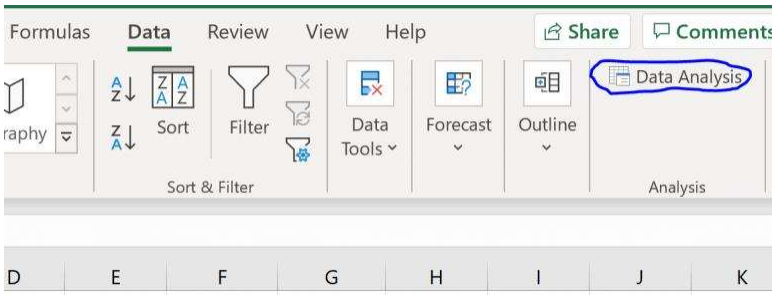
Suppose the IQ levels among individuals in two different cities are known to be normally distributed each with population standard deviations of 15.

A scientist wants to know if the mean IQ level between individuals in city A and city B are different, so she selects a simple random sample of 20 individuals from each city and records their IQ levels.

The following screenshot shows the IQ levels for the individuals in each sample:

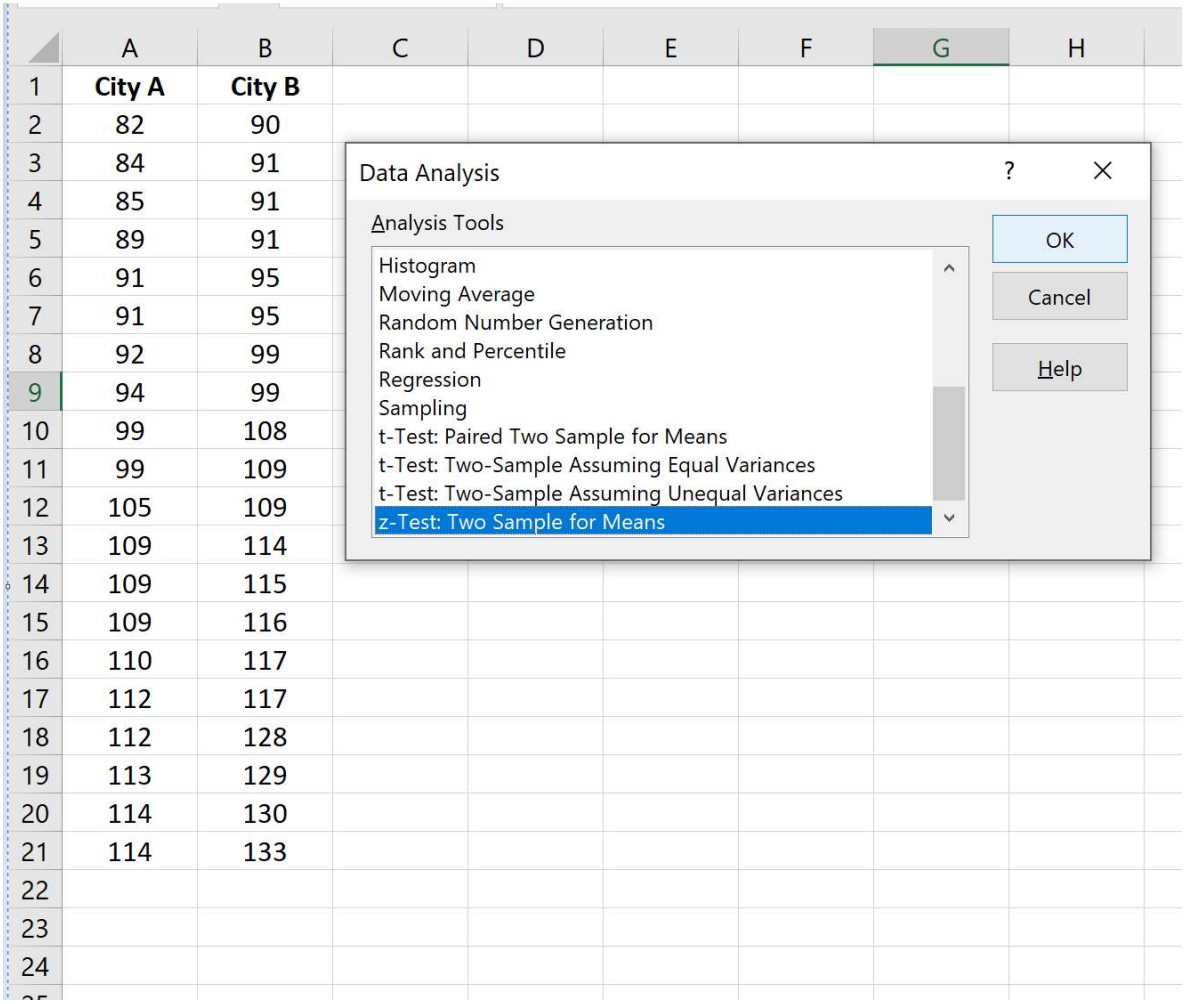
	A	B	C	D	E	F	G
1	**City A**	**City B**					
2	82	90					
3	84	91					
4	85	91					
5	89	91					
6	91	95					
7	91	95					
8	92	99					
9	94	99					
10	99	108					
11	99	109					
12	105	109					
13	109	114					
14	109	115					
15	109	116					
16	110	117					
17	112	117					
18	112	128					
19	113	129					
20	114	130					
21	114	133					
22							
23							
24							
25							

To perform a two sample z-test to determine if the mean IQ level is different between the two cities, click the **Data** tab along the top ribbon, then click the **Data Analysis** button within the **Analysis** group.



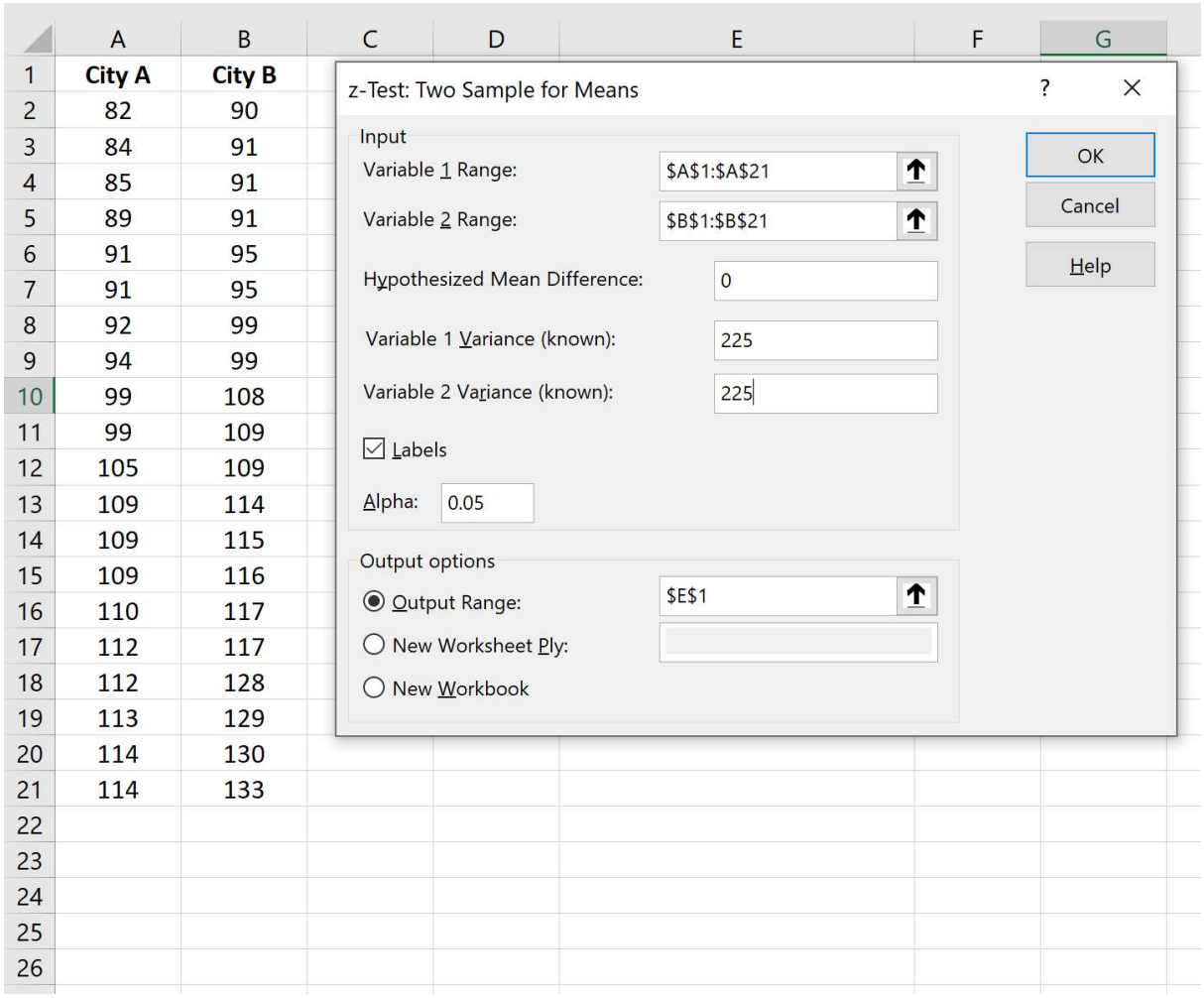
If you don't see **Data Analysis** as an option, you need to first load the Analysis ToolPak in Excel.

Once you click this button, select **z-Test: Two Sample for Means** in the new window that appears:



	A	B
1	**City A**	**City B**
2	82	90
3	84	91
4	85	91
5	89	91
6	91	95
7	91	95
8	92	99
9	94	99
10	99	108
11	99	109
12	105	109
13	109	114
14	109	115
15	109	116
16	110	117
17	112	117
18	112	128
19	113	129
20	114	130
21	114	133
22		
23		
24		

Once you click **OK**, you can fill in the following information:



Once you click **OK**, the results will appear in cell E1:

	A	B	C	D	E	F	G	H
1	**City A**	**City B**			z-Test: Two Sample for Means			
2	82	90						
3	84	91				*City A*	*City B*	
4	85	91			Mean	100.65	108.8	
5	89	91			Known Variance	225	225	
6	91	95			Observations	20	20	
7	91	95			Hypothesized Mean Difference	0		
8	92	99			z	-1.71817		
9	94	99			P(Z<=z) one-tail	0.042883		
10	99	108			z Critical one-tail	1.644854		
11	99	109			P(Z<=z) two-tail	0.085765		
12	105	109			z Critical two-tail	1.959964		
13	109	114						
14	109	115						
15	109	116						
16	110	117						
17	112	117						
18	112	128						
19	113	129						
20	114	130						
21	114	133						
22								
23								
24								

The test statistic for the two sample z-test is **-1.71817**and the corresponding p-value is **.085765.**

Since this p-value is not less than .05, we do not have sufficient evidence to reject the null hypothesis.

Thus, we conclude that the mean IQ level is not significantly different between the two cities.

Practical No : 3 Perform a T.Test in excel

Example #1

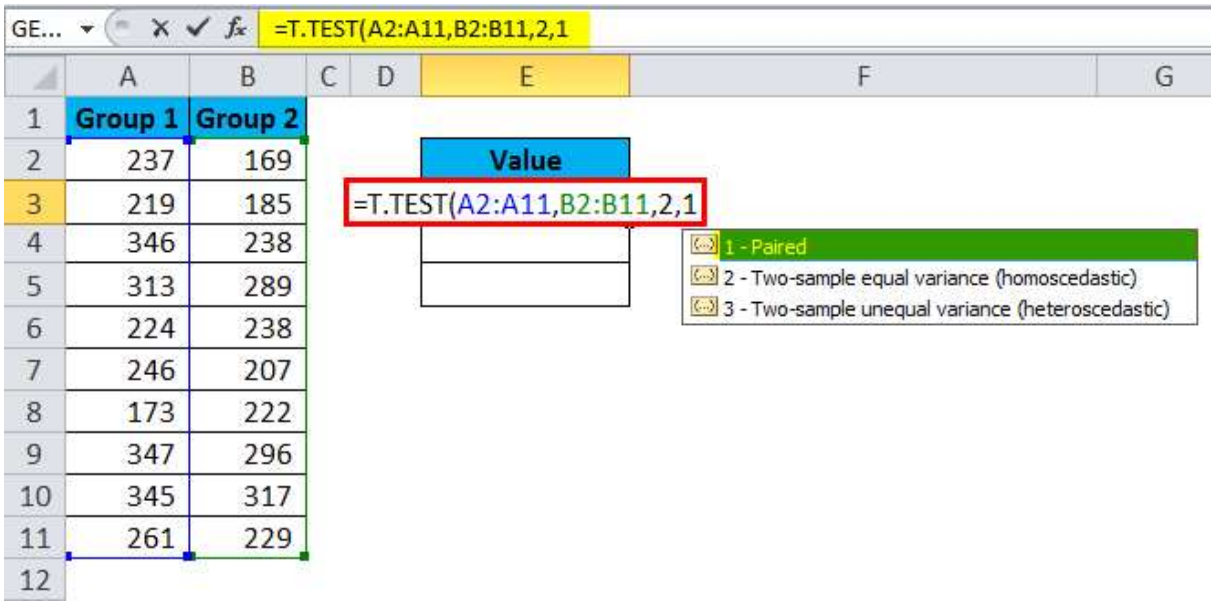
The functionality of the T.TEST in Excel can be best explained by using an example dataset to get the logic of the T.TEST.

I have Group 1 and Group 2 test scores for a classroom. I need to run T.TEST to find a significant difference between these two groups.

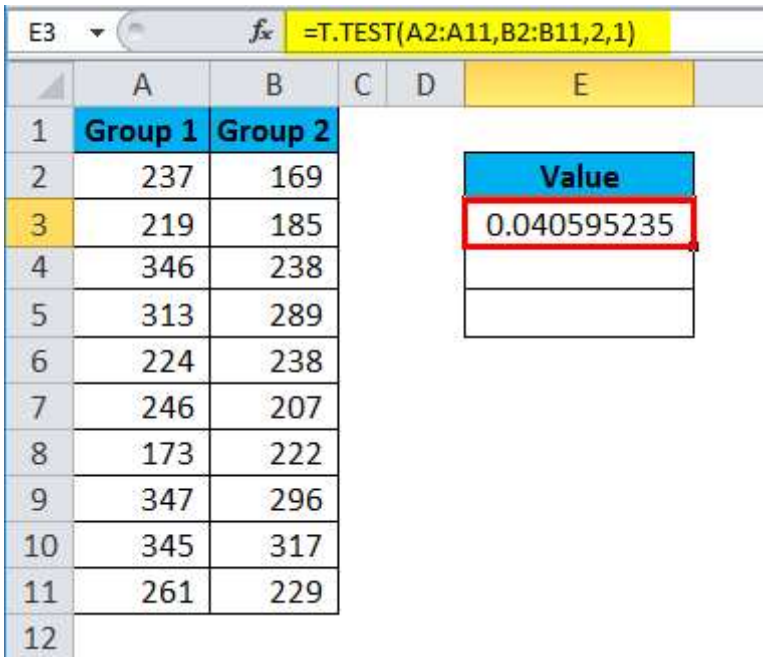
	A	B	C
1	**Group 1**	**Group 2**	
2	237	169	
3	219	185	
4	346	238	
5	313	289	
6	224	238	
7	246	207	
8	173	222	
9	347	296	
10	345	317	
11	261	229	
12			

Apply T.TEST to get the difference.

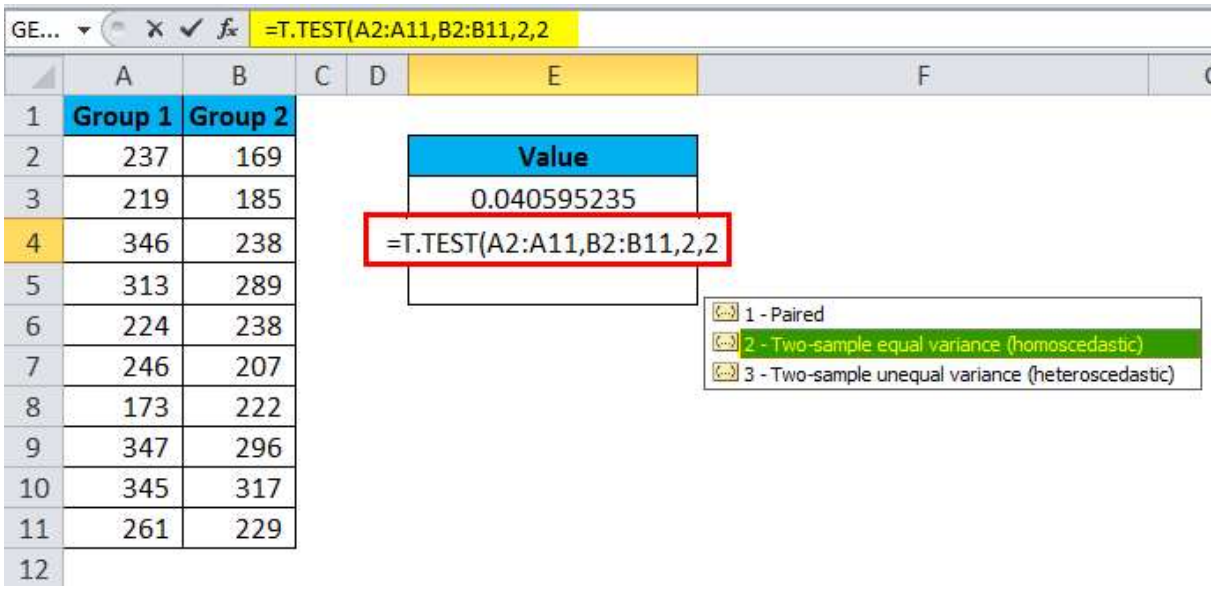
The first test is a type of Paired.



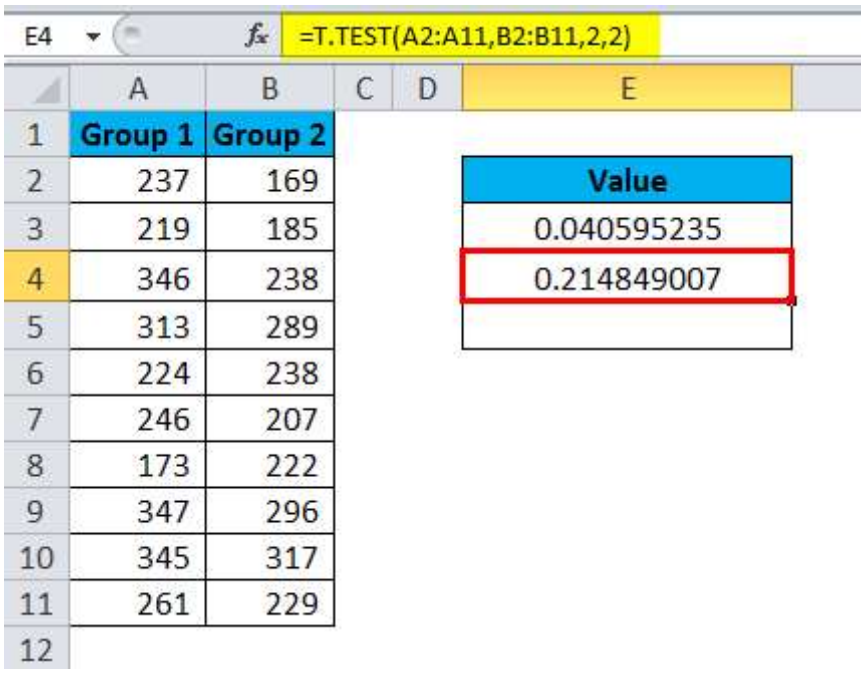
The result is 0.04059.



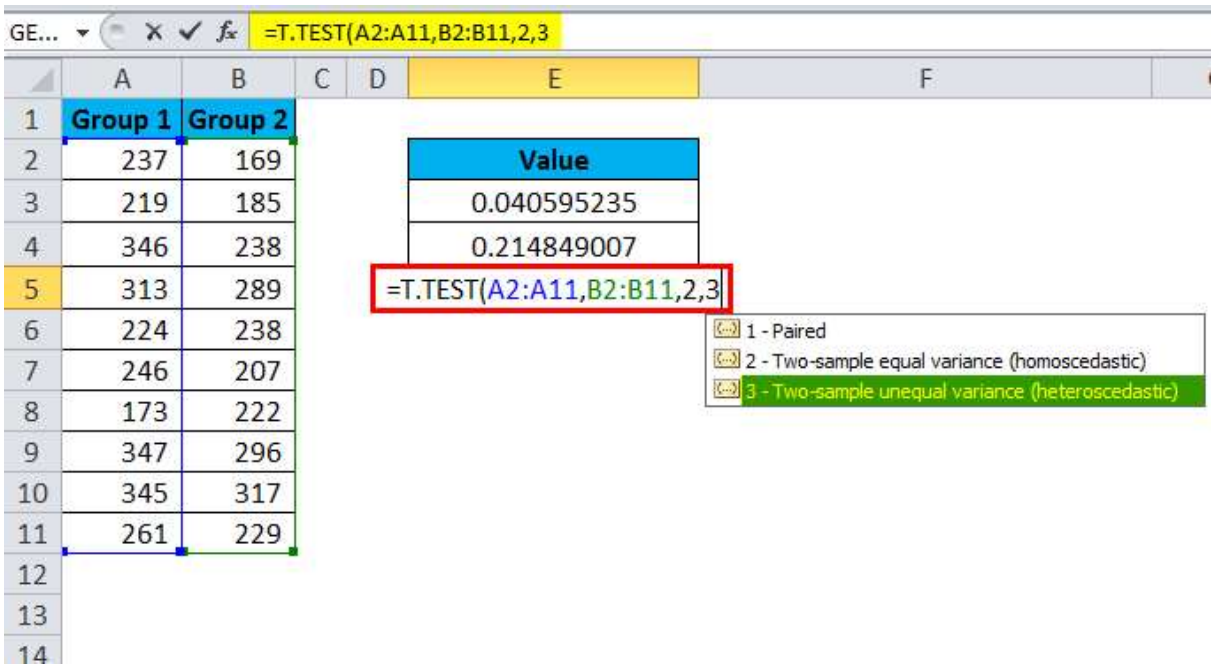
The second test is a type of Two Samples equal variance.



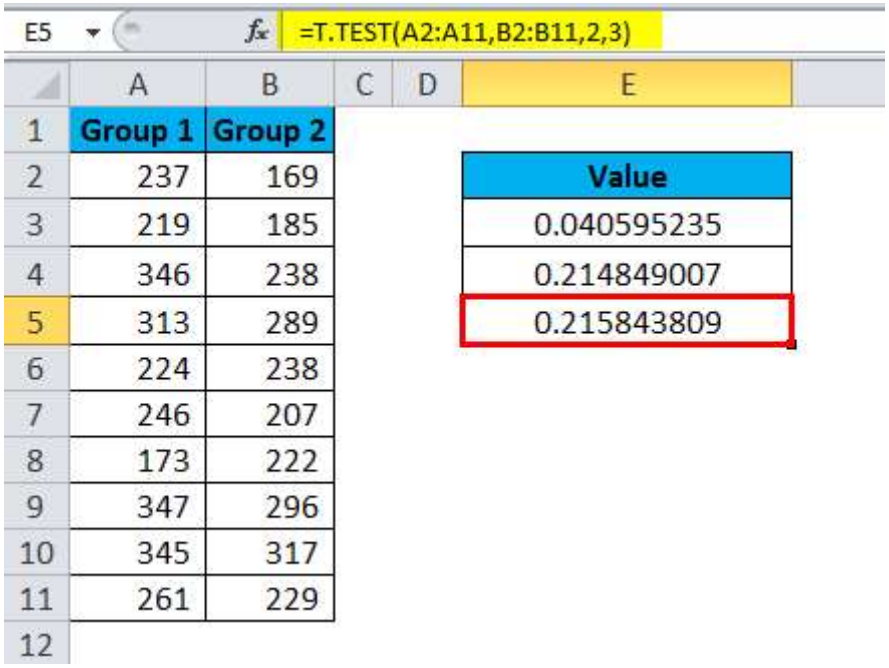
The result is 0.2148.



The third test is a type of Two Sample unequal variance.



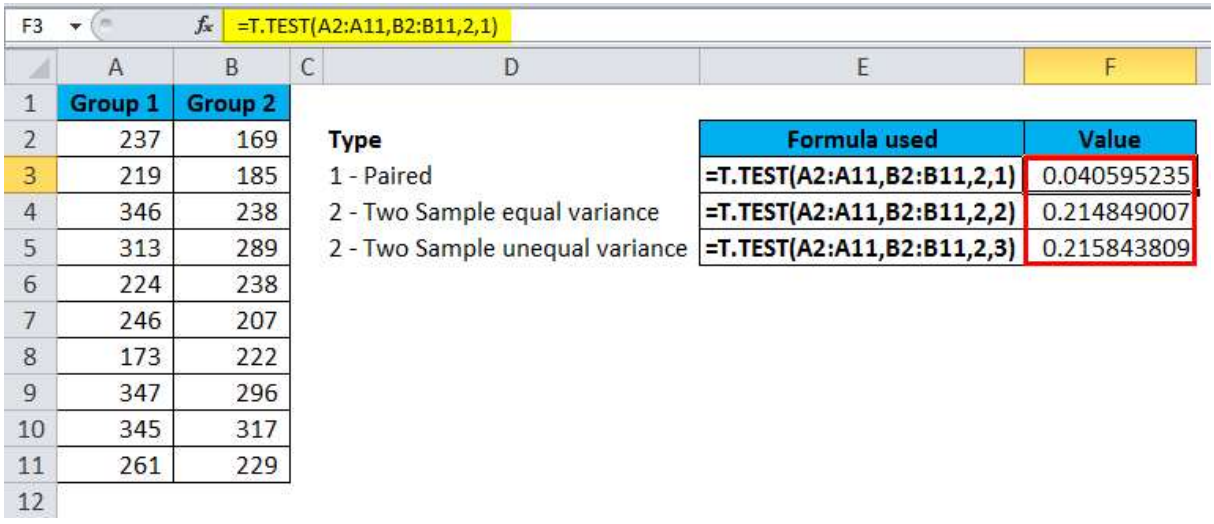
The result is 0.2158.



E5 =T.TEST(A2:A11,B2:B11,2,3)

	A	B	C	D	E
1	Group 1	Group 2			
2	237	169			Value
3	219	185			0.040595235
4	346	238			0.214849007
5	313	289			0.215843809
6	224	238			
7	246	207			
8	173	222			
9	347	296			
10	345	317			
11	261	229			
12					

The returned value is generally called the P-value. If the P-value is <0.05, we can come to the conclusion that the two sets of data have a different mean. Otherwise, the two means are not significantly different from each other.



F3 =T.TEST(A2:A11,B2:B11,2,1)

	A	B	C	D	E	F
1	Group 1	Group 2				
2	237	169		Type	Formula used	Value
3	219	185		1 - Paired	=T.TEST(A2:A11,B2:B11,2,1)	0.040595235
4	346	238		2 - Two Sample equal variance	=T.TEST(A2:A11,B2:B11,2,2)	0.214849007
5	313	289		2 - Two Sample unequal variance	=T.TEST(A2:A11,B2:B11,2,3)	0.215843809
6	224	238				
7	246	207				
8	173	222				
9	347	296				
10	345	317				
11	261	229				
12						

Example #2

I have the salary numbers of two different departments. I want to find out if the mean of the two departments' salaries is significantly different.

	A	B
1	**Sales**	**Marketing**
2	607,830	729,468
3	638,729	1,019,985
4	570,067	961,383
5	267,357	1,002,869
6	610,916	786,574
7	660,227	1,045,148
8	264,191	801,772
9	362,070	687,686
10	339,602	868,793
11	517,509	662,938
12		

Apply the T.TEST function to see the difference.

	A	B	C	D	E
1	**Sales**	**Marketing**			
2	607,830	729,468		**Type**	**Formula**
3	638,729	1,019,985		1 - Paired	**=T.TEST(A2:A11,B2:B11,2,1)**
4	570,067	961,383		2 - Two Sample equal variance	**=T.TEST(A2:A11,B2:B11,2,2)**
5	267,357	1,002,869		2 - Two Sample unequal variance	**=T.TEST(A2:A11,B2:B11,2,3)**
6	610,916	786,574			
7	660,227	1,045,148			
8	264,191	801,772			
9	362,070	687,686			
10	339,602	868,793			
11	517,509	662,938			
12					

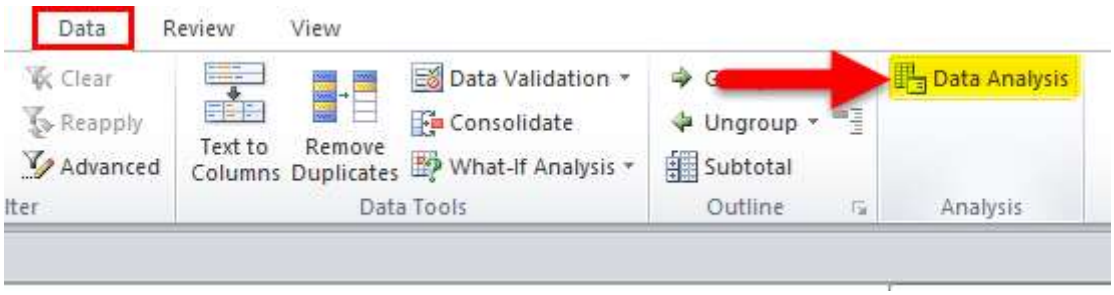
The above formula returns the result as:

F3 =T.TEST(A2:A11,B2:B11,2,1)

	A	B	C	D	E	F
1	**Sales**	**Marketing**				
2	607,830	729,468		**Type**	**Formula used**	**Value**
3	638,729	1,019,985		1 - Paired	**=T.TEST(A2:A11,B2:B11,2,1)**	0.000186102
4	570,067	961,383		2 - Two Sample equal variance	**=T.TEST(A2:A11,B2:B11,2,2)**	3.06844E-05
5	267,357	1,002,869		2 - Two Sample unequal variance	**=T.TEST(A2:A11,B2:B11,2,3)**	3.17896E-05
6	610,916	786,574				
7	660,227	1,045,148				
8	264,191	801,772				
9	362,070	687,686				
10	339,602	868,793				
11	517,509	662,938				
12						

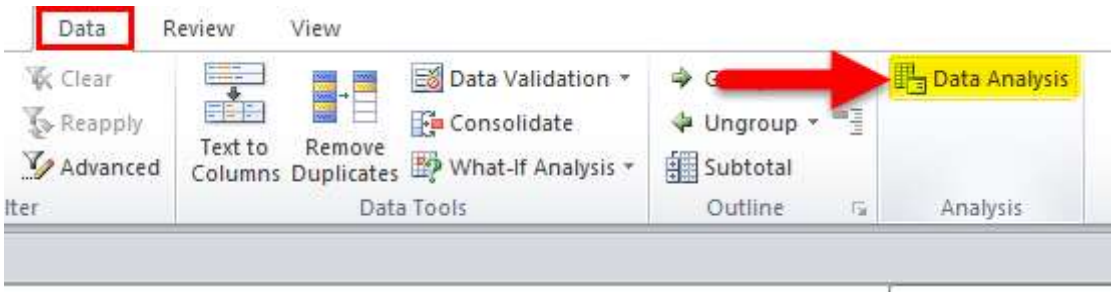
Run T.TEST using Analysis Tool Pack

We can run the T.TEST using the analysis tool pack under the Data ribbon tab.

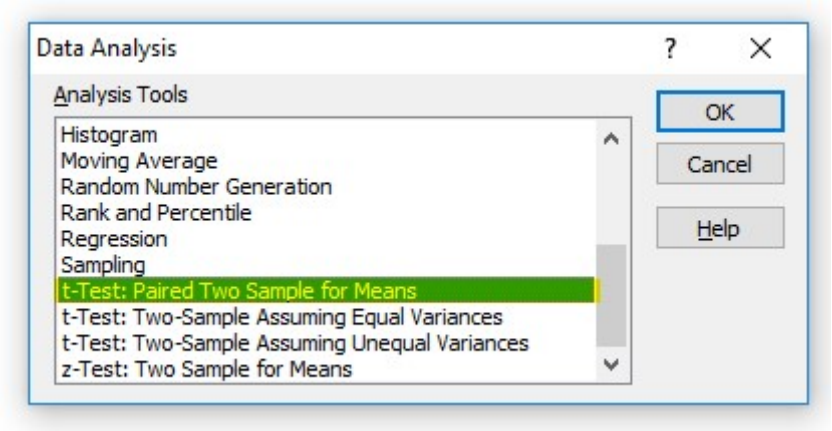


Follow the below Steps to Run T.TEST using Data Analysis ToolPak

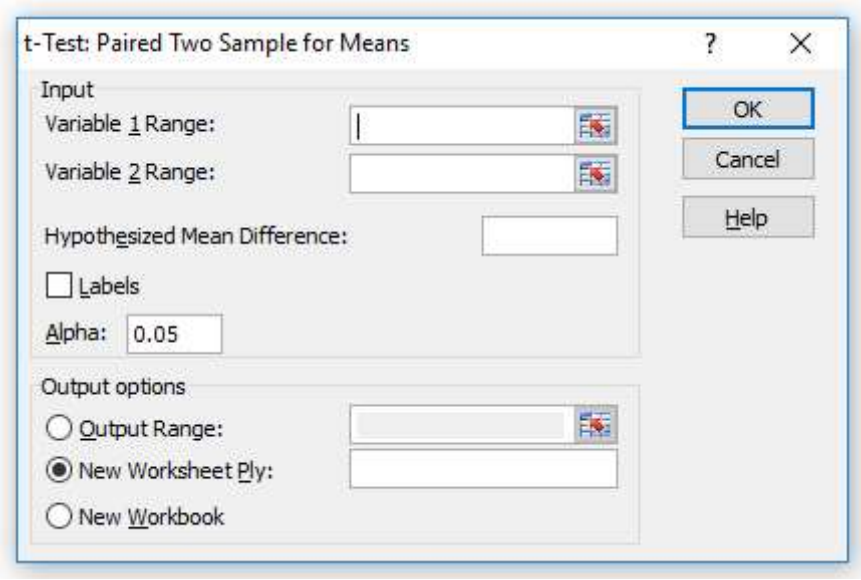
Step 1: Click on Data Analysis under the Data tab.



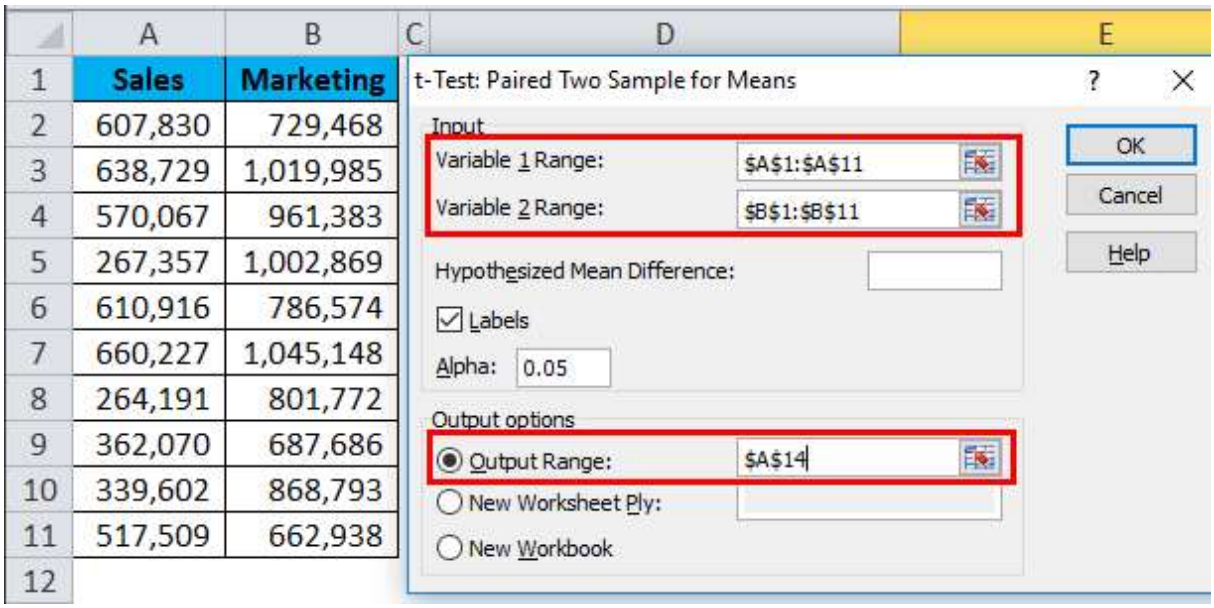
Step 2: Scroll down find the t-Test.



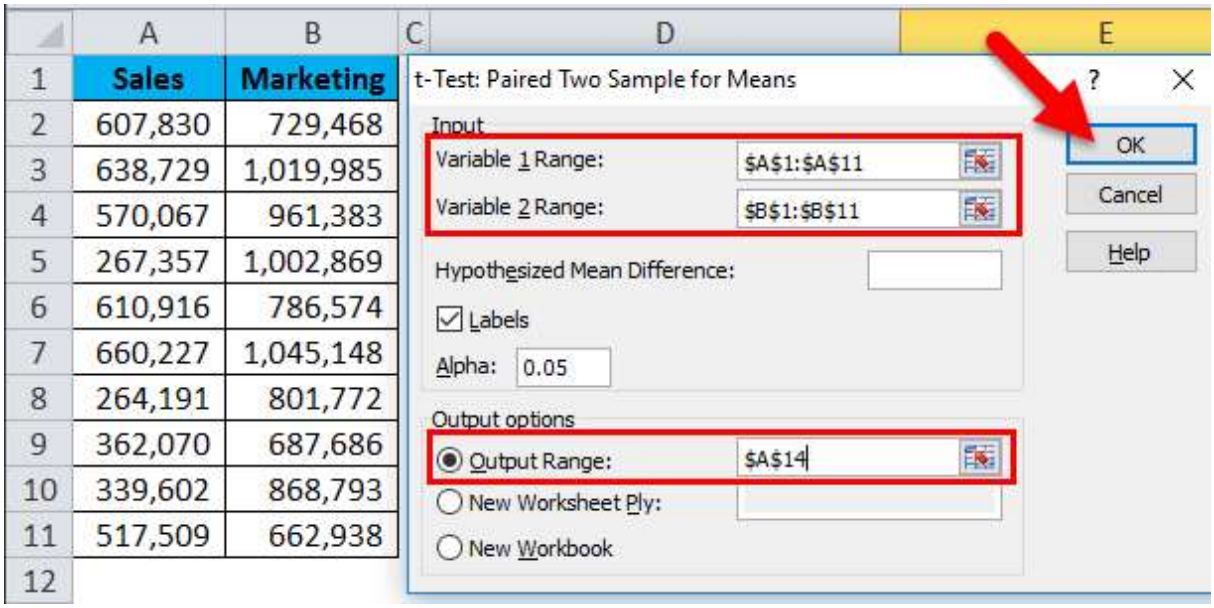
Step 3: Click on the first t-test, and it will open the below dialogue box.



Step 4: Select the Variable 1 range and Variable 2 range.



Step 5: Once all the above boxes filled, click on the OK button.



It will show a detailed report.

B26 f_x 0.00018610207926053

	A	B	C
1	**Sales**	**Marketing**	
2	607,830	729,468	
3	638,729	1,019,985	
4	570,067	961,383	
5	267,357	1,002,869	
6	610,916	786,574	
7	660,227	1,045,148	
8	264,191	801,772	
9	362,070	687,686	
10	339,602	868,793	
11	517,509	662,938	
12			
13			
14	t.Test: Paired Two Sample Means		
15			

16		**Sales**	**Marketing**
17	Mean	483849.8	856661.6
18	Variance	25104067570	20541882290
19	Observations	10	10
20	Pearson Correlation	0.174211658	
21	Hypothesized Mean Difference	0	
22	df	9	
23	t Stat	-6.069108325	
24	P(T<=t) one-tail	9.3051E-05	
25	t Critical one-tail	1.833112933	
26	P(T<=t) two-tail	0.000186102	
27	t Critical two-tail	2.262157163	

This will show the mean of each data set, their variance, how many observations are considered, correlation, and P-value.

We need to see the P-value (refer to B26), i.e. 0.000186102, which is way below the expected P-value of 0.05.

Our data is significant as long as the P-value is less than 0.05.

Practical No : 4 Perform a Chi square goodness of fit Test in excel

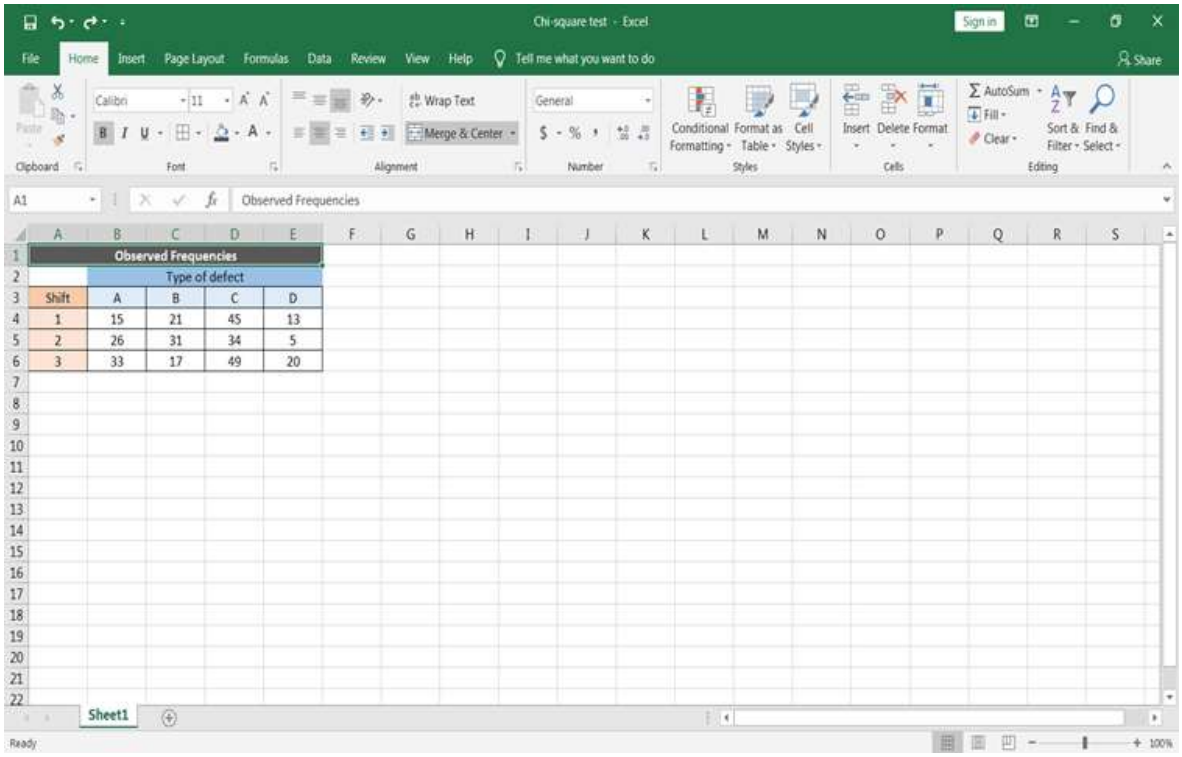
Example of the Chi-square test

Suppose you wish to classify defects in the furniture produced by a manufacturing plant based on the type of defects and the production shift. A total of 390 furniture defects were recorded, and the defects were classified as one of four types A, B, C, and D. At the same time, each piece of defected furniture was identified according to the production shift.

Shift	Type of defect			
	A	B	C	D
1	15	21	45	13
2	26	31	34	5
3	33	17	49	20

Solving the example using the Chi-square test in Spreadsheets

Let's first put this data into the Spreadsheet



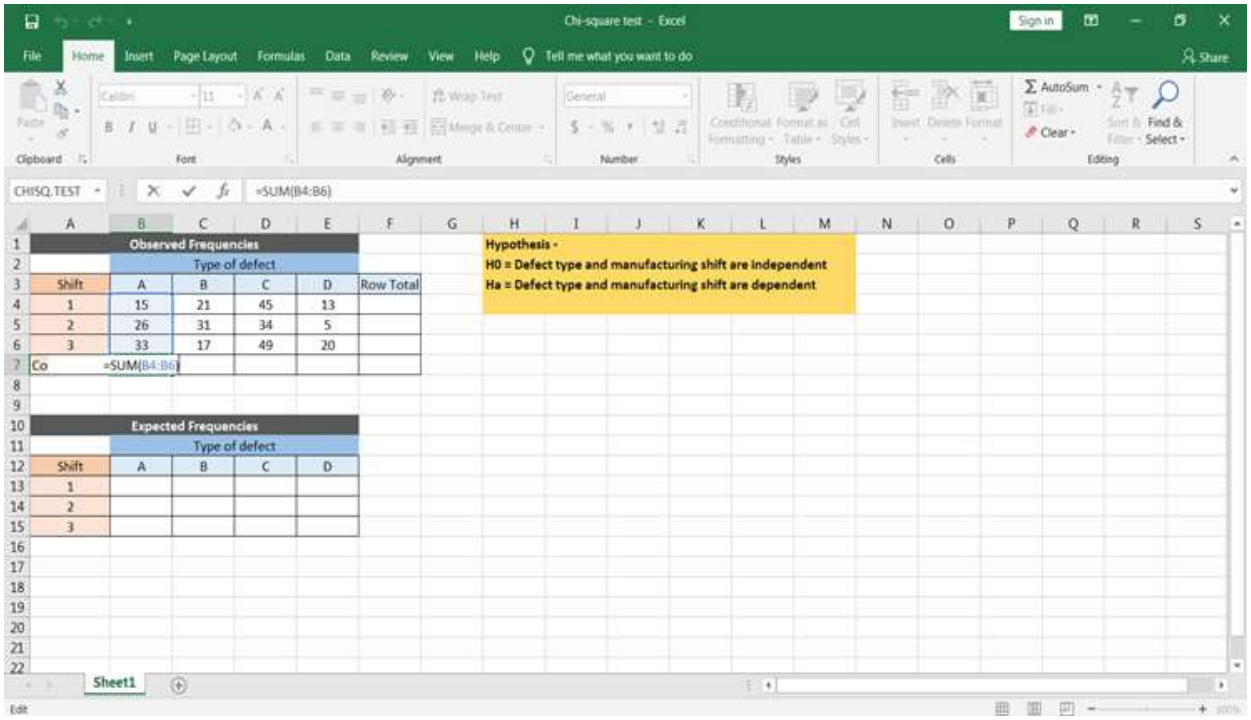
Defining the null hypothesis and the alternative hypothesis

To define the null and alternate hypothesis in the above section. The main objective is to check whether the furniture defects are independent of the production shift or not:

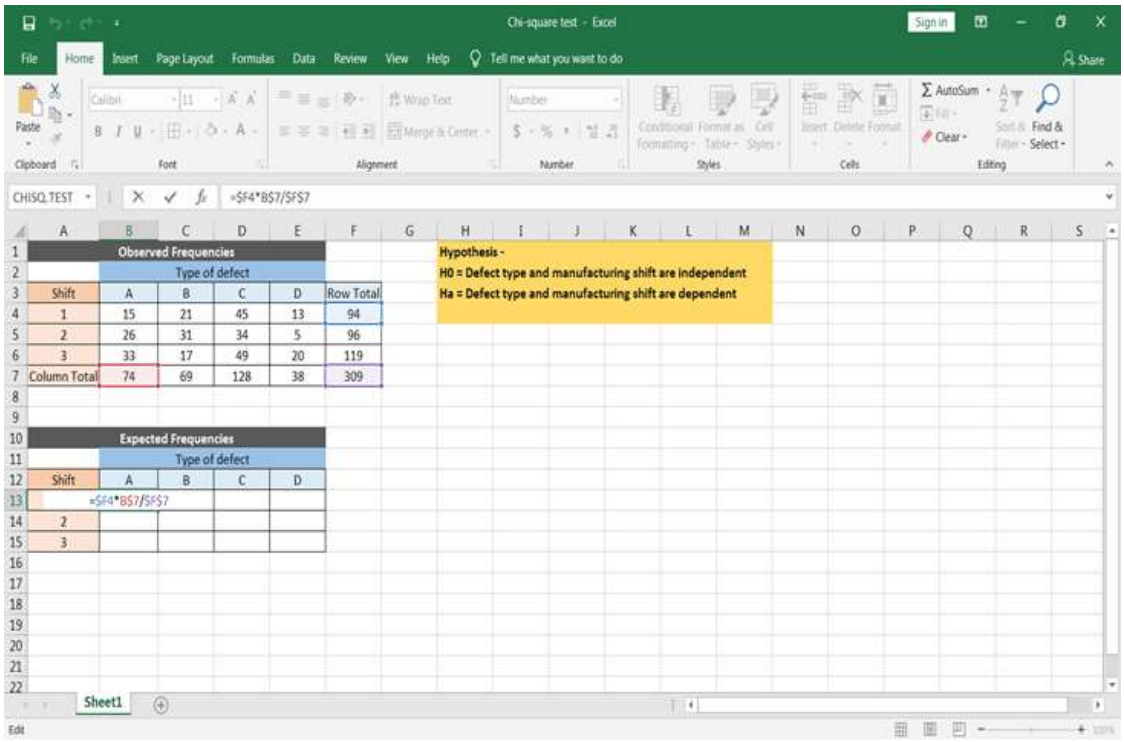
- H0 = Defect type and manufacturing shift are independent
- Ha = Defect type and manufacturing shift are dependent

Calculated expected frequencies

- Before calculating the expected frequencies. First, calculate the row-wise sum of items for each row and the column-wise sum of items for each column using the SUM() function, which is known as row total and column total, respectively. Also, calculate the total of row total and column total. Row and column total both will be the same.



- As you know expected frequency = (row total * column total) / total



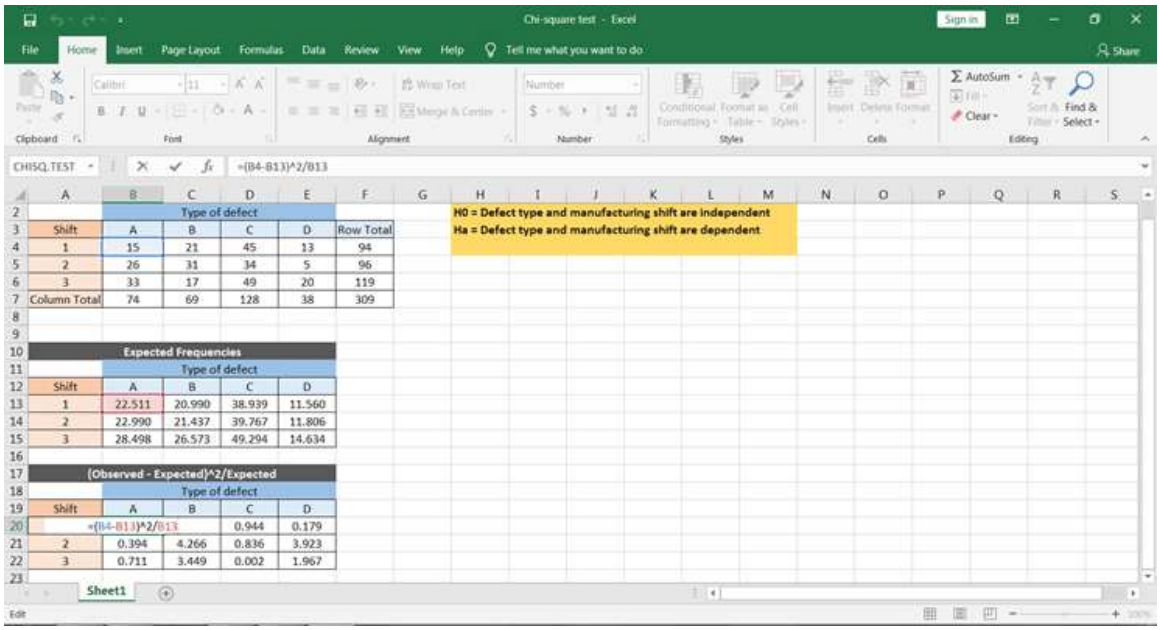
Don't forget to make cells absolute while applying the formula, so that you can copy & paste the formula for all of the expected values.

Calculate Chi-statistic value

Now before you calculate Chi – statistic value or p-value, lets first assume the significance level. This means at what significance level you want to know the answer. Let's assume significance level α = 0.05. Also, the degree of freedom would be = (r-1)*(c-1) = (3-1)*(4-1) = 6.

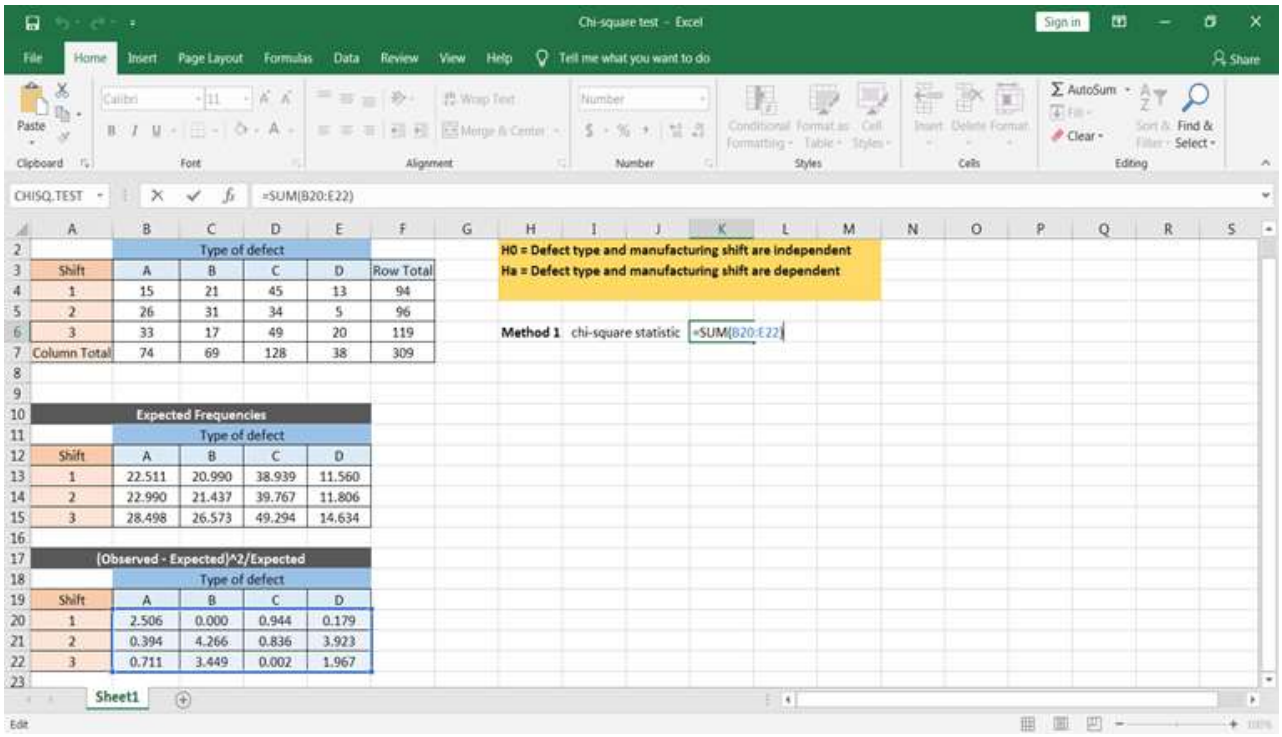
Now there are two ways to calculate chi-statistic value one by the formula $\chi^2 = \sum(O-E)^2/E$ or use the excel function to get the chi-square statistic value.

Let's first calculate using the formula. For this, you need to calculate $\sum(O-E)^2/E$ using excel. This can be done by using the below step –

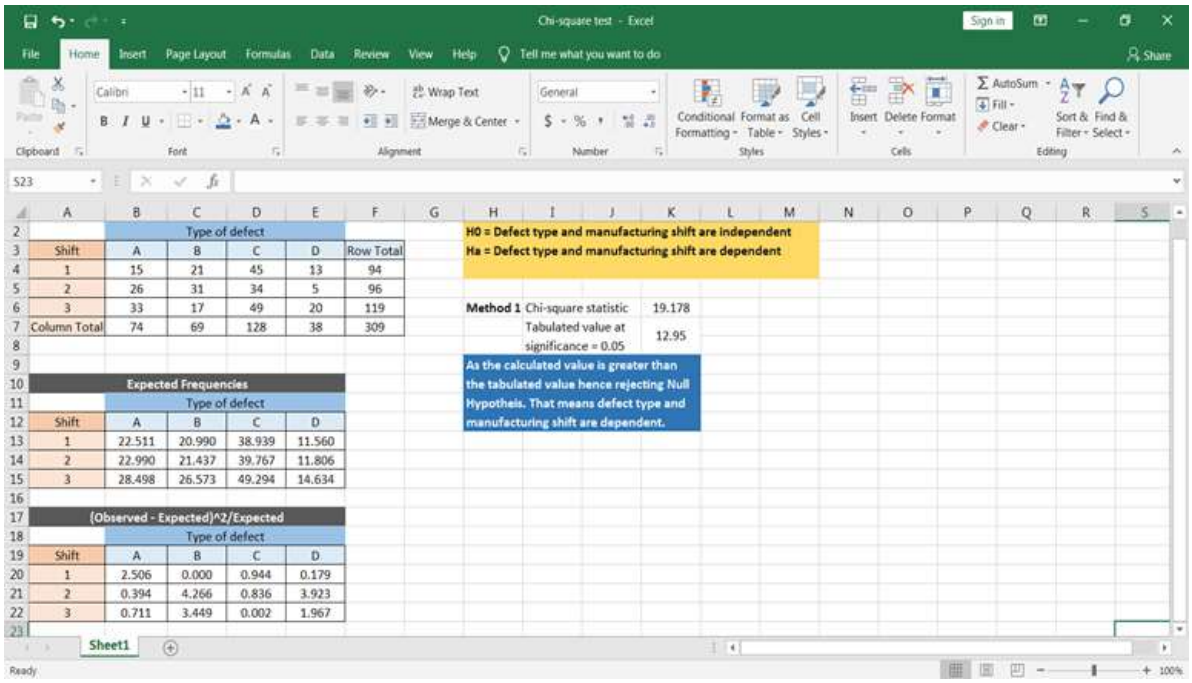


You can get all the values by copy and paste this formula to all the cells.

To get the χ^2 values to take the sum of all the values, this would give us the chi-square statistic calculated value.



Based on the tabulated and calculated value, you can conclude that the defect types and shift times are dependent.



	A	B	C	D	E	F
2		Type of defect				
3	Shift	A	B	C	D	Row Total
4	1	15	21	45	13	94
5	2	26	31	34	5	96
6	3	33	17	49	20	119
7	Column Total	74	69	128	38	309

H0 = Defect type and manufacturing shift are independent
Ha = Defect type and manufacturing shift are dependent

Method 1	Chi-square statistic	19.178
	Tabulated value at significance = 0.05	12.95

As the calculated value is greater than the tabulated value hence rejecting Null Hypotheis. That means defect type and manufacturing shift are dependent.

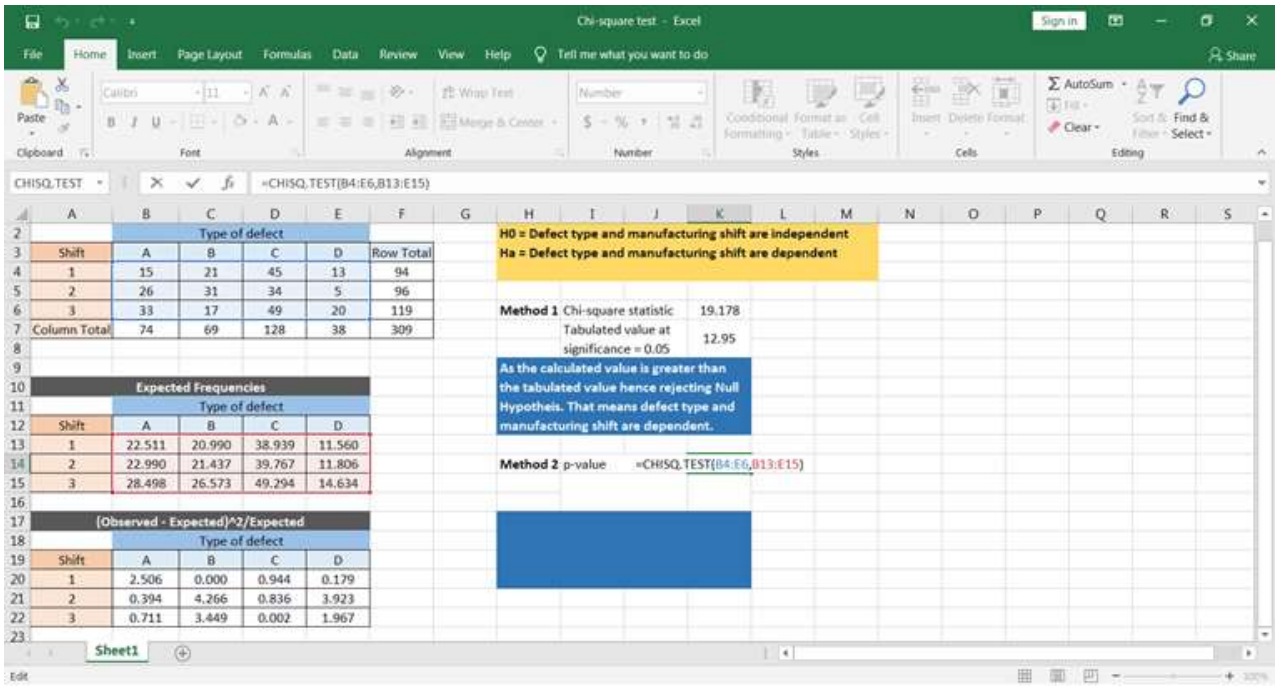
Expected Frequencies

	A	B	C	D
Type of defect				
Shift				
1	22.511	20.990	38.939	11.560
2	22.990	21.437	39.767	11.806
3	28.498	26.573	49.294	14.634

(Observed - Expected)^2/Expected

	A	B	C	D
Type of defect				
Shift				
1	2.506	0.000	0.944	0.179
2	0.394	4.266	0.836	3.923
3	0.711	3.449	0.002	1.967

Now let's calculate using excel function. CHISQ.TEST() function will give the p-value, which can directly be compared with the significance level to conclude the results.



	A	B	C	D	E	F
2		Type of defect				
3	Shift	A	B	C	D	Row Total
4	1	15	21	45	13	94
5	2	26	31	34	5	96
6	3	33	17	49	20	119
7	Column Total	74	69	128	38	309

H0 = Defect type and manufacturing shift are independent
Ha = Defect type and manufacturing shift are dependent

Method 1	Chi-square statistic	19.178
	Tabulated value at significance = 0.05	12.95

As the calculated value is greater than the tabulated value hence rejecting Null Hypotheis. That means defect type and manufacturing shift are dependent.

Method 2 p-value =CHISQ.TEST(B4:E6,B13:E15)

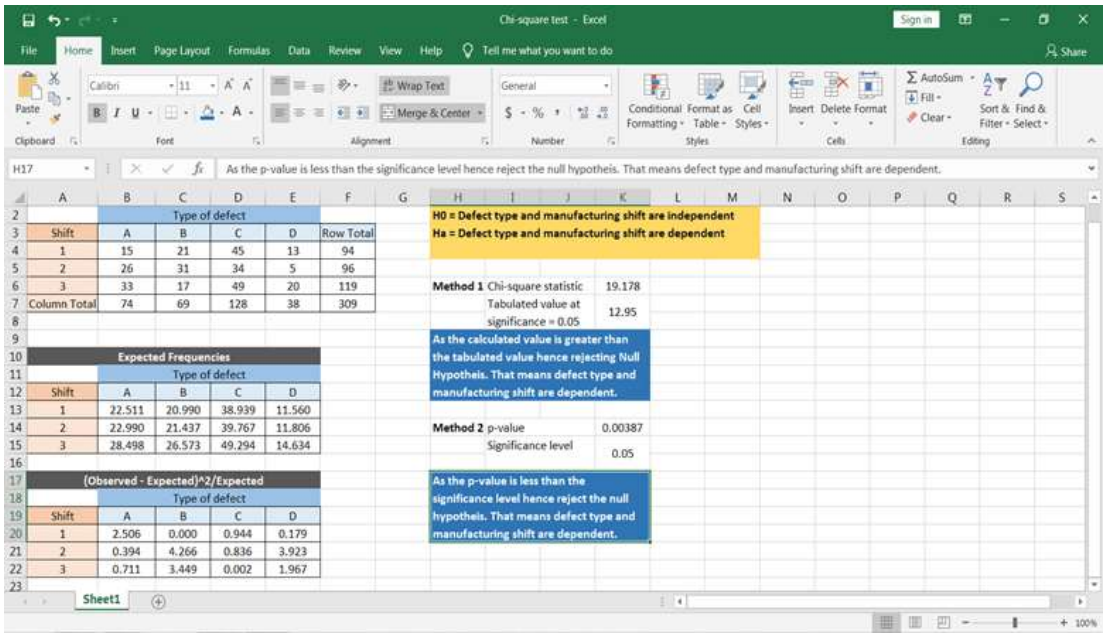
Expected Frequencies

	A	B	C	D
Type of defect				
Shift				
1	22.511	20.990	38.939	11.560
2	22.990	21.437	39.767	11.806
3	28.498	26.573	49.294	14.634

(Observed - Expected)^2/Expected

	A	B	C	D
Type of defect				
Shift				
1	2.506	0.000	0.944	0.179
2	0.394	4.266	0.836	3.923
3	0.711	3.449	0.002	1.967

Based on the p-value, you can conclude that the defect is dependent on manufacturing shift time.



Pros:

- It is easier to compute.
- It can also be used with nominal data.
- It does not assume anything about the data distribution.

Cons:

- The number of observations should be more than 20.
- Data must be frequency data.
- It assumes random sampling. It means the sample should be selected randomly.
- It is sensitive to small frequencies, which leads to erroneous conclusions.
- It is also sensitive to sample size.

Practical No : 5 Perform a Chi square Test in excel

A restaurant manager wants to find the relationship between quality of service and the salary of customers waiting to be served.

She organizes the task in the following way:

- A random sample of 100 customers is considered.
- Every customer is asked to rate the service of the restaurant as “excellent,” “good,” and “poor.”

She constructs the following hypothesis:

- **Null hypothesis** (H_0)–The quality of service is not dependent on the salary of customers waiting to be served.

- Alternative hypothesis (H_1)–The quality of service is dependent on the salary of customers waiting to be served.

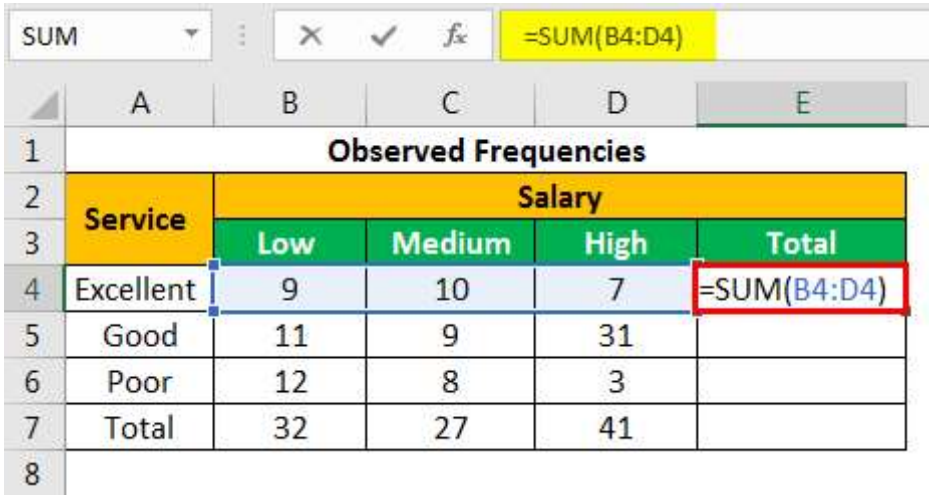
  The manager divides the customers into three categories based on their salaries–"low," "medium," and "high." The level of significance (α) is 0.05.

  The findings are presented as nine data points shown in the following table.

	A	B	C	D	E
1	**Observed Frequencies**				
2	**Service**	**Salary**			
3		**Low**	**Medium**	**High**	**Total**
4	Excellent	9	10	7	
5	Good	11	9	31	
6	Poor	12	8	3	
7	Total	32	27	41	
8					

Let us calculate the sum of all the rows and columns. We apply the following SUM formula to add the numbers of the fourth row.

"=SUM(B4:D4)"



Press the "Enter" key and the sum appears in cell E4. The output is 26.

Similarly, we apply the **SUM formula** to the remaining **rows and columns**. There are 27 respondents with medium salary and 51 respondents who rated the service quality as "good."

E4 =SUM(B4:D4)

	A	B	C	D	E
1	Observed Frequencies				
2	Service	Salary			
3		Low	Medium	High	Total
4	Excellent	9	10	7	26
5	Good	11	9	31	51
6	Poor	12	8	3	23
7	Total	32	27	41	100
8					

We apply the formula “(r-1)(c-1)” to calculate the degrees of freedom (df).

df=(3-1)(3-1)=2*2=4

We apply the following formula to calculate the expected frequency for column B and row 4.

“(=B7*E4/B9)”

The calculation is shown in the following image.

SUM =B7*E4/B9

	A	B	C	D	E
1	Observed Frequencies				
2	Service	Salary			
3		Low	Medium	High	Total
4	Excellent	9	10	7	26
5	Good	11	9	31	51
6	Poor	12	8	3	23
7	Total	32	27	41	100
8					
9	N	100			
10					
11	Expected Frequencies(Variables Perfectly Independent)				
12	Service	Salary			
13		Low	Medium	High	
14	Excellent	=B7*E4/B9			
15	Good				
16	Poor				
17					

The expected number of customers who have “low” salary but rated the restaurant service as “excellent” is 8.32.

In the following calculations, E_{11} is the expected frequency of the first row and the first column. E_{12} is the expected frequency of the first row and the second column.

- E_{11}=(26*32)/100=8.32, E_{12}=7.02, E_{13}=10.66
- E_{21}=16.32, E_{22}=13.77, E_{23}=20.91
- E_{31}=7.36, E_{32}=6.21, E_{33}=9.43

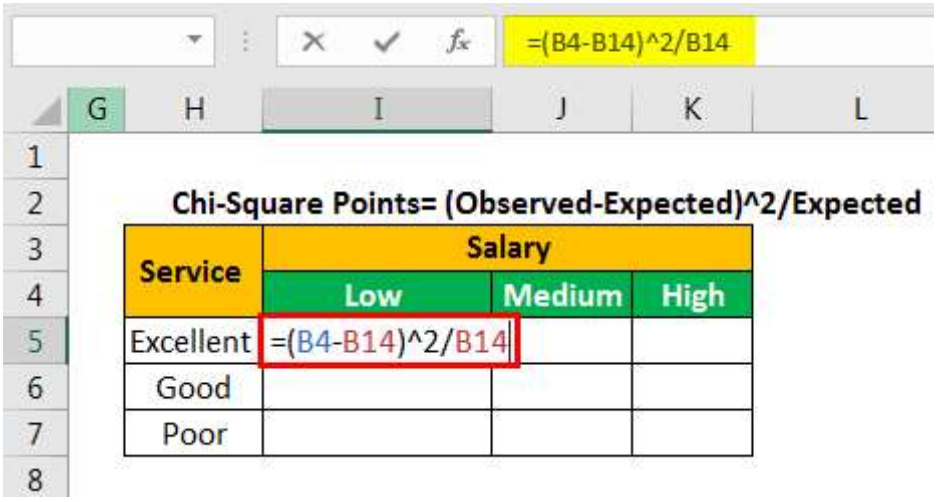
Similarly, we calculate the expected frequencies for the entire table, as shown in the succeeding image.

	A	B	C	D	E
11	**Expected Frequencies(Variables Perfectly Independent)**				
12	**Service**	**Salary**			
13		**Low**	**Medium**	**High**	
14	Excellent	8.32	7.02	10.66	
15	Good	16.32	13.77	20.91	
16	Poor	7.36	6.21	9.43	
17					

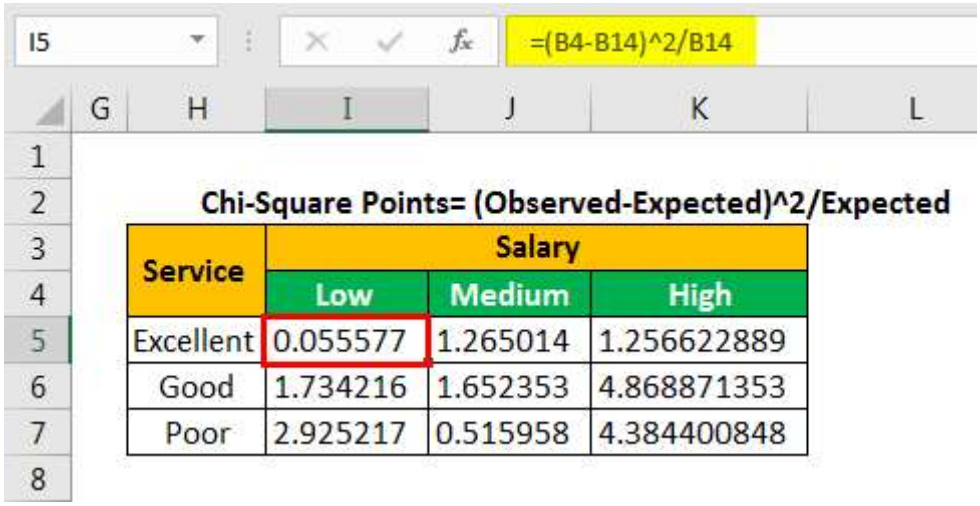
Let us calculate the chi-square data points by using the following formula.

Chi-square points=(observed-expected)^2/expected

We apply the formula "=(B4-B14)^2/B14" to calculate the first chi-square point.

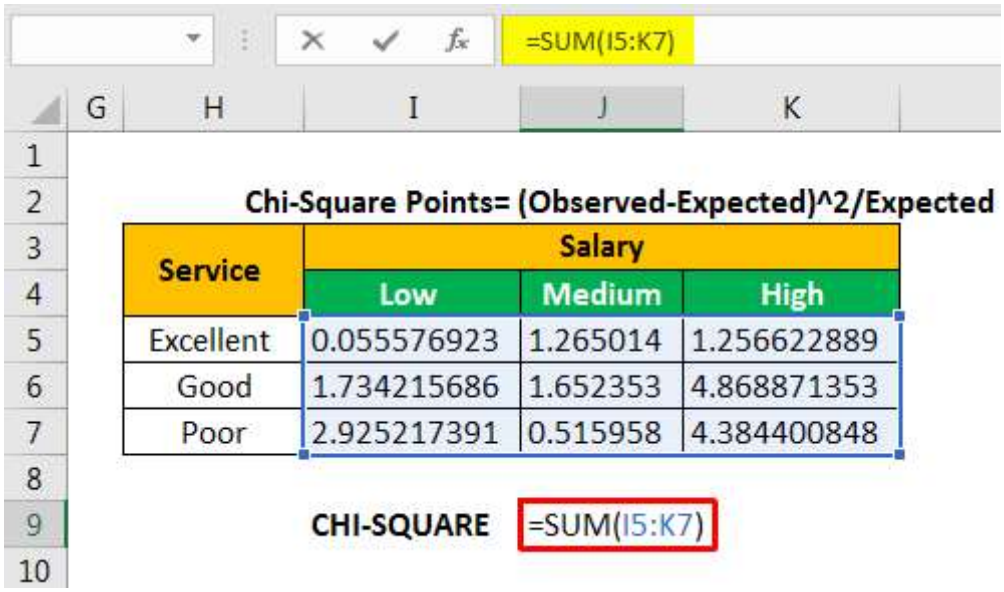


We copy and paste the formula to the remaining cells. This is done to fill values in the entire table, as shown in the following image.



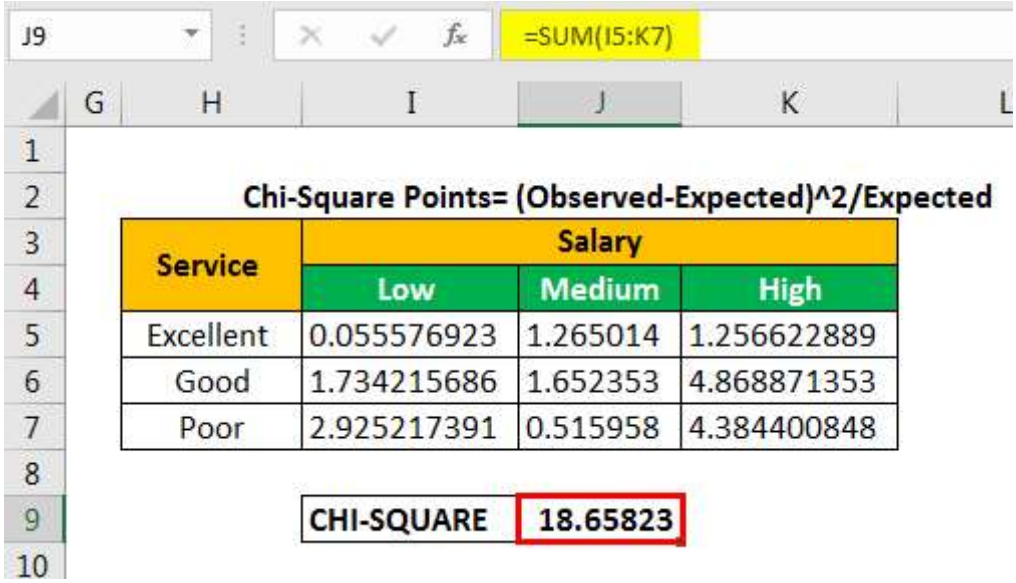
Service	Salary		
	Low	Medium	High
Excellent	0.055577	1.265014	1.256622889
Good	1.734216	1.652353	4.868871353
Poor	2.925217	0.515958	4.384400848

Let us calculate the chi-square calculated value by adding all the values given in the succeeding table.



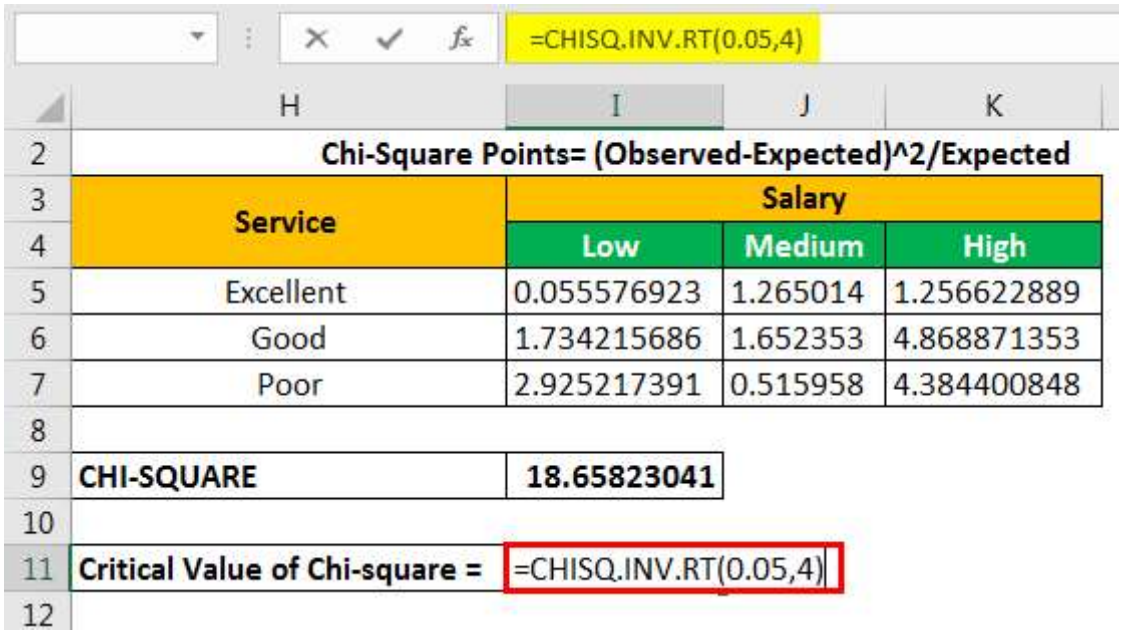
Service	Salary		
	Low	Medium	High
Excellent	0.055576923	1.265014	1.256622889
Good	1.734215686	1.652353	4.868871353
Poor	2.925217391	0.515958	4.384400848

The chi-square calculated value is 18.65823.

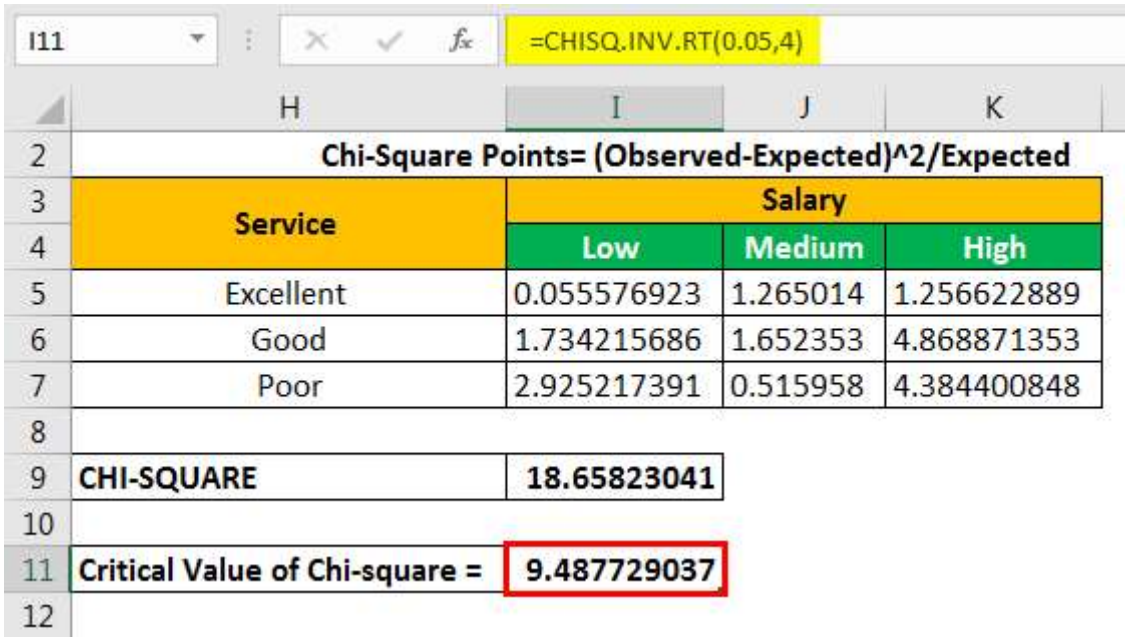


Service	Salary		
	Low	Medium	High
Excellent	0.055576923	1.265014	1.256622889
Good	1.734215686	1.652353	4.868871353
Poor	2.925217391	0.515958	4.384400848

To calculate the critical value, we use either the chi-square critical value table or the CHISQ formula. The formula "CHISQ.INV.RT" contains two parameters–the probability and the **degrees of freedom**. The probability is 0.05, which is a significant value. The df is equal to 4.



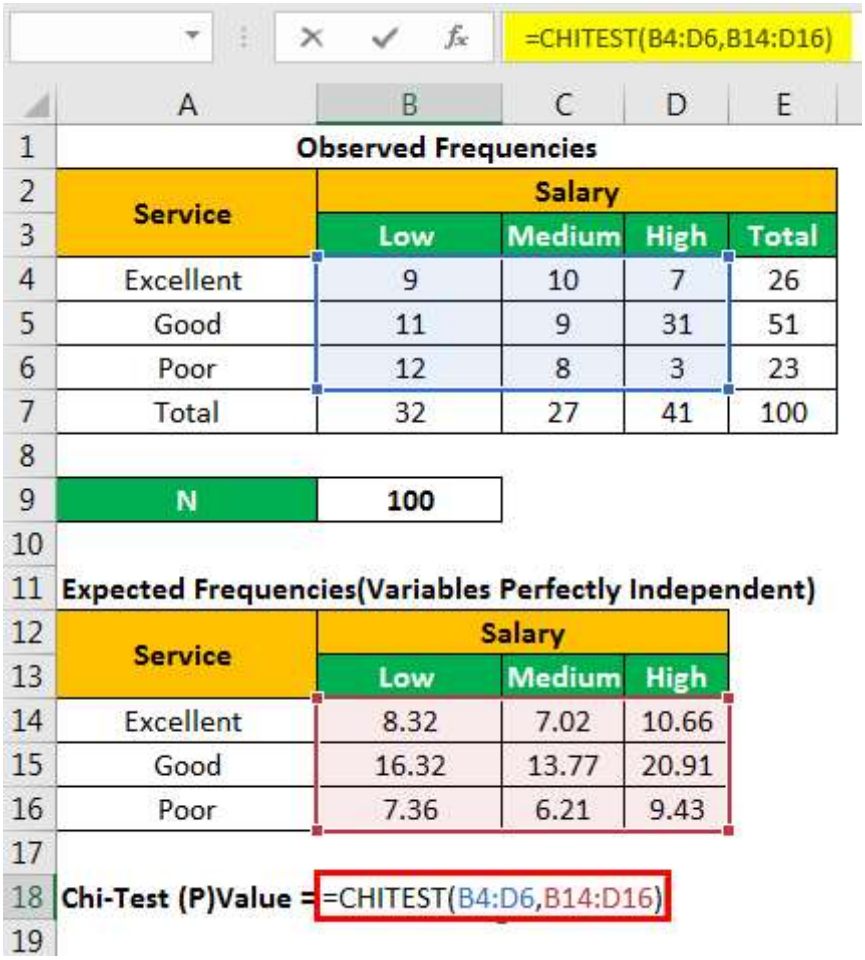
The chi-square critical value is 9.487729037.



Let us find the chi-square p-value with the help of the following formula.

"=CHITEST(actual_range,expected_range)"

We apply the formula "=CHITEST(B4:D6,B14:D16)."



The chi-square p-value is= 0.00091723.

B18 =CHITEST(B4:D6,B14:D16)

	A	B	C	D	E
1	**Observed Frequencies**				
2	**Service**	**Salary**			
3		**Low**	**Medium**	**High**	**Total**
4	Excellent	9	10	7	26
5	Good	11	9	31	51
6	Poor	12	8	3	23
7	Total	32	27	41	100
8					
9	**N**	**100**			
10					
11	**Expected Frequencies(Variables Perfectly Independent)**				
12	**Service**	**Salary**			
13		**Low**	**Medium**	**High**	
14	Excellent	8.32	7.02	10.66	
15	Good	16.32	13.77	20.91	
16	Poor	7.36	6.21	9.43	
17					
18	**Chi-Test (P)Value =**	**0.00091723**			
19					

The chi-square calculated value is significant when equal to or more than the chi-square critical value (tabulated value). The null hypothesis (H_0) is rejected if the chi-square calculated value is greater than the chi-square critical value.

Here x^2 (calculated)>x^2 (tabulated) or 18.65>9.48. Hence, we reject the null hypothesis and accept the alternative hypothesis.

The p-value can also determine whether the null hypothesis must be accepted or rejected. For this, the p-value is compared with alpha (α) in the following way:

- If p-value <= α, the null hypothesis is rejected.
- If p-value > α, the null hypothesis is accepted.

In this example, p-value<α or 0.0009172<0.05. So, we reject H_0 and accept H_1.

We conclude that the quality of service is dependent on the salary of customers waiting to be served.

Practical No : 6 Perform a One way Anova in excel

This example teaches you how to perform a single factor ANOVA (analysis of variance) in Excel. A single factor or one-way ANOVA is used to test the null hypothesis that the means of several populations are all equal.

Below you can find the salaries of people who have a degree in economics, medicine or history.

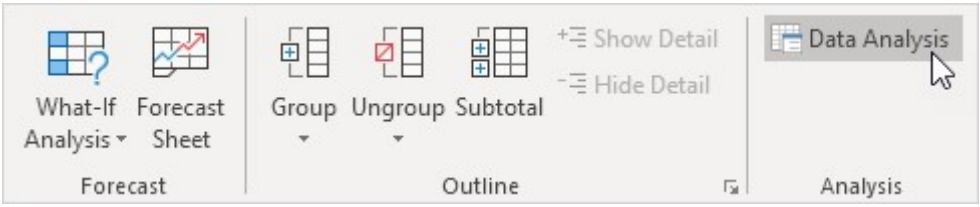
H_0: $\mu_1 =$ $\mu_2 =$ μ_3

H_1: at least one of the means is different.

	A	B	C	D
1	economics	medicine	history	
2	42	69	35	
3	53	54	40	
4	49	58	53	
5	53	64	42	
6	43	64	50	
7	44	55	39	
8	45	56	55	
9	52		39	
10	54		40	
11				

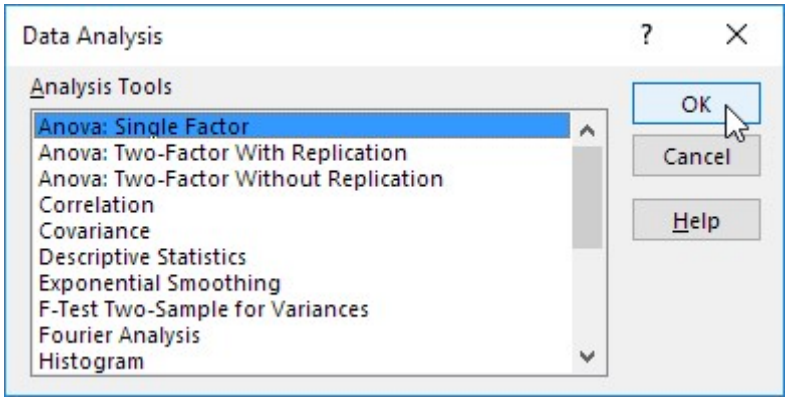
To perform a single factor ANOVA, execute the following steps.

1. On the Data tab, in the Analysis group, click Data Analysis.



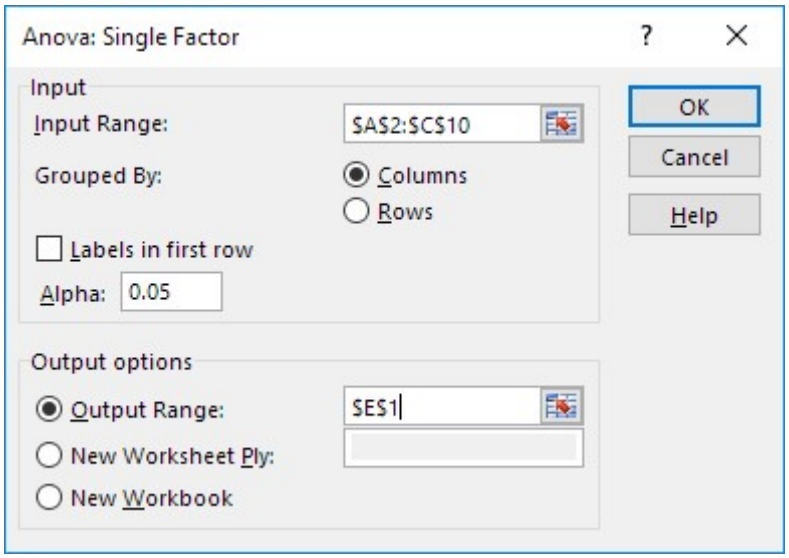
Note: can't find the Data Analysis button? Click here to load the Analysis ToolPak add-in.

2. Select Anova: Single Factor and click OK.



3. Click in the Input Range box and select the range A2:C10.

4. Click in the Output Range box and select cell E1.



5. Click OK.

Result:

E	F	G	H	I	J	K
Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Column 1	9	435	48.33333	23.5		
Column 2	7	420	60	32.33333		
Column 3	9	393	43.66667	50.5		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1085.84	2	542.92	15.19623	7.16E-05	3.443357
Within Groups	786	22	35.72727			
Total	1871.84	24				

Conclusion: if F > F crit, we reject the null hypothesis. This is the case, 15.196 > 3.443. Therefore, we reject the null hypothesis. The means of the three populations are not all equal. At least one of the means is different. However, the ANOVA does not tell you where the difference lies. You need a t-Test to test each pair of means.

Practical No : 7 Perform a One way and Two way Anova in excel

Data arrangement for one-way ANOVA in Excel

If you've been using Excel for a long time, you've gotten used to the idea that the spreadsheet is cell-based. That is, there's very little difference between putting numbers in the spreadsheet in rows or in columns.

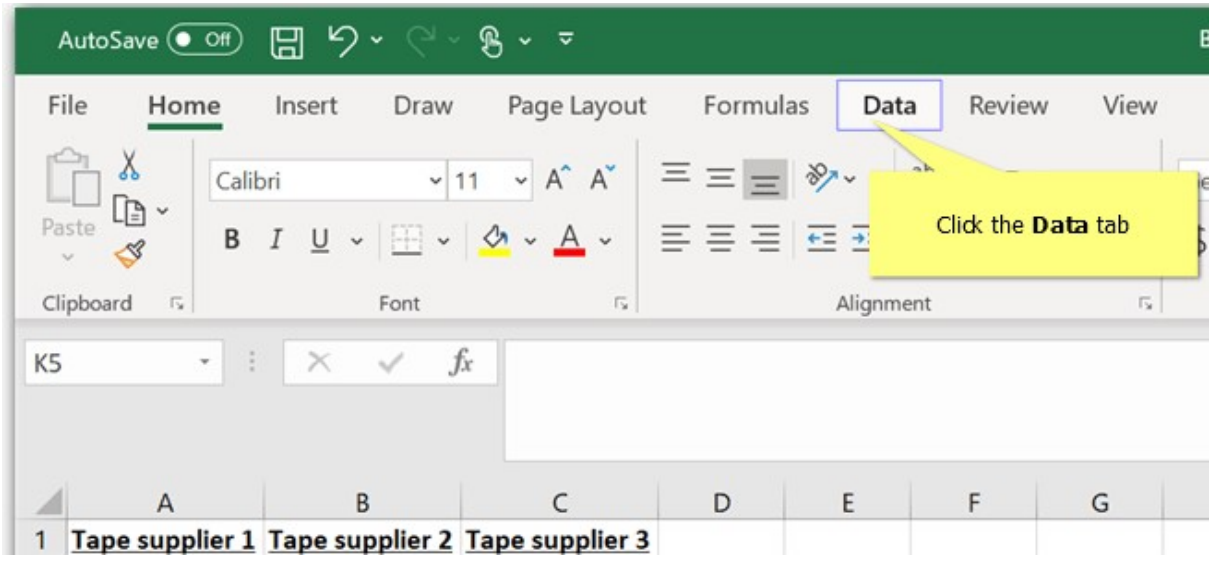
Data in columns:

	A	B	C
1	Tape supplier 1	Tape supplier 2	Tape supplier 3
2	9.5	9.8	10.0
3	10.5	9.8	9.6
4	9.6	9.6	9.9
5	10.1	9.1	9.9
6	9.4	9.5	9.1
7	9.6	9.2	9.6
8	8.9	9.4	10.1
9	10.0	9.9	10.3

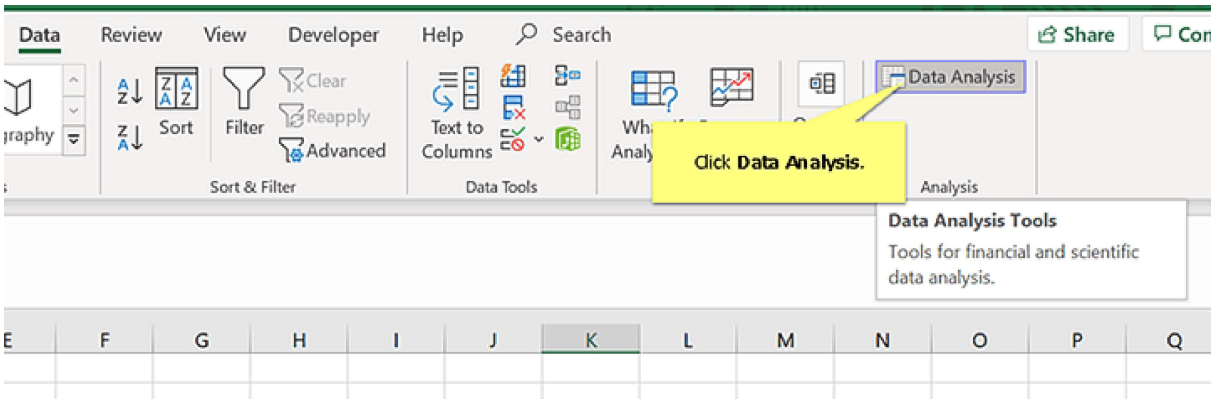
How to use one-way ANOVA in Excel

With the Data Analysis Toolpak installed and your data in columns, you can perform the following steps in Excel to get the results of the one-way ANOVA analysis.

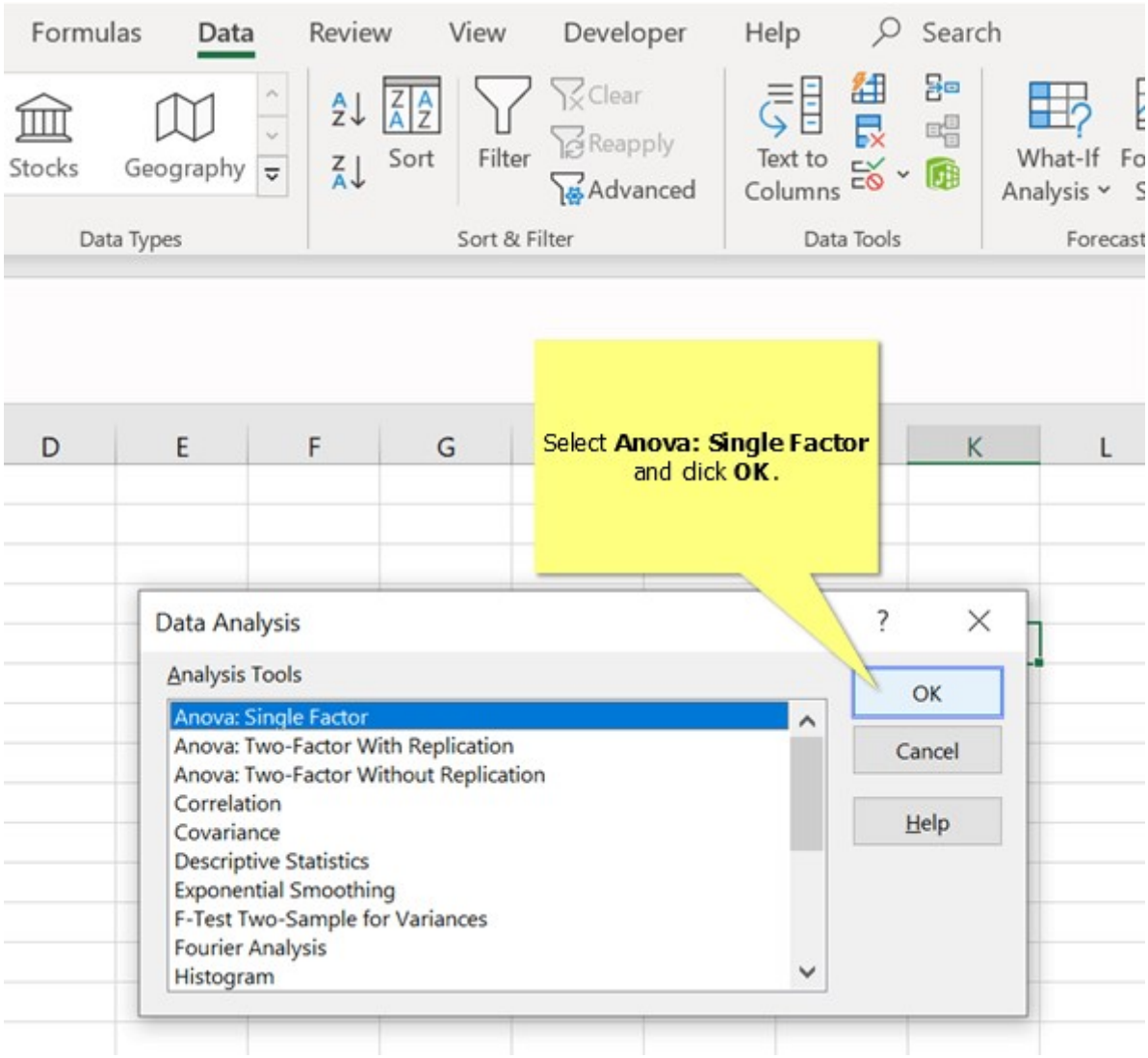
1. Click the **Data** tab



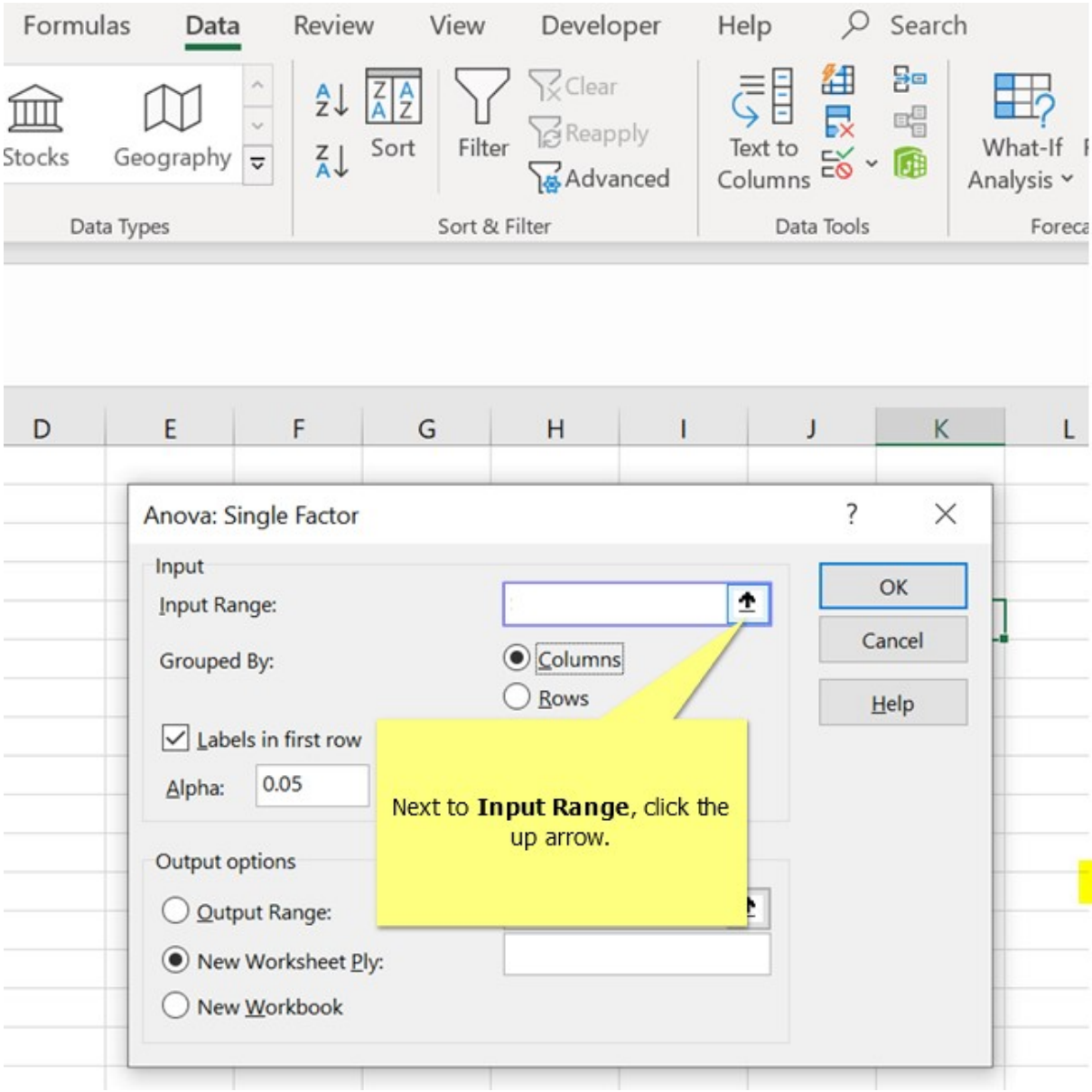
2. Click **Data Analysis**



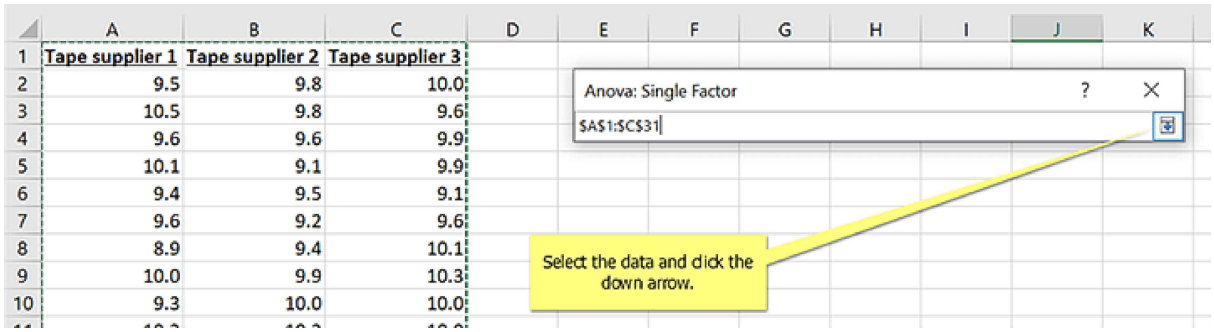
3. Select **Anova: Single Factor** and click **OK**



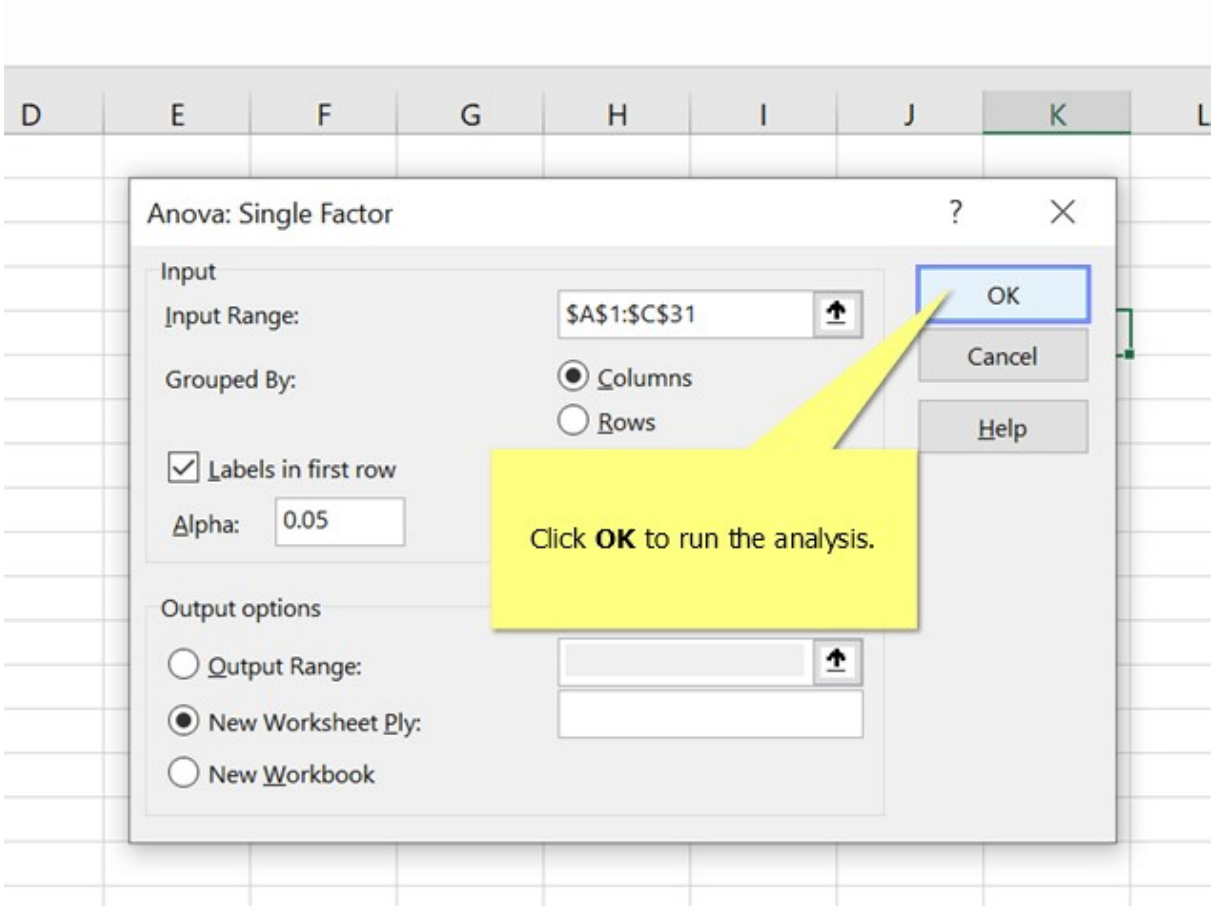
4. Next to **Input Range** click the **up arrow**



5. Select the data and click the **down arrow**



6. Click **OK** to run the analysis



Results for one-way ANOVA in Excel: Summary statistics

The results will look like this

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	Tape supplier 1	30	293.1303	9.771009	0.265686		
6	Tape supplier 2	30	290.2275	9.674249	0.125547		
7	Tape supplier 3	30	295.4848	9.849493	0.126999		
8							
9							
10	ANOVA						
11	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
12	Between Groups	0.462329	2	0.231165	1.33819	0.267665	3.101296
13	Within Groups	15.02874	87	0.172744			
14							
15	Total	15.49107	89				
16							

First, let's take a minute to look at the summary statistics of each group.

In particular, the averages, in ascending order, are about 9.67, 9.77, and 9.84. That is, each of the tapes holds almost 10 kg before breaking. The difference between the largest mean and the smallest mean is about 0.17 kg. If kilograms aren't very familiar to you, you can think of the tape with the lowest average being strong enough to hold about 60 apples and the tape with the highest average being strong enough to hold about 62 apples.

That should be enough for us to start to think about what we expect about the null hypothesis for the ANOVA. If you think that the means are similar, then you'll expect to see a larger p-value for the hypothesis test.

Data Arrangement for Two-Way ANOVA in Excel

Excel can be flexible with your data arrangement for one-way ANOVA, but is strict about the data arrangement when you do a two-way ANOVA with replication through the Data Analysis Toolpak. Data for one factor need to be in different columns.

Data for the second factor need to be in consecutive rows.

For Excel to work, you'll need to have the same number of measurements for all of your groups.

You don't necessarily have to provide the factor label for the rows, but it's good practice, especially if you might want to graph your data in Excel later. This data arrangement, called a two-way table, would look like this:

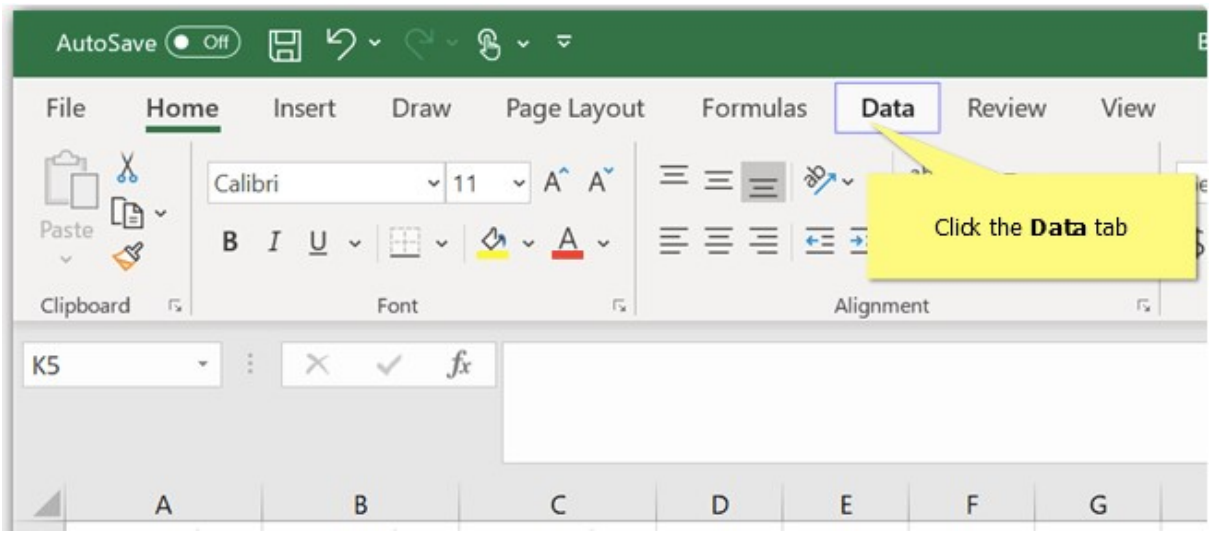
	A	B	C	D
1	Box Type	Tape supplier 1	Tape supplier 2	Tape supplier 3
2	Type 1	9.5	9.8	10.0
3	Type 1	10.5	9.8	9.6
4	Type 1	9.6	9.6	9.9
5	Type 1	10.1	9.1	9.9
6	Type 1	9.4	9.5	9.1
7	Type 1	9.6	9.2	9.6
8	Type 1	8.9	9.4	10.1
9	Type 1	10.0	9.9	10.3
10	Type 1	9.3	10.0	10.0
11	Type 1	10.2	10.2	10.0
12	Type 1	9.3	9.2	9.7
13	Type 1	9.6	9.4	10.0
14	Type 1	11.6	9.4	10.7
15	Type 1	9.7	10.0	9.9
16	Type 2	8.5	8.8	10.3
17	Type 2	9.0	8.6	9.5
18	Type 2	8.6	8.8	9.8
19	Type 2	8.9	8.8	10.1
20	Type 2	8.3	8.5	9.5
21	Type 2	9.0	8.6	9.6
22	Type 2	9.1	9.5	9.7
23	Type 2	8.8	8.8	10.2
24	Type 2	8.2	9.0	9.7
25	Type 2	8.4	7.9	9.3
26	Type 2	8.3	8.2	9.9
27	Type 2	9.4	9.1	9.4
28	Type 2	8.6	8.9	10.3
29	Type 2	8.6	8.9	9.4
30	Type 2	9.2	8.7	10.0
31	Type 2	9.0	8.3	9.8

How to use two-way ANOVA in Excel

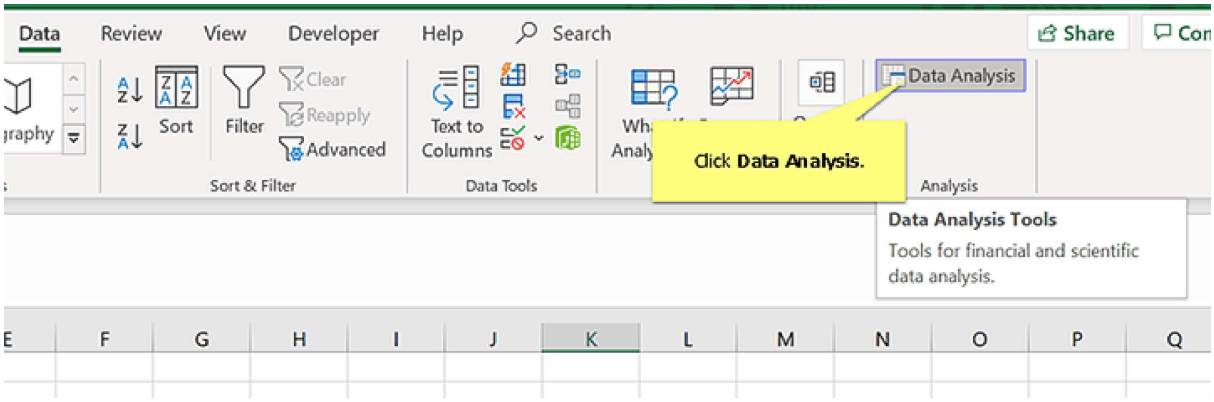
With the Data Analysis Toolpak installed and your data in columns, you can perform the following steps in Excel to get the results of the two-way ANOVA analysis. You'll begin as you did for one-way ANOVA.

Follow along with the two-way ANOVA steps

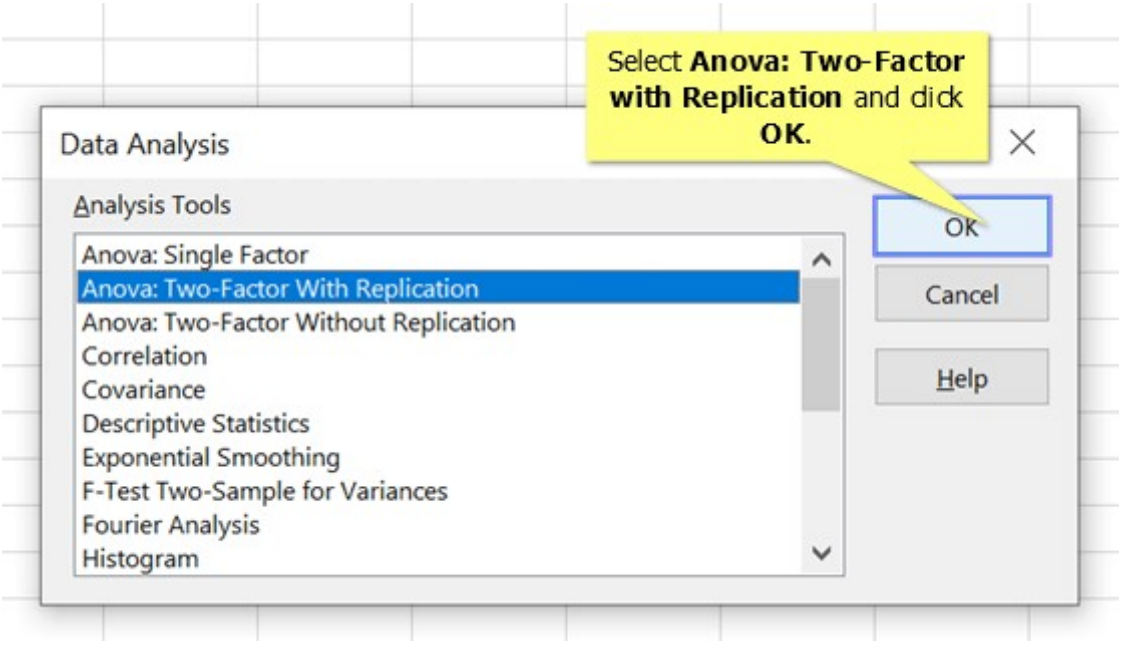
1. Click the **Data** tab



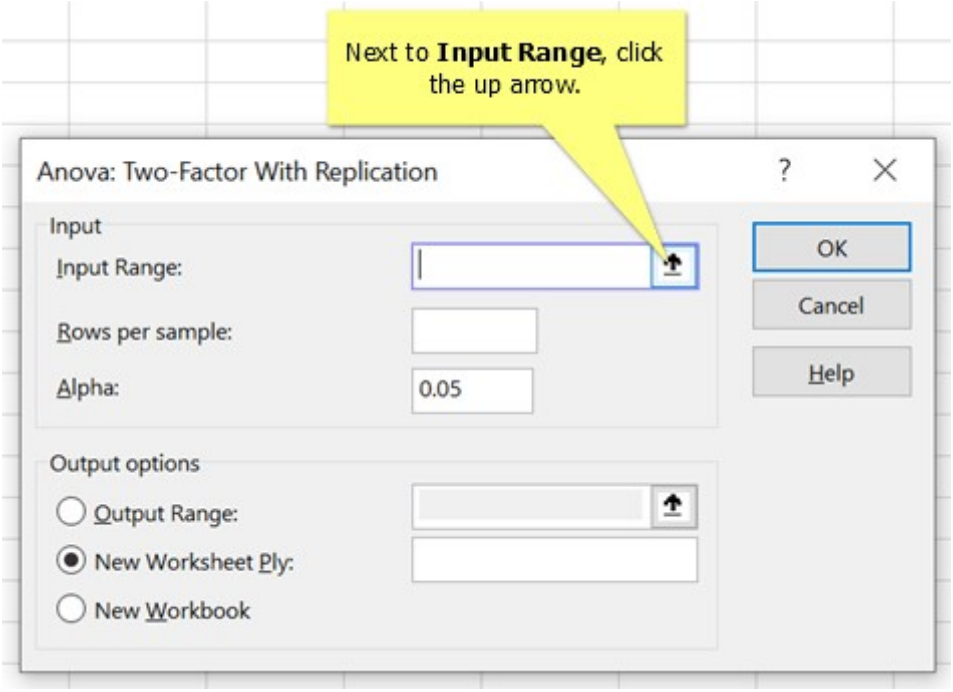
2. Click **Data Analysis**



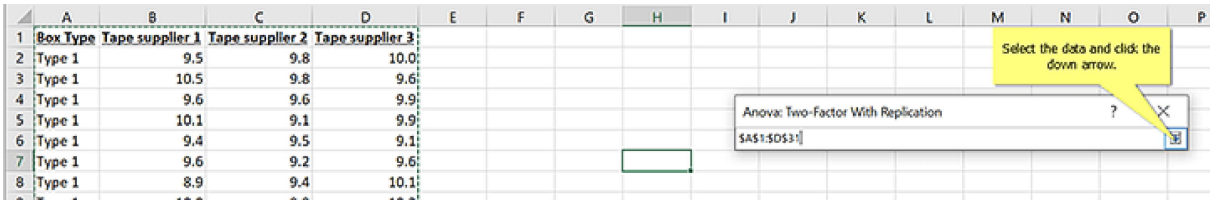
3. Select **Anova: Two Factor with Replication** and click **OK**



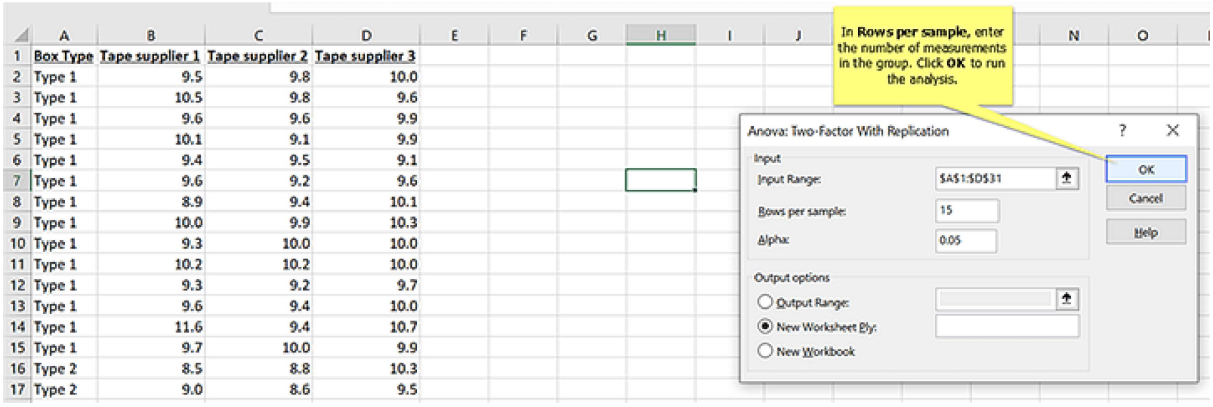
4. Next to **Input Range**, click the **up arrow**



5. Select the data and click the **down arrow**



6. In **Rows per sample**, enter the number of measurements in the group, then click **OK** to run



In this data, you can see that rows 2 to 15 have the measurements for the first box type. Those rows have 15 data points. Since the groups all have to have the same amount of data for the analysis to work in Excel, we know that the second box type must also have 15 rows.

Results for two-way ANOVA in Excel: Summary statistics

As with one-way ANOVA, your results will come in two parts. The first part will be summary statistics about your groups. I've added the highlighting.

	A	B	C	D	E
1	Anova: Two-Factor With Replication				
2					
3	SUMMARY	Tape supplier 1	Tape supplier 2	Tape supplier 3	Total
4	*Type 1*				
5	Count	15	15	15	45
6	Sum	145.7547785	143.6043962	149.1854396	438.5446
7	Average	9.716985235	9.573626414	9.945695974	9.745436
8	Variance	0.533654834	0.144809627	0.144865911	0.28598
9					
10	*Type 2*				
11	Count	15	15	15	45
12	Sum	131.3754857	130.6230684	146.2993576	408.2979
13	Average	8.758365711	8.708204557	9.753290507	9.073287
14	Variance	0.129259258	0.150741122	0.098371318	0.357277
15					
16	*Total*				
17	Count	30	30	30	
18	Sum	277.1302642	274.2274646	295.4847972	
19	Average	9.237675473	9.140915486	9.84949324	
20	Variance	0.557687335	0.336374928	0.126998972	
21					

The blue highlighting shows the overall averages for the two different box types in the data. The difference is about 0.67 kilograms. The gray highlighting shows the averages for the 3 different tape suppliers. The averages for tape supplier 3 is closest to 10, while the averages for tape suppliers 1 and 2 are closer to 9.

The averages for the individual groups have gold highlighting. If the tapes from the different suppliers all work the same on both types of boxes, then the averages for the individual groups should follow the same patterns: The average for box type 1 should be higher and the average for tape supplier 3 should be higher.

The group averages show a different pattern than the overall averages for the two factors. Tape supplier 3's average is higher than the other two because there is a larger difference between the suppliers for the second box type.

This comparison of the averages should prepare us for what to expect about the null hypothesis for two-way ANOVA that the factors do not affect the response variable.

Results for two-way ANOVA in Excel: Hypothesis tests

For our one-way ANOVA analysis, the p-value was relatively large. That value led us to conclude that we couldn't be certain whether there was any difference between the tape suppliers.

For the two-way ANOVA, our largest p-value is about 0.002. That is much smaller than the traditional cutoff value for statistical significance of 0.05.

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	10.16514471	1	10.16514471	50.75373	3.32E-10	3.954568
Columns	8.857659099	2	4.428829549	22.11278	1.93E-08	3.105157
Interaction	2.621802115	2	1.310901057	6.545222	0.002282	3.105157
Within	16.82382898	84	0.200283678			
Total	38.46843491	89				

Because the p-value for the interaction is small, we cannot make a simple statement that one supplier or box type leads to a higher peel strength.

The hypothesis test confirms what we might have expected from the examination of the averages: The effect of the different tapes depends on the box type. (We could equivalently say that the effect of the different box types depends on the tape.)

From the default results in Excel, you can conclude that not all of the groups have the same peel strength. To make a more precise statement about the relationships among the groups, you should proceed to a multiple comparisons analysis.

Practical No : 8 Perform a One way Anova in excel

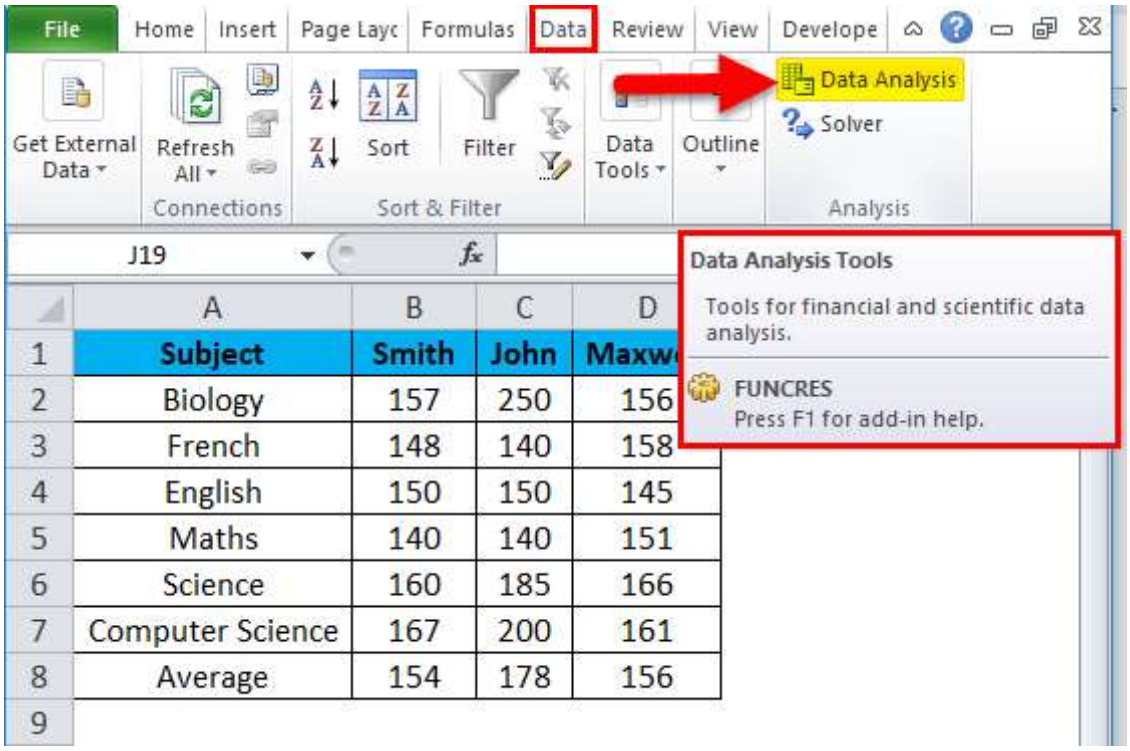
In this example, we are going to see how to apply Excel ANOVA single factor by following the below example.

Consider the below example, which shows students' marks scored in each subject.

	A	B	C	D	E
1	**Subject**	**Smith**	**John**	**Maxwell**	
2	Biology	157	250	156	
3	French	148	140	158	
4	English	150	150	145	
5	Maths	140	140	151	
6	Science	160	185	166	
7	Computer Science	167	200	161	
8	Average	154	178	156	
9					

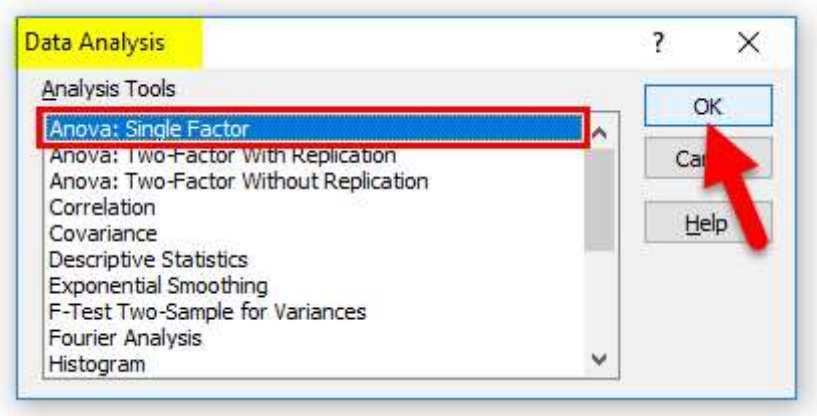
Now we are going to check that students' marks are significantly different by using the ANOVA tool by following the below steps.

- First, Go to the **DATA** menu and then click on **DATA ANALYSIS**.

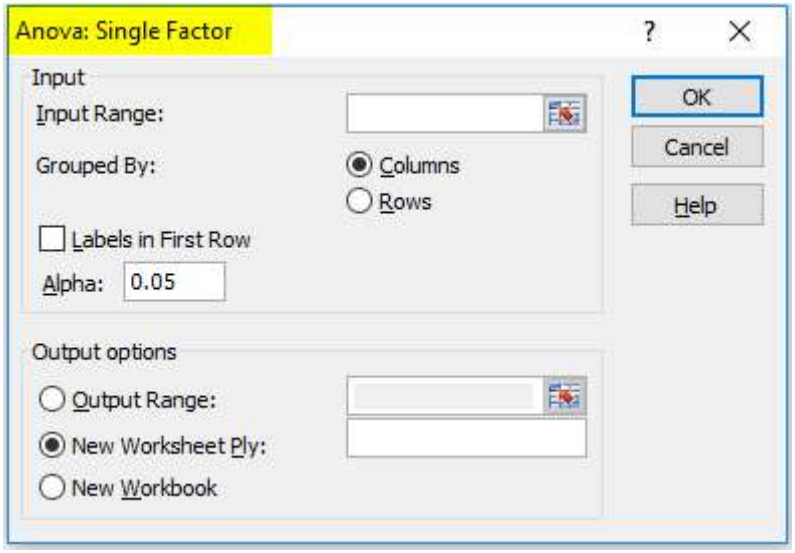


	A	B	C	D
1	Subject	Smith	John	Maxw
2	Biology	157	250	156
3	French	148	140	158
4	English	150	150	145
5	Maths	140	140	151
6	Science	160	185	166
7	Computer Science	167	200	161
8	Average	154	178	156
9				

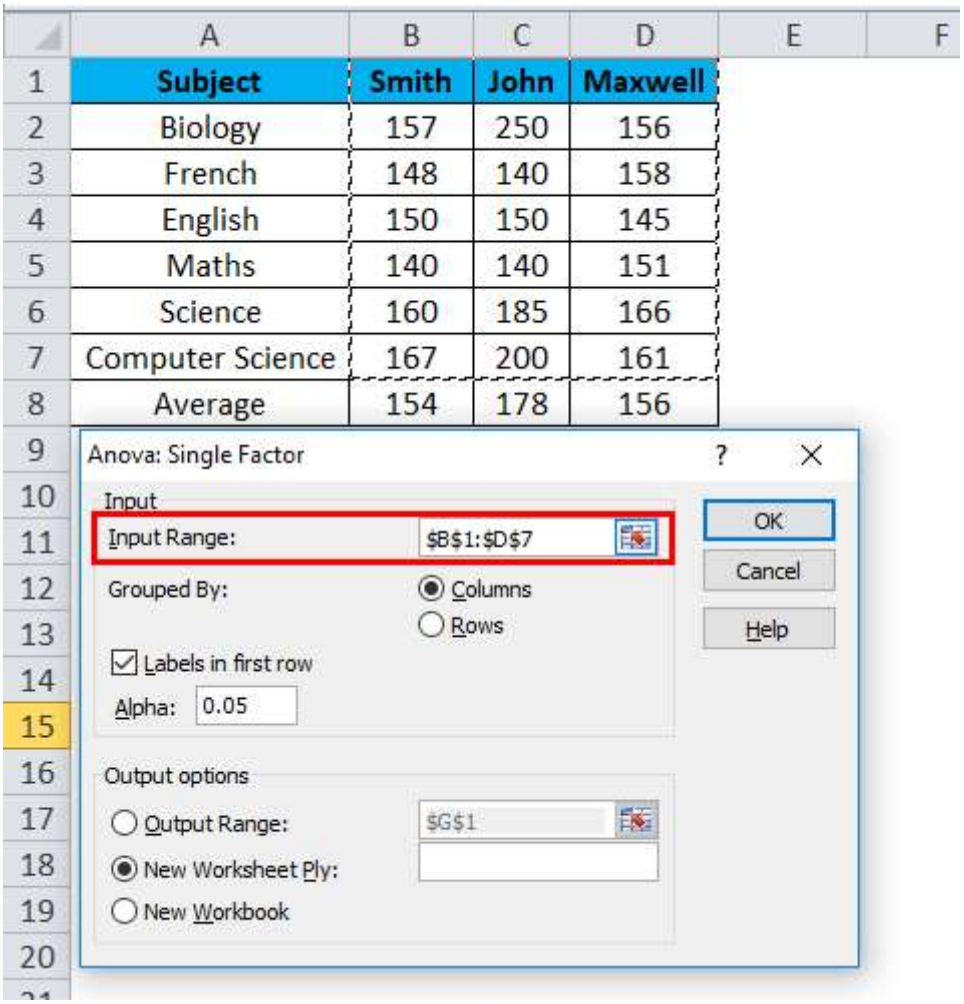
- We will also get the analysis dialogue box.
- In the below screenshot, we can see the list of analysis tools where we can see the ANOVA-Single-factor tool.
- Click on the **ANOVA: Single-factor** tool and then click **OK**.



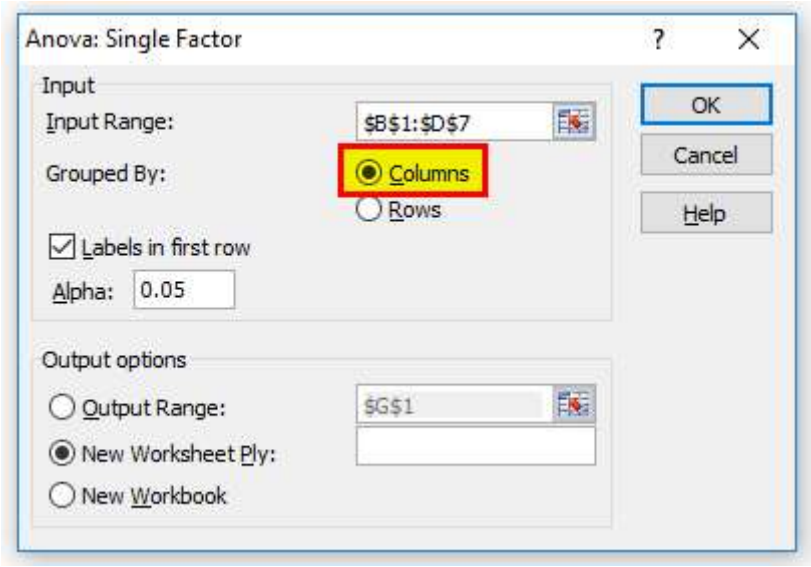
- So that we will get the ANOVA: Single-factor dialogue box as shown in the below screenshot.



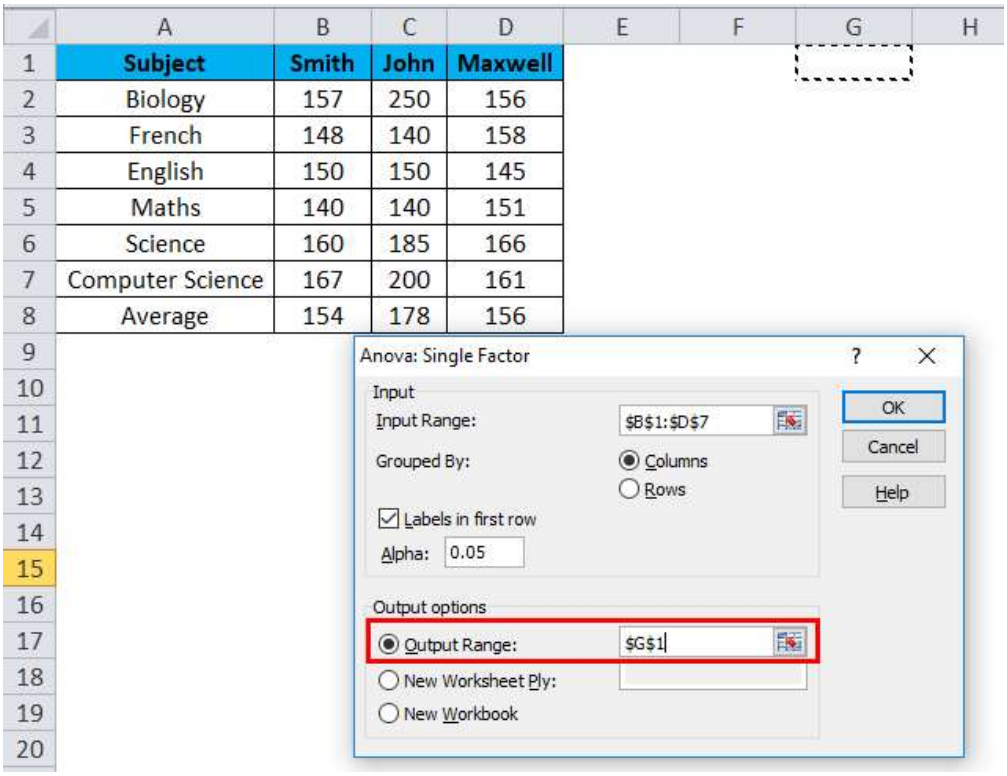
- Now we can see the input range in the dialogue box.
- Click on the input range box to select the range $B$1:$D$7 as shown below.



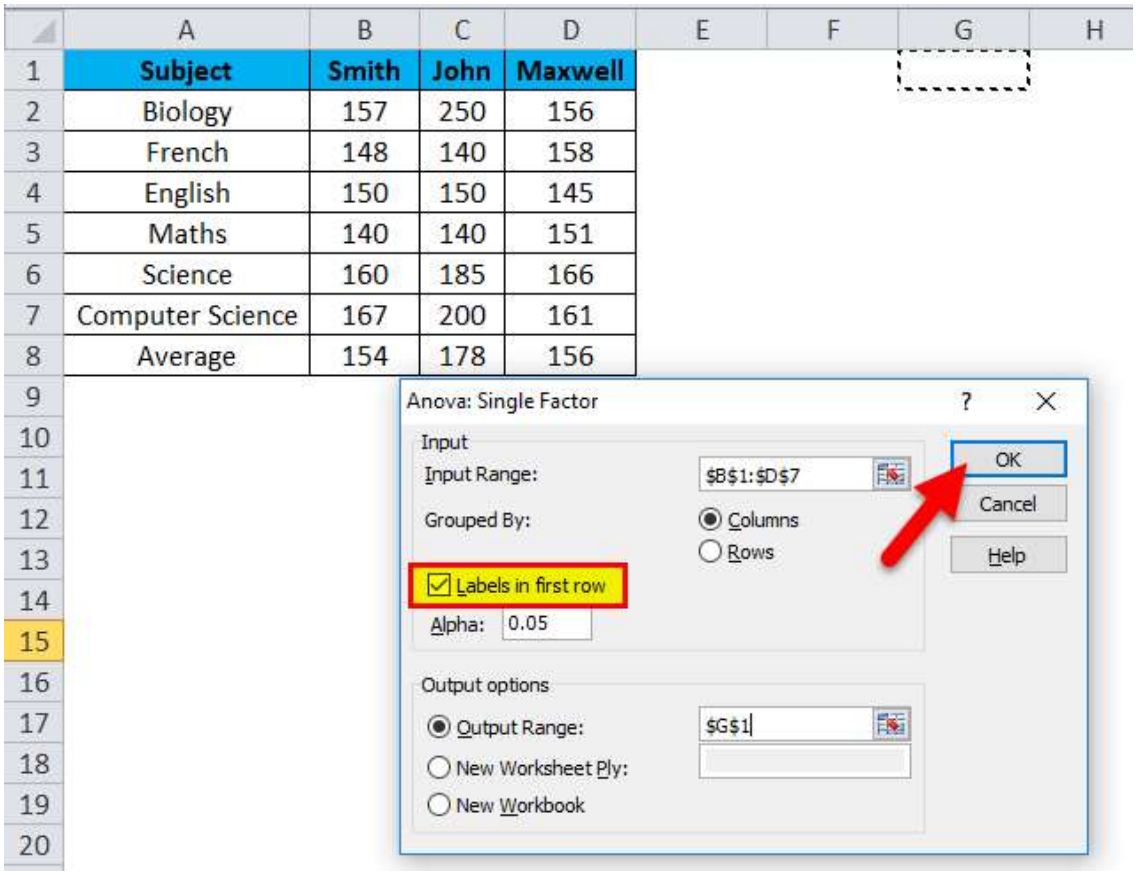
- As we can see in the above screenshot, we have selected ranges along with the student name to get the exact output.
- Now the input range has been selected, make sure that **Column** Checkbox is selected.



- Next step, we want to select the output range where our output needs to be displayed.
- Click on the output range box and select the output cell in the worksheet, as shown below.



- We have the output range cell as G1, where the output is going to be under the display.
- Make sure to select **Labels in the first-row** Checkbox, and then click on **OK**.



- Thus, we will get the below output as follows.

Subject	Smith	John	Maxwell
Biology	157	250	156
French	148	140	158
English	150	150	145
Maths	140	140	151
Science	160	185	166
Computer Science	167	200	161
Average	154	178	156

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Smith	6	922	153.6667	92.26667
John	6	1065	177.5	1877.5
Maxwell	6	937	156.1667	54.96667

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2058.778	2	1029.389	1.525221	0.249474	3.68232
Within Groups	10123.67	15	674.9111			
Total	12182.44	17				

The above screenshot shows the summary part and Anova where the summary part contains the Group Name, No of Count, Sum, Average, and Variance, and the Anova shows a list of summaries where we need to check the **F value** and **F Crit value**.

F= Between Group /Within Group

- **F Statistic:** The F statistic is nothing but values we get when we execute the ANOVA, which is useful to determine the means between two populations significantly.

F values are always used along with the "P" value to check the results are significant, and it is enough to reject the null hypothesis.

If we get the F value greater than the F crit value, then we can reject the null hypothesis, which means that something is significant, but in the above screenshot, we cannot reject the null hypothesis because the F value is smaller than the F critic and the student's marks are not significant which is in highlighted format and shown in the below screenshot.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2058.778	2	1029.389	1.525221	0.249474	3.68232
Within Groups	10123.67	15	674.9111			
Total	12182.44	17				

If we are running an Excel ANOVA single factor, make sure that variance 1 is smaller than variance 2.

In the above screenshot, we can see that the first variance (92.266) is smaller than variance 2 (1877.5).

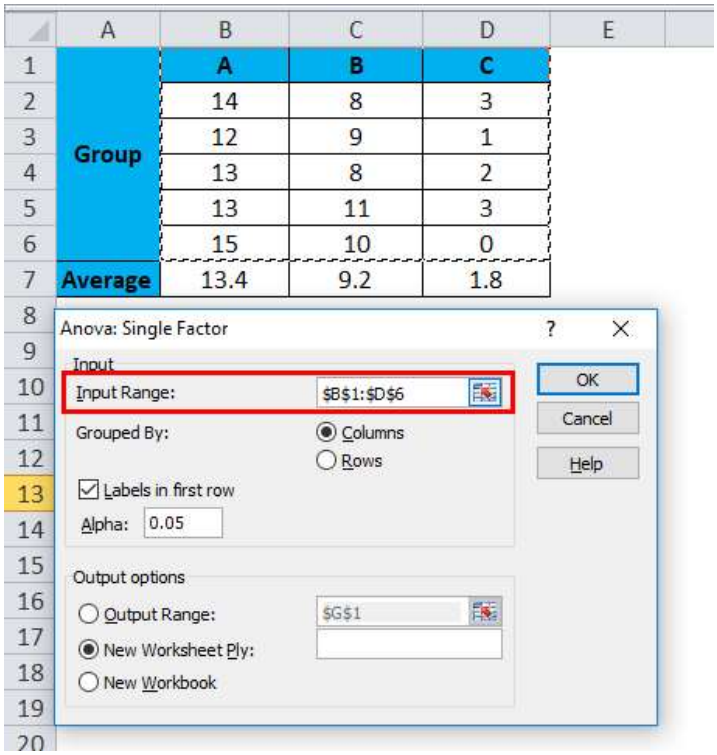
Excel ANOVA – Example #2

In this example, we will see how to reject the null hypothesis by following the below steps.

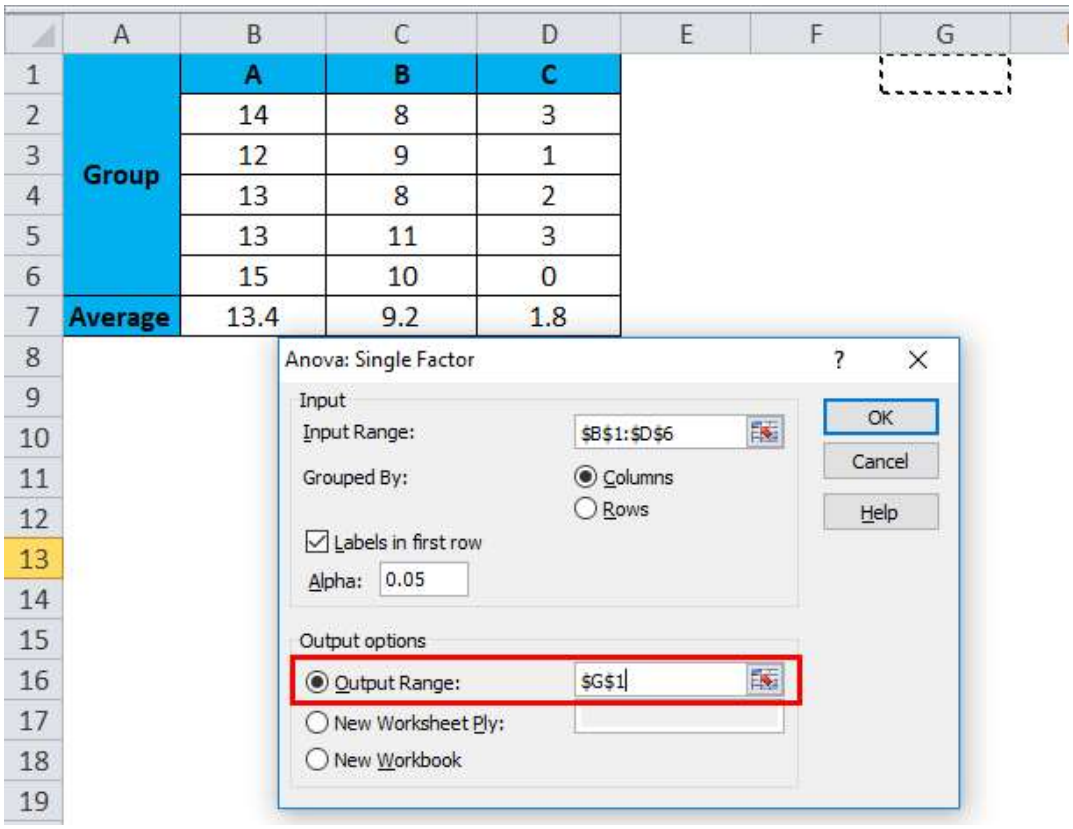
	A	B	C	D	E
1	Group	A	B	C	
2		14	8	3	
3		12	9	1	
4		13	8	2	
5		13	11	3	
6		15	10	0	
7	Average	13.4	9.2	1.8	
8					
9					

In the above screenshot, we can see the three groups A, B, and C, and we are going to determine how these groups are significantly different by running the ANOVA test.

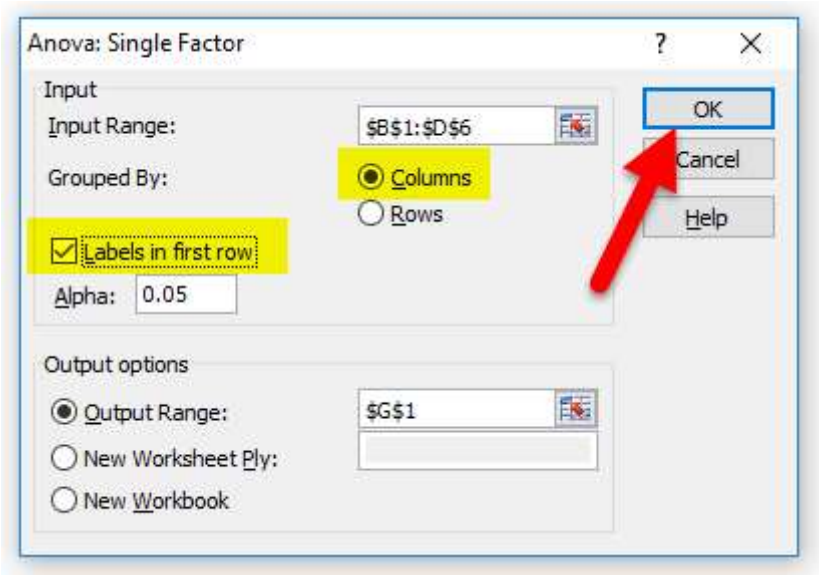
- First, click on the **DATA** menu.
- Click on the **data analysis** tab.
- Choose **Anova Single-factor** from the Analysis dialogue box.
- Now select the **input range** as shown below.



- Next, select the **output range** as G1 to get the output.



- Make sure to select Columns and Labels in the first-row Checkbox, and then click on Ok.



- We will get the below result as shown below.

	A	B	C	D	G	H	I	J	K	L	M
1		**A**	**B**	**C**	Anova: Single Factor						
2		14	8	3							
3	**Group**	12	9	1	SUMMARY						
4		13	8	2	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5		13	11	3	A	5	67	13.4	1.3		
6		15	10	0	B	5	46	9.2	1.7		
7	**Average**	13	9	2	C	5	9	1.8	1.7		
8											
9											
10					ANOVA						
11					*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
12					Between Groups	344.9333	2	172.467	110.1	1.90655E-08	3.9
13					Within Groups	18.8	12	1.56667			
14											
15					Total	363.7333	14				
16											

In the below screenshot, we can see that the F value is greater than the F crit value so we can reject the null hypothesis, and we can also say that at least one of the groups is significantly different.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	344.9333	2	172.467	110.1	1.90655E-08	3.9
Within Groups	18.8	12	1.56667			
Total	363.7333	14				

Practical No : 9 Perform a Wilcoxon Signed Rank Test in Excel.
(Step-by-Step)

Step 1: Create the Data

Suppose an engineer want to know if a new fuel treatment leads to a change in the average miles per gallon of a certain car. To test this, he measures the mpg of 12 cars with and without the fuel treatment.

We'll create the following data in Excel to hold the mpg values for each car with the fuel treatment (group1) and without the fuel treatment (group 2):

	A	B	C	D	E	F
1	**group1**	**group2**				
2	20	24				
3	23	25				
4	21	21				
5	25	22				
6	18	23				
7	17	18				
8	18	17				
9	24	28				
10	20	24				
11	24	27				
12	23	21				
13	19	23				
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						

Step 2: Calculate the Difference Between the Groups

Next, we'll calculate the difference between the groups:

	A	B	C	D	E	F
1	group1	group2	difference			
2	20	24	=A2-B2			
3	23	25	-2			
4	21	21	0			
5	25	22	3			
6	18	23	-5			
7	17	18	-1			
8	18	17	1			
9	24	28	-4			
10	20	24	-4			
11	24	27	-3			
12	23	21	2			
13	19	23	-4			
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						

Step 3: Calculate the Absolute Differences

Next, we'll calculate the absolute difference between the groups, returning a blank if the absolute difference is zero:

	A	B	C	D	E	F
1	group1	group2	difference	abs difference		
2	20	24	-4	=IF(C2=0, "", ABS(C2))		
3	23	25	-2	2		
4	21	21	0			
5	25	22	3	3		
6	18	23	-5	5		
7	17	18	-1	1		
8	18	17	1	1		
9	24	28	-4	4		
10	20	24	-4	4		
11	24	27	-3	3		
12	23	21	2	2		
13	19	23	-4	4		
14						
15						
16						
17						
18						
19						
20						
21						
22						

Step 4: Calculate the Rank of the Absolute Differences

Next, we'll use the **RANK.AVG()** function to calculate the rank of the absolute differences between the groups, returning a blank if the absolute difference is zero:

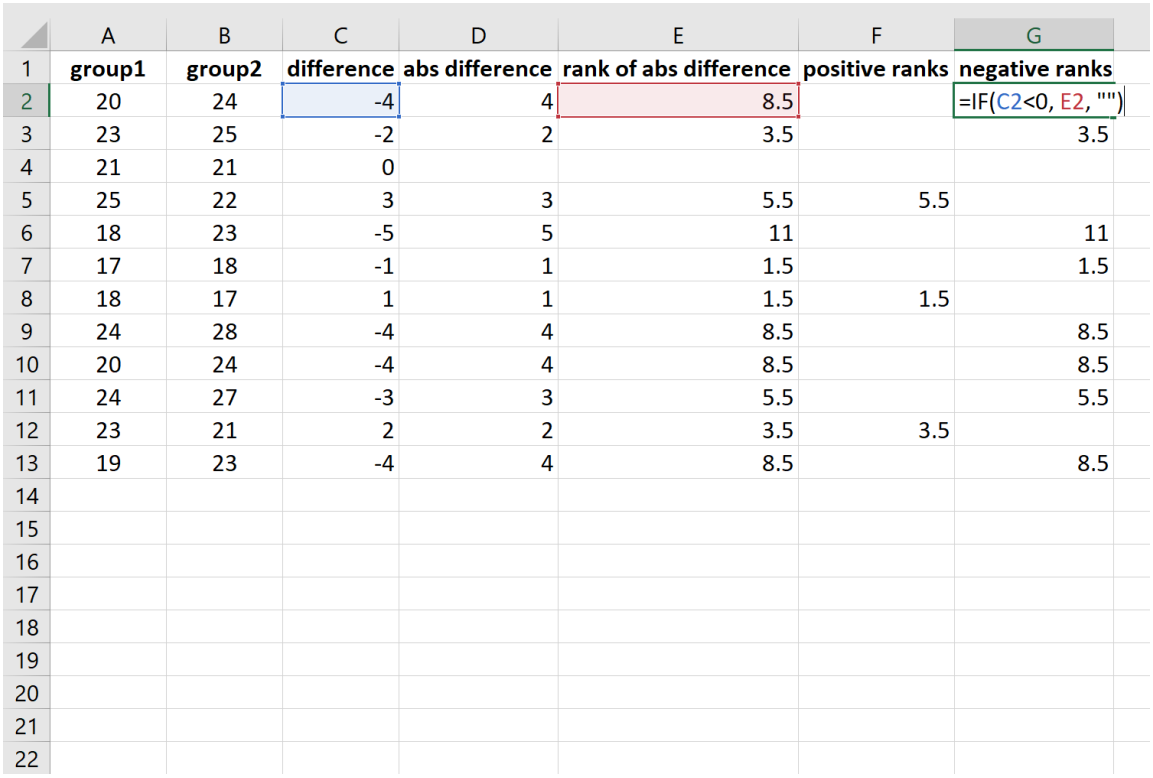
	A	B	C	D	E	F	G
1	**group1**	**group2**	**difference**	**abs difference**	**rank of abs difference**		
2	20	24	-4	4	=IF(C2=0, "", RANK.AVG(D2, $D$2:$D$13, 1))		
3	23	25	-2	2	3.5		
4	21	21	0				
5	25	22	3	3	5.5		
6	18	23	-5	5	11		
7	17	18	-1	1	1.5		
8	18	17	1	1	1.5		
9	24	28	-4	4	8.5		
10	20	24	-4	4	8.5		
11	24	27	-3	3	5.5		
12	23	21	2	2	3.5		
13	19	23	-4	4	8.5		
14							
15							
16							
17							
18							
19							
20							
21							

Step 5: Calculate the Positive & Negative Ranks

Next, we'll calculate the positive ranks:

	A	B	C	D	E	F
1	**group1**	**group2**	**difference**	**abs difference**	**rank of abs difference**	**positive ranks**
2	20	24	-4	4	8.5	=IF(C2>0, E2, "")
3	23	25	-2	2	3.5	
4	21	21	0			
5	25	22	3	3	5.5	5.5
6	18	23	-5	5	11	
7	17	18	-1	1	1.5	
8	18	17	1	1	1.5	1.5
9	24	28	-4	4	8.5	
10	20	24	-4	4	8.5	
11	24	27	-3	3	5.5	
12	23	21	2	2	3.5	3.5
13	19	23	-4	4	8.5	
14						
15						
16						
17						
18						
19						
20						

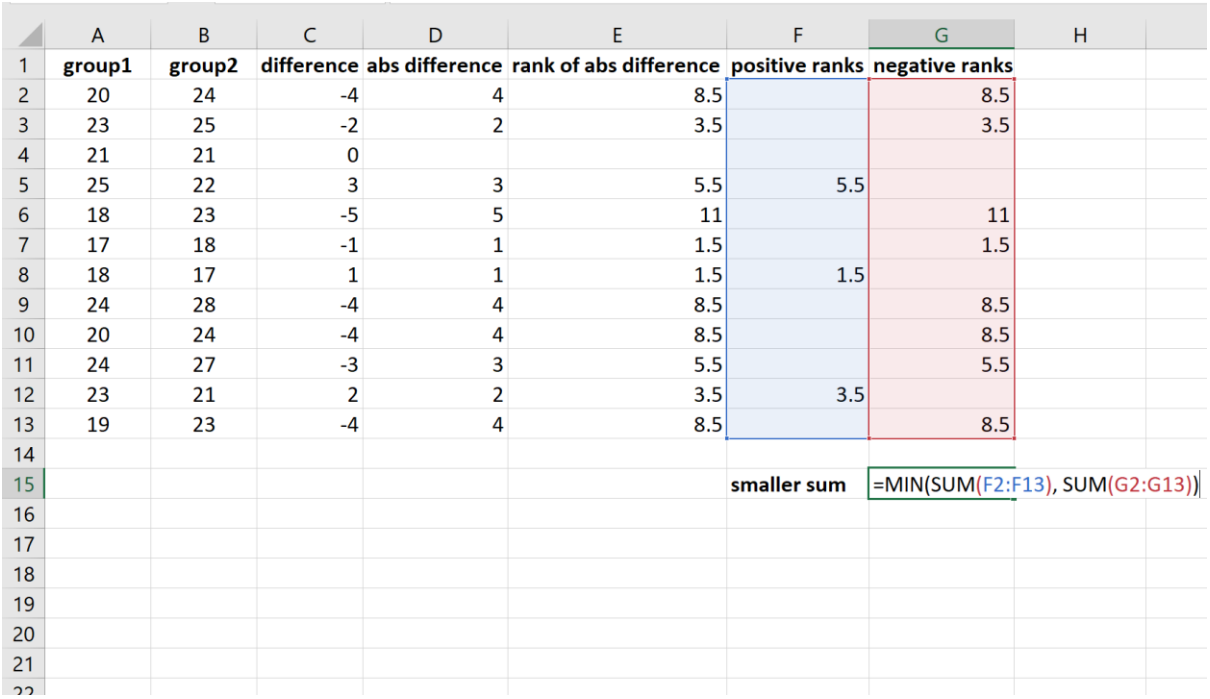
And we'll calculate the negative ranks:



	A	B	C	D	E	F	G
1	**group1**	**group2**	**difference**	**abs difference**	**rank of abs difference**	**positive ranks**	**negative ranks**
2	20	24	-4	4	8.5		=IF(C2<0, E2, "")
3	23	25	-2	2	3.5		3.5
4	21	21	0				
5	25	22	3	3	5.5	5.5	
6	18	23	-5	5	11		11
7	17	18	-1	1	1.5		1.5
8	18	17	1	1	1.5	1.5	
9	24	28	-4	4	8.5		8.5
10	20	24	-4	4	8.5		8.5
11	24	27	-3	3	5.5		5.5
12	23	21	2	2	3.5	3.5	
13	19	23	-4	4	8.5		8.5
14							
15							
16							
17							
18							
19							
20							
21							
22							

Step 6: Calculate the Test Statistic & Sample Size

Lastly, we'll calculate the test statistic which is simply the smaller of the sum of the positive ranks or the sum of the negative ranks:



	A	B	C	D	E	F	G	H
1	**group1**	**group2**	**difference**	**abs difference**	**rank of abs difference**	**positive ranks**	**negative ranks**	
2	20	24	-4	4	8.5		8.5	
3	23	25	-2	2	3.5		3.5	
4	21	21	0					
5	25	22	3	3	5.5	5.5		
6	18	23	-5	5	11		11	
7	17	18	-1	1	1.5		1.5	
8	18	17	1	1	1.5	1.5		
9	24	28	-4	4	8.5		8.5	
10	20	24	-4	4	8.5		8.5	
11	24	27	-3	3	5.5		5.5	
12	23	21	2	2	3.5	3.5		
13	19	23	-4	4	8.5		8.5	
14								
15						**smaller sum**	=MIN(SUM(F2:F13), SUM(G2:G13))	
16								
17								
18								
19								
20								
21								

And we'll calculate the sample size, which is the total number of ranks that aren't equal to zero:

	A	B	C	D	E	F	G
1	**group1**	**group2**	**difference**	**abs difference**	**rank of abs difference**	**positive ranks**	**negative ranks**
2	20	24	-4	4	8.5		8.5
3	23	25	-2	2	3.5		3.5
4	21	21	0				
5	25	22	3	3	5.5	5.5	
6	18	23	-5	5	11		11
7	17	18	-1	1	1.5		1.5
8	18	17	1	1	1.5	1.5	
9	24	28	-4	4	8.5		8.5
10	20	24	-4	4	8.5		8.5
11	24	27	-3	3	5.5		5.5
12	23	21	2	2	3.5	3.5	
13	19	23	-4	4	8.5		8.5
14							
15						**smaller sum**	10.5
16						**sample size**	=COUNT(F2:G13)
17							
18							
19							
20							

The test statistic turns out to be **10.5** and the sample size is **11**.

In this example, the Wilcoxon Signed-Rank Test uses the following null and alternative hypotheses:

H_0: The mpg is equal between the two groups

H_A: The mpg is *not* equal between the two groups

To determine if we should reject or fail to reject the null hypothesis, we can find the critical value that corresponds to $\alpha = .05$ and a sample size of 11 in the following Wilcoxon Signed Rank Test Critical Values Table:

	Alpha value				
n	**0.005**	**0.01**	**0.025**	**0.05**	**0.10**
5	-	-	-	-	0
6	-	-	-	0	2
7	-	-	0	2	3
8	-	0	2	3	5
9	0	1	3	5	8
10	1	3	5	8	10
11	3	5	8	10	13
12	5	7	10	13	17
13	7	9	13	17	21
14	9	12	17	21	25
15	12	15	20	25	30
16	15	19	25	29	35
17	19	23	29	34	41
18	23	27	34	40	47
19	27	32	39	46	53
20	32	37	45	52	60
21	37	42	51	58	67
22	42	48	57	65	75
23	48	54	64	73	83
24	54	61	72	81	91
25	60	68	79	89	100
26	67	75	87	98	110
27	74	83	96	107	119
28	82	91	105	116	130
29	90	100	114	126	140
30	98	109	124	137	151

The critical value that corresponds to $\alpha = .05$ and a sample size of 11 is **10**.

Since the test statistic (10.5) is not less than the critical value of 10, we fail to reject the null hypothesis.

We do not have sufficient evidence to say that the mean mpg is not equal between the two groups.

Practical No : 10 Perform a Kruskal-Wallis Test in Excel

Step 1: Enter the data.

Enter the following data, which shows the total growth (in inches) for each of the 10 plants in each group:

	A	B	C	D	E	F
1	**Fertilizer 1**	**Fertilizer 2**	**Fertilizer 3**			
2	7	15	6			
3	14	17	8			
4	14	13	8			
5	13	15	9			
6	12	15	5			
7	9	13	14			
8	6	9	13			
9	14	12	8			
10	12	10	10			
11	8	8	9			
12						
13						
14						
15						

Step 2: Rank the data.

Next, we will use the **RANK.AVG()** function to assign a rank to the growth of each plant out of all 30 plants. The following formula shows how to calculate the rank for the first plant in the first group:

	A	B	C	D	E	F	G
1	**Fertilizer 1**	**Fertilizer 2**	**Fertilizer 3**		**Fertilizer 1 Ranks**	**Fertilizer 2 Ranks**	**Fertilizer 3 Ranks**
2	7	15	6		=RANK.AVG(A2, $A$2:$C$11, 1)		
3	14	17	8				
4	14	13	8				
5	13	15	9				
6	12	15	5				
7	9	13	14				
8	6	9	13				
9	14	12	8				
10	12	10	10				
11	8	8	9				
12							
13							
14							
15							

Copy this formula to the rest of the cells:

	A	B	C	D	E	F	G
1	**Fertilizer 1**	**Fertilizer 2**	**Fertilizer 3**		**Fertilizer 1 Ranks**	**Fertilizer 2 Ranks**	**Fertilizer 3 Ranks**
2	7	15	6		4	28	2.5
3	14	17	8		24.5	30	7
4	14	13	8		24.5	20.5	7
5	13	15	9		20.5	28	11.5
6	12	15	5		17	28	1
7	9	13	14		11.5	20.5	24.5
8	6	9	13		2.5	11.5	20.5
9	14	12	8		24.5	17	7
10	12	10	10		17	14.5	14.5
11	8	8	9		7	7	11.5
12							
13							
14							
15							

Then, calculate the sum of the ranks for each column along with the sample size and the squared sum of ranks divided by the sample size:

	A	B	C	D	E	F	G
1	**Fertilizer 1**	**Fertilizer 2**	**Fertilizer 3**		**Fertilizer 1 Ranks**	**Fertilizer 2 Ranks**	**Fertilizer 3 Ranks**
2	7	15	6		4	28	2.5
3	14	17	8		24.5	30	7
4	14	13	8		24.5	20.5	7
5	13	15	9		20.5	28	11.5
6	12	15	5		17	28	1
7	9	13	14		11.5	20.5	24.5
8	6	9	13		2.5	11.5	20.5
9	14	12	8		24.5	17	7
10	12	10	10		17	14.5	14.5
11	8	8	9		7	7	11.5
12				**R**	153	205	107
13				**n**	10	10	10
14				**R^2 / n**	2340.9	4202.5	1144.9
15							
16							

Step 3: Calculate the test statistic and the corresponding p-value.

The test statistic is defined as:

$H = 12/(n(n+1)) * \Sigma R_j^2/n_j - 3(n+1)$

where:

- n = total sample size
- R_j^2 =sum of ranks for the j^{th} group
- n_j =sample size of j^{th} group

Under the null hypothesis, H follows a Chi-square distribution with k-1 degrees of freedom.

The following screenshot shows the formulas used to calculate the test statistic, H, and the corresponding p-value:

D	E	F	G	H	I	J	K	L	M	N
	Fertilizer 1 Ranks	**Fertilizer 2 Ranks**	**Fertilizer 3 Ranks**							
	4	28	2.5		**n**	30	=COUNT(E2:G11)			
	24.5	30	7		**k**	3	=COUNTA(A1:C1)			
	24.5	20.5	7		**H**	6.204	=12/(J2*(J2+1))*SUM(E14:G14)-3*(J2+1)			
	20.5	28	11.5		**p-value**	0.045	0.044962			
	17	28	1							
	11.5	20.5	24.5							
	2.5	11.5	20.5							
	24.5	17	7							
	17	14.5	14.5							
	7	7	11.5							
R	153	205	107							
n	10	10	10							
R^2 / n	2340.9	4202.5	1144.9							

The test statistic is H = **6.204** and the corresponding p-value is p = **0.045**. Since this p-value is less than 0.05, we can reject the null hypothesis that the median plant growth is the same for all three fertilizers. We have sufficient evidence to conclude that the type of fertilizer used leads to statistically significant differences in plant growth.

Step 4: Report the results.

Lastly, we want to report the results of the Kruskal-Wallis Test. Here is an example of how to do so:

A Kruskal-Wallist Test was performed to determine if median plant growth was the same for three different plant fertilizers. A total of 30 plants were used in the analysis. Each fertilizer was applied to 10 different plants.

The test revealed that the median plant growth was not the same (H = 6.204, p = 0.045) among the three fertilizers. That is, there was a statistically significant difference in median plant growth among two or more of the fertilizers.

Practical no : 11 Perform the Friedman Test in Excel

Step 1: Enter the data.

Enter the following data, which shows the reaction time (in seconds) of 10 patients on three different drugs. Since each patient is measured on each of the three drugs, we will use the Friedman Test to determine if the mean reaction time differs between drugs.

	A	B	C	D	E	F
1	**Patient**	**Drug 1**	**Drug 2**	**Drug 3**		
2	Patient 1	4	5	2		
3	Patient 2	6	6	4		
4	Patient 3	3	8	4		
5	Patient 4	4	7	3		
6	Patient 5	3	7	2		
7	Patient 6	2	8	2		
8	Patient 7	2	4	1		
9	Patient 8	7	6	4		
10	Patient 9	6	4	3		
11	Patient 10	5	5	2		
12						
13						
14						
15						
16						

Step 2: Rank the data.

Next, rank the data values in each row in ascending order using the **=RANK.AVG()** function. The following formula shows how to calculate the rank for the response time of patient 1 on drug 1:

	A	B	C	D	E	F	G	H
1	**Patient**	**Drug 1**	**Drug 2**	**Drug 3**		**Drug 1 Ranks**	**Drug 2 Ranks**	**Drug 3 Ranks**
2	Patient 1	4	5	2		=RANK.AVG(B2, $B2:$D2, 1)		
3	Patient 2	6	6	4				
4	Patient 3	3	8	4				
5	Patient 4	4	7	3				
6	Patient 5	3	7	2				
7	Patient 6	2	8	2				
8	Patient 7	2	4	1				
9	Patient 8	7	6	4				
10	Patient 9	6	4	3				
11	Patient 10	5	5	2				
12								
13								
14								
15								

Copy this formula to the rest of the cells:

	A	B	C	D	E	F	G	H
1	**Patient**	**Drug 1**	**Drug 2**	**Drug 3**		**Drug 1 Ranks**	**Drug 2 Ranks**	**Drug 3 Ranks**
2	Patient 1	4	5	2		2	3	1
3	Patient 2	6	6	4		2.5	2.5	1
4	Patient 3	3	8	4		1	3	2
5	Patient 4	4	7	3		2	3	1
6	Patient 5	3	7	2		2	3	1
7	Patient 6	2	8	2		1.5	3	1.5
8	Patient 7	2	4	1		2	3	1
9	Patient 8	7	6	4		3	2	1
10	Patient 9	6	4	3		3	2	1
11	Patient 10	5	5	2		2.5	2.5	1
12								
13								
14								
15								

Then, calculate the sum of the ranks for each column along with the squared sum of ranks:

	A	B	C	D	E	F	G	H
1	**Patient**	**Drug 1**	**Drug 2**	**Drug 3**		**Drug 1 Ranks**	**Drug 2 Ranks**	**Drug 3 Ranks**
2	Patient 1	4	5	2		2	3	1
3	Patient 2	6	6	4		2.5	2.5	1
4	Patient 3	3	8	4		1	3	2
5	Patient 4	4	7	3		2	3	1
6	Patient 5	3	7	2		2	3	1
7	Patient 6	2	8	2		1.5	3	1.5
8	Patient 7	2	4	1		2	3	1
9	Patient 8	7	6	4		3	2	1
10	Patient 9	6	4	3		3	2	1
11	Patient 10	5	5	2		2.5	2.5	1
12					**Sum of Ranks**	21.5	27	11.5
13					**Sum of Ranks Squared**	462.25	729	132.25
14								
15								

Step 3: Calculate the test statistic and the corresponding p-value.

The test statistic is defined as:

$Q = 12/nk(k+1) * \Sigma R_j^2 - 3n(k+1)$

where:

- n = number of patients
- k = number of treatment groups
- R_j^2 =sum of ranks for the j^{th} group

Under the null hypothesis, Q follows a chi-square distribution with k-1 degrees of freedom.

The following screenshot shows the formulas used to calculate the test statistic, Q, and the corresponding p-value:

F	G	H	I	J	K	L	M	N	O	P
Drug 1 Ranks	**Drug 2 Ranks**	**Drug 3 Ranks**								
2	3	1		**k**	3	=COUNTA(B1:D1)				
2.5	2.5	1		**n**	10	=COUNTA(A2:A11)				
1	3	2		**Q**	12.35	=12/(K3*K2*(K2+1))*(SUM(F13:H13))-3*K3*(K2+1)				
2	3	1		**p-value**	0.00208	=CHISQ.DIST.RT(K4, K2-1)				
2	3	1								
1.5	3	1.5								
2	3	1								
3	2	1								
3	2	1								
2.5	2.5	1								
21.5	27	11.5								
462.25	729	132.25								

The test statistic is Q = **12.35** and the corresponding p-value is p = **0.00208**. Since this value is less than 0.05, we can reject the null hypothesis that the mean response time is the same for all three drugs. We have sufficient evidence to conclude that the type of drug used leads to statistically significant differences in response time.

Step 4: Report the results.

Lastly, we want to report the results of the test. Here is an example of how to do so:

A Friedman Test was conducted on 10 patients to examine the effect that three different drugs had on response time. Each patient used each drug once.

Results showed that the type of drug used lead to statistically significant differences in response time (Q = 12.35, p = 0.00208).

Practical no : 12 Perform Multiple Linear Regression in Excel

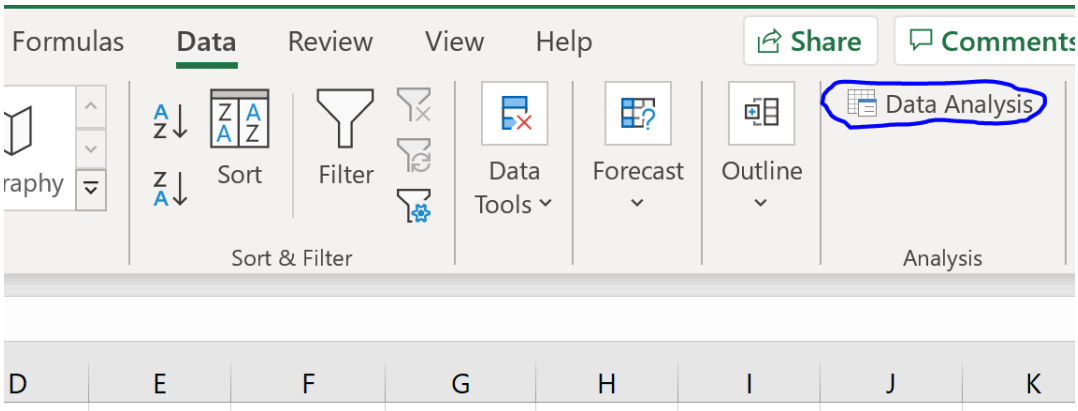
Step 1: Enter the data.

Enter the following data for the number of hours studied, prep exams taken, and exam score received for 20 students:

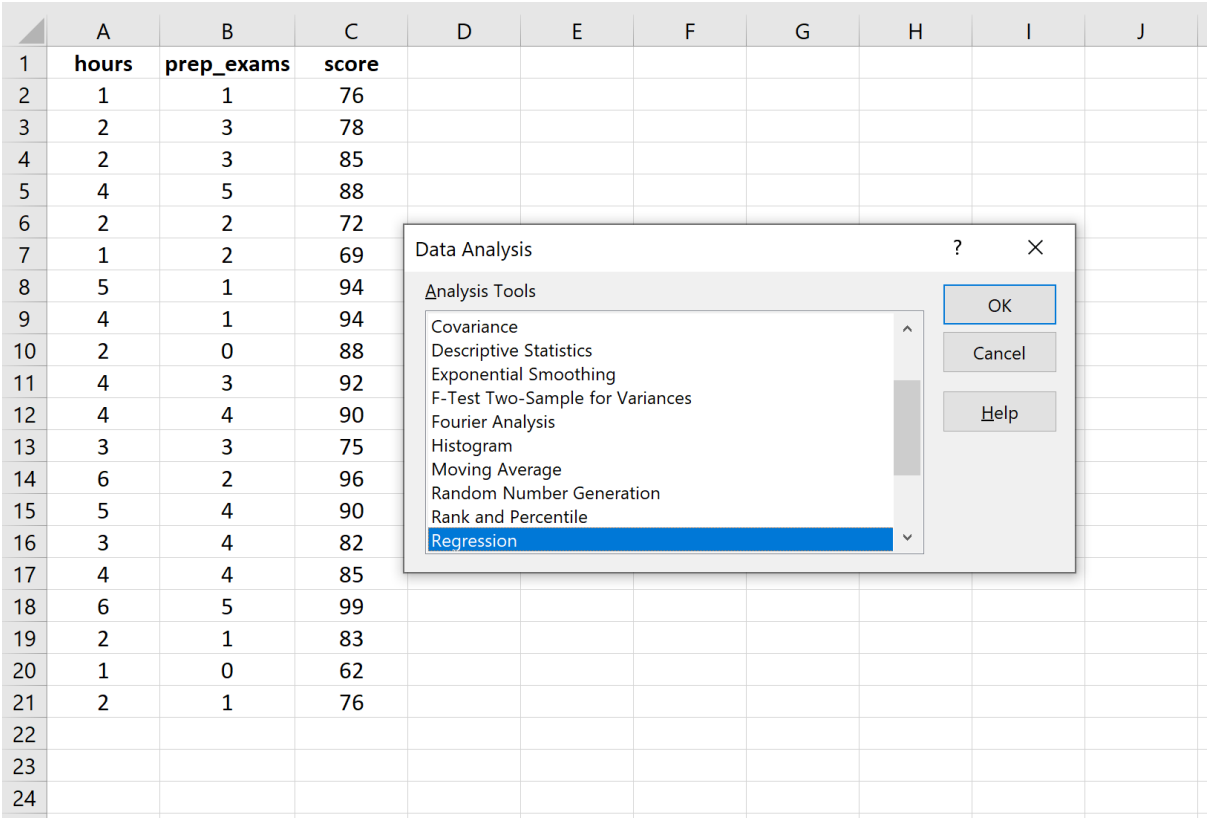
	A	B	C	D	E	F
1	**hours**	**prep_exams**	**score**			
2	1	1	76			
3	2	3	78			
4	2	3	85			
5	4	5	88			
6	2	2	72			
7	1	2	69			
8	5	1	94			
9	4	1	94			
10	2	0	88			
11	4	3	92			
12	4	4	90			
13	3	3	75			
14	6	2	96			
15	5	4	90			
16	3	4	82			
17	4	4	85			
18	6	5	99			
19	2	1	83			
20	1	0	62			
21	2	1	76			
22						
23						
24						

Step 2: Perform multiple linear regression.

Along the top ribbon in Excel, go to the **Data** tab and click on **Data Analysis**. If you don't see this option, then you need to first install the free Analysis ToolPak.

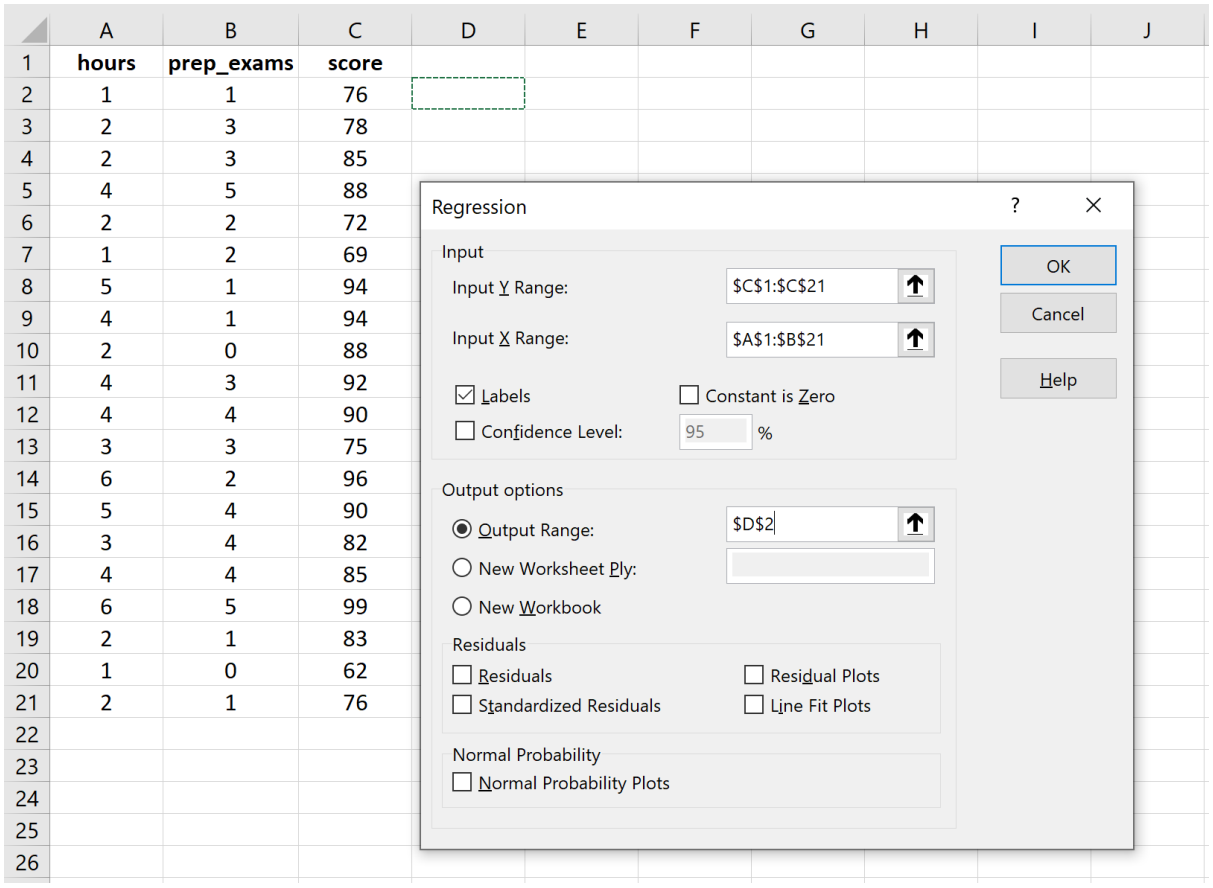


Once you click on **Data Analysis,** a new window will pop up. Select **Regression** and click OK.



	A	B	C	D	E	F	G	H	I	J
1	**hours**	**prep_exams**	**score**							
2	1	1	76							
3	2	3	78							
4	2	3	85							
5	4	5	88							
6	2	2	72							
7	1	2	69							
8	5	1	94							
9	4	1	94							
10	2	0	88							
11	4	3	92							
12	4	4	90							
13	3	3	75							
14	6	2	96							
15	5	4	90							
16	3	4	82							
17	4	4	85							
18	6	5	99							
19	2	1	83							
20	1	0	62							
21	2	1	76							
22										
23										
24										

For **Input Y Range**, fill in the array of values for the response variable. For **Input X Range**, fill in the array of values for the two explanatory variables. Check the box next to **Labels** so Excel knows that we included the variable names in the input ranges. For **Output Range**, select a cell where you would like the output of the regression to appear. Then click **OK**.



	A	B	C
1	**hours**	**prep_exams**	**score**
2	1	1	76
3	2	3	78
4	2	3	85
5	4	5	88
6	2	2	72
7	1	2	69
8	5	1	94
9	4	1	94
10	2	0	88
11	4	3	92
12	4	4	90
13	3	3	75
14	6	2	96
15	5	4	90
16	3	4	82
17	4	4	85
18	6	5	99
19	2	1	83
20	1	0	62
21	2	1	76

The following output will automatically appear:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.857
R Square	0.734
Adjusted R Square	0.703
Standard Error	5.366
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	1350.76	675.38	23.46	0.00
Residual	17	489.44	28.79		
Total	19	1840.20			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	67.67	2.82	24.03	0.00	61.73	73.61
hours	5.56	0.90	6.18	0.00	3.66	7.45
prep_exams	-0.60	0.91	-0.66	0.52	-2.53	1.33

Step 3: Interpret the output.

Here is how to interpret the most relevant numbers in the output:

R Square: 0.734. This is known as the coefficient of determination. It is the proportion of the variance in the response variable that can be explained by the explanatory variables. In this example, 73.4% of the variation in the exam scores can be explained by the number of hours studied and the number of prep exams taken.

Standard error: 5.366. This is the average distance that the observed values fall from the regression line. In this example, the observed values fall an average of 5.366 units from the regression line.

F: 23.46. This is the overall F statistic for the regression model, calculated as regression MS / residual MS.

Significance F: 0.0000. This is the p-value associated with the overall F statistic. It tells us whether or not the regression model as a whole is statistically significant. In other words, it tells us if the two explanatory variables combined have a statistically significant association with the response variable. In this case the p-value is less than 0.05, which indicates that the explanatory variables **hours studied** and **prep exams taken** combined have a statistically significant association with **exam score**.

P-values. The individual p-values tell us whether or not each explanatory variable is statistically significant. We can see that **hours studied** is statistically significant (p = 0.00) while **prep exams taken** (p = 0.52) is not statistically signifciant at α = 0.05. Since **prep exams taken** is not statistically significant, we may end up deciding to remove it from the model.

Coefficients: The coefficients for each explanatory variable tell us the average expected change in the response variable, assuming the other explanatory variable remains constant. For example, for each additional hour spent studying, the average exam score is expected to increase by **5.56**, assuming that **prep exams taken** remains constant.

Here's another way to think about this: If student A and student B both take the same amount of prep exams but student A studies for one hour more, then student A is expected to earn a score that is **5.56** points higher than student B.

We interpret the coefficient for the intercept to mean that the expected exam score for a student who studies zero hours and takes zero prep exams is **67.67**.

Estimated regression equation: We can use the coefficients from the output of the model to create the following estimated regression equation:

exam score = 67.67 + 5.56*(hours) – 0.60*(prep exams)

We can use this estimated regression equation to calculate the expected exam score for a student, based on the number of hours they study and the number of prep exams they take. For example, a student who studies for three hours and takes one prep exam is expected to receive a score of **83.75**:

exam score = 67.67 + 5.56*(3) – 0.60*(1) = 83.75

Keep in mind that because **prep exams taken** was not statistically significant (p = 0.52), we may decide to remove it because it doesn't add any improvement to the overall model. In this case, we could perform simple linear regression using only **hours studied** as the explanatory variable.

Practical no : 13 How to Perform Simple Linear Regression in Excel

Step 1: Enter the data.

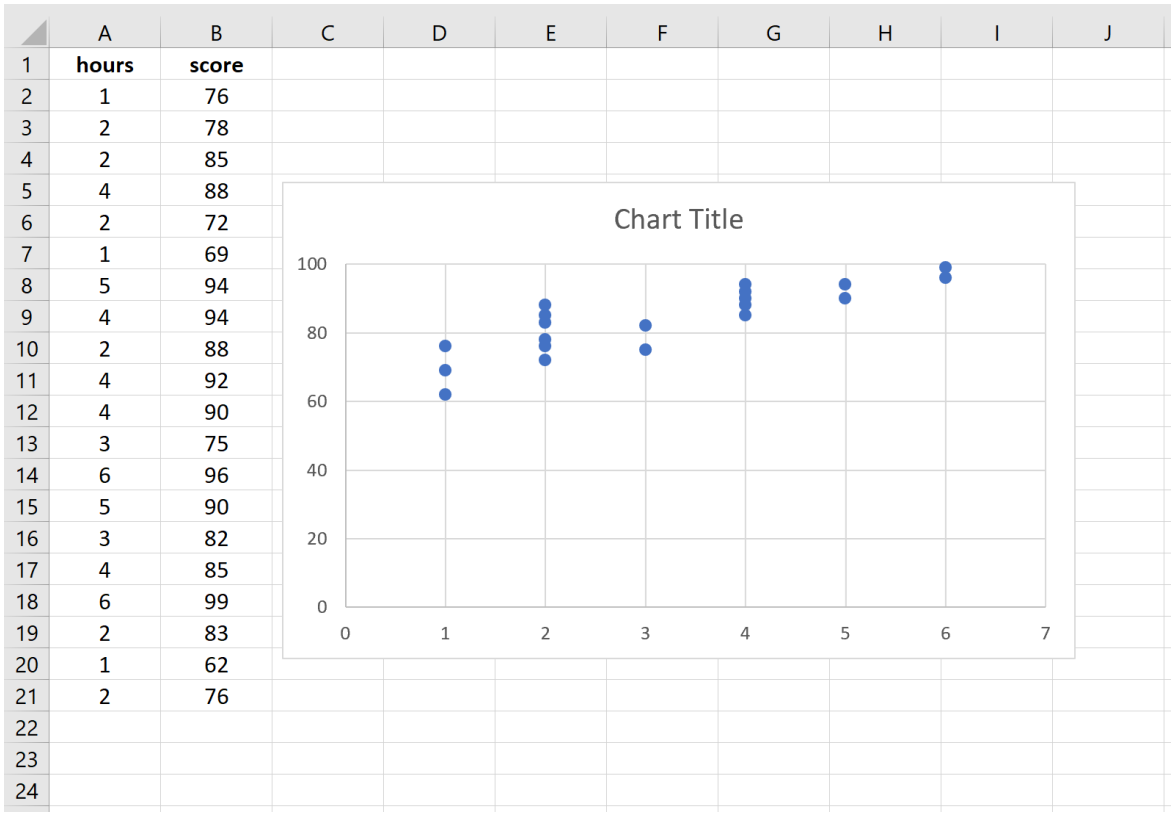
Enter the following data for the number of hours studied and the exam score received for 20 students:

	A	B	C	D	E
1	**hours**	**score**			
2	1	76			
3	2	78			
4	2	85			
5	4	88			
6	2	72			
7	1	69			
8	5	94			
9	4	94			
10	2	88			
11	4	92			
12	4	90			
13	3	75			
14	6	96			
15	5	90			
16	3	82			
17	4	85			
18	6	99			
19	2	83			
20	1	62			
21	2	76			
22					
23					
24					

Step 2: Visualize the data.

Before we perform simple linear regression, it's helpful to create a scatterplot of the data to make sure there actually exists a linear relationship between hours studied and exam score.

Highlight the data in columns A and B. Along the top ribbon in Excel go to the **Insert** tab. Within the **Charts** group, click **Insert Scatter (X, Y)** and click on the first option titled **Scatter**. This will automatically produce the following scatterplot:



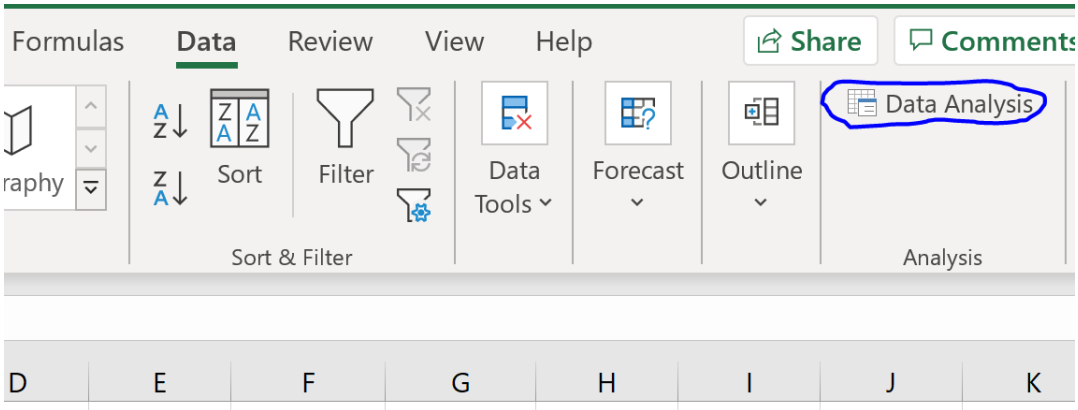
	A	B
1	**hours**	**score**
2	1	76
3	2	78
4	2	85
5	4	88
6	2	72
7	1	69
8	5	94
9	4	94
10	2	88
11	4	92
12	4	90
13	3	75
14	6	96
15	5	90
16	3	82
17	4	85
18	6	99
19	2	83
20	1	62
21	2	76
22		
23		
24		

The number of hours studied is shown on the x-axis and the exam scores are shown on the y-axis. We can see that there is a linear relationship between the two variables – more hours studied is associated with higher exam scores.

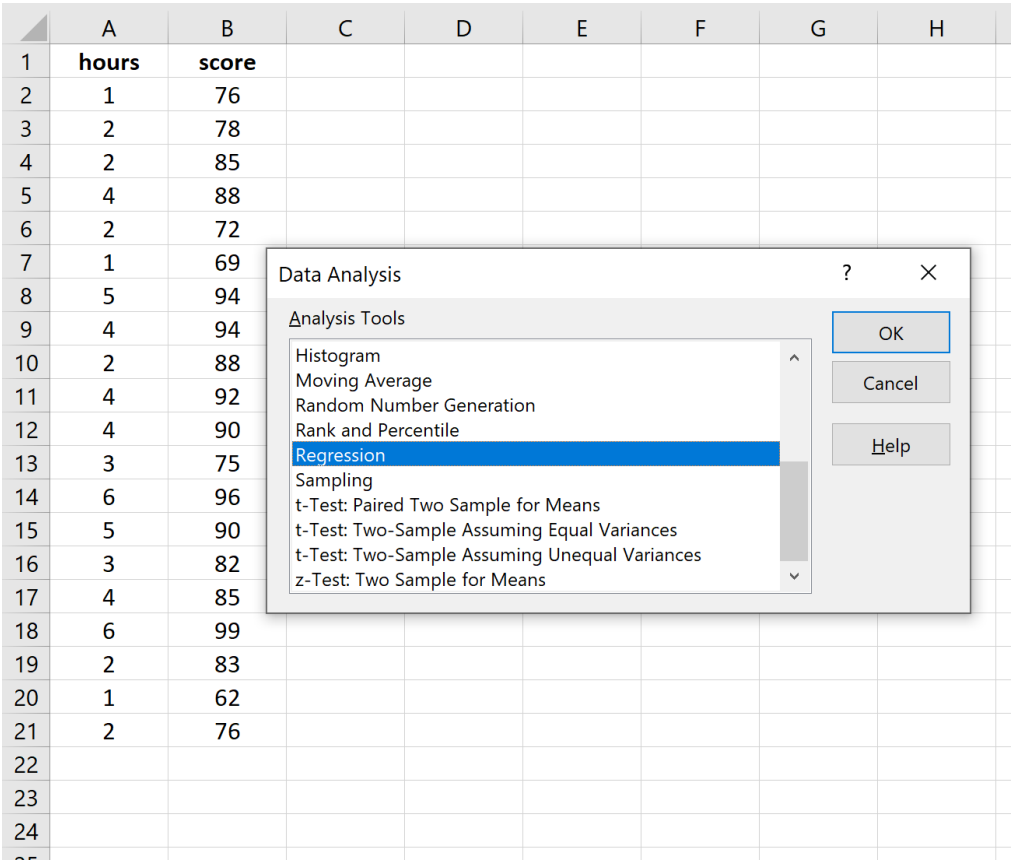
To quantify the relationship between these two variables, we can perform simple linear regression.

Step 3: Perform simple linear regression.

Along the top ribbon in Excel, go to the **Data** tab and click on **Data Analysis**. If you don't see this option, then you need to first install the free Analysis ToolPak.



Once you click on **Data Analysis,** a new window will pop up. Select **Regression** and click OK.



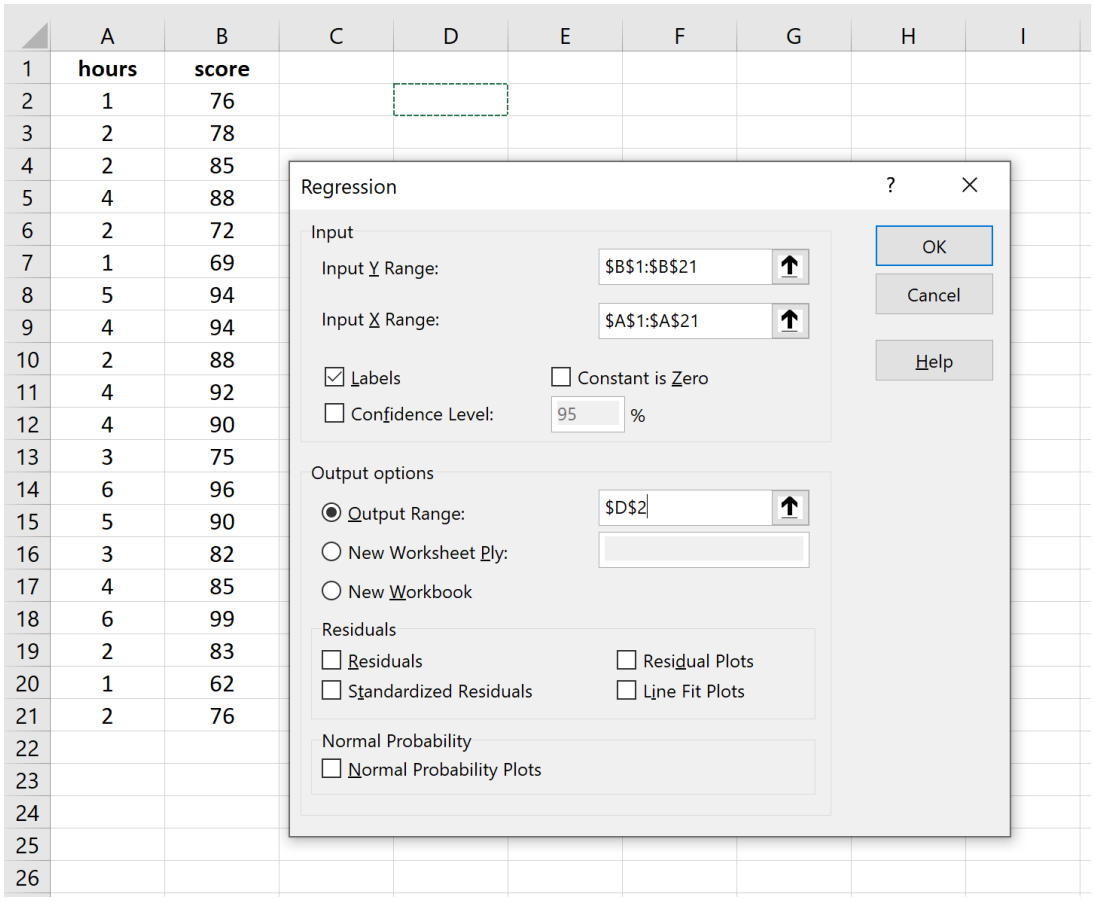
	A	B
1	**hours**	**score**
2	1	76
3	2	78
4	2	85
5	4	88
6	2	72
7	1	69
8	5	94
9	4	94
10	2	88
11	4	92
12	4	90
13	3	75
14	6	96
15	5	90
16	3	82
17	4	85
18	6	99
19	2	83
20	1	62
21	2	76

For **Input Y Range**, fill in the array of values for the response variable. For **Input X Range**, fill in the array of values for the explanatory variable.

Check the box next to **Labels** so Excel knows that we included the variable names in the input ranges.

For **Output Range**, select a cell where you would like the output of the regression to appear.

Then click **OK**.



	A	B
1	**hours**	**score**
2	1	76
3	2	78
4	2	85
5	4	88
6	2	72
7	1	69
8	5	94
9	4	94
10	2	88
11	4	92
12	4	90
13	3	75
14	6	96
15	5	90
16	3	82
17	4	85
18	6	99
19	2	83
20	1	62
21	2	76

The following output will automatically appear:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8528
R Square	0.7273
Adjusted R Square	0.7121
Standard Error	5.2805
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1338.2906	1338.2906	47.9952	0.0000
Residual	18	501.9094	27.8839		
Total	19	1840.2000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	67.1617	2.6633	25.2178	0.0000	61.5664	72.7570	61.5664	72.7570
hours	5.2503	0.7578	6.9279	0.0000	3.6581	6.8424	3.6581	6.8424

Step 4: Interpret the output.

Here is how to interpret the most relevant numbers in the output:

R Square: 0.7273. This is known as the coefficient of determination. It is the proportion of the variance in the response variable that can be explained by the explanatory variable. In this example, 72.73% of the variation in the exam scores can be explained by the number of hours studied.

Standard error: 5.2805. This is the average distance that the observed values fall from the regression line. In this example, the observed values fall an average of 5.2805 units from the regression line.

F: 47.9952. This is the overall F statistic for the regression model, calculated as regression MS / residual MS.

Significance F: 0.0000. This is the p-value associated with the overall F statistic. It tells us whether or not the regression model is statistically significant. In other words, it tells us if the explanatory variable has a statistically significant association with the response variable. In this case the p-value is less than 0.05, which indicates that there is a statistically significant association between hours studied and exam score received.

Coefficients: The coefficients give us the numbers necessary to write the estimated regression equation. In this example the estimated regression equation is:

exam score = 67.16 + 5.2503*(hours)

We interpret the coefficient for hours to mean that for each additional hour studied, the exam score is expected to increase by **5.2503**, on average. We interpret the coefficient for the intercept to mean that the expected exam score for a student who studies zero hours is **67.16**.

We can use this estimated regression equation to calculate the expected exam score for a student, based on the number of hours they study.

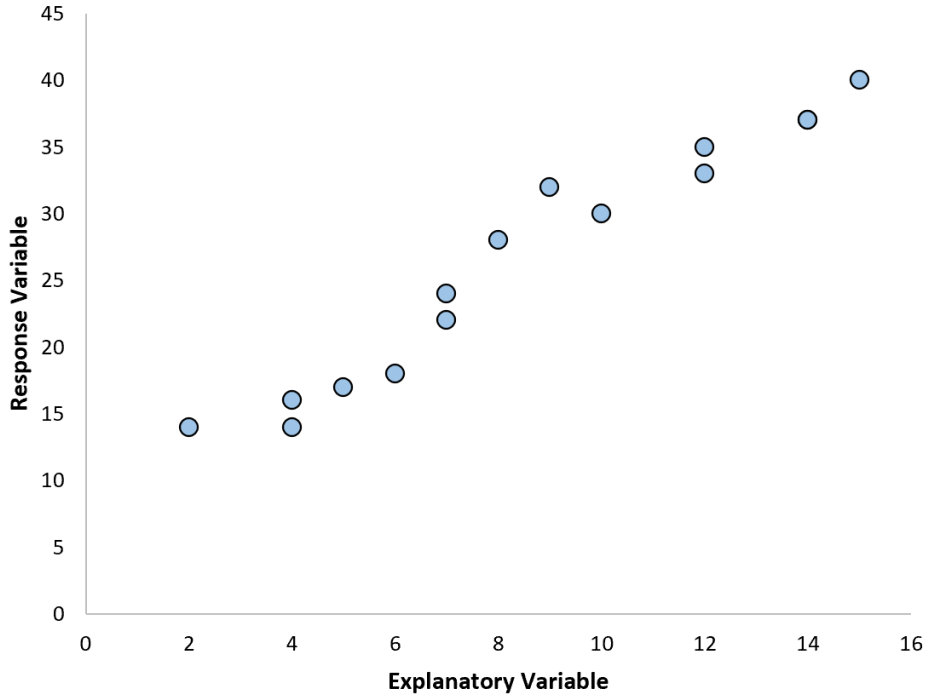
For example, a student who studies for three hours is expected to receive an exam score of **82.91**:

exam score = 67.16 + 5.2503*(3) = 82.91

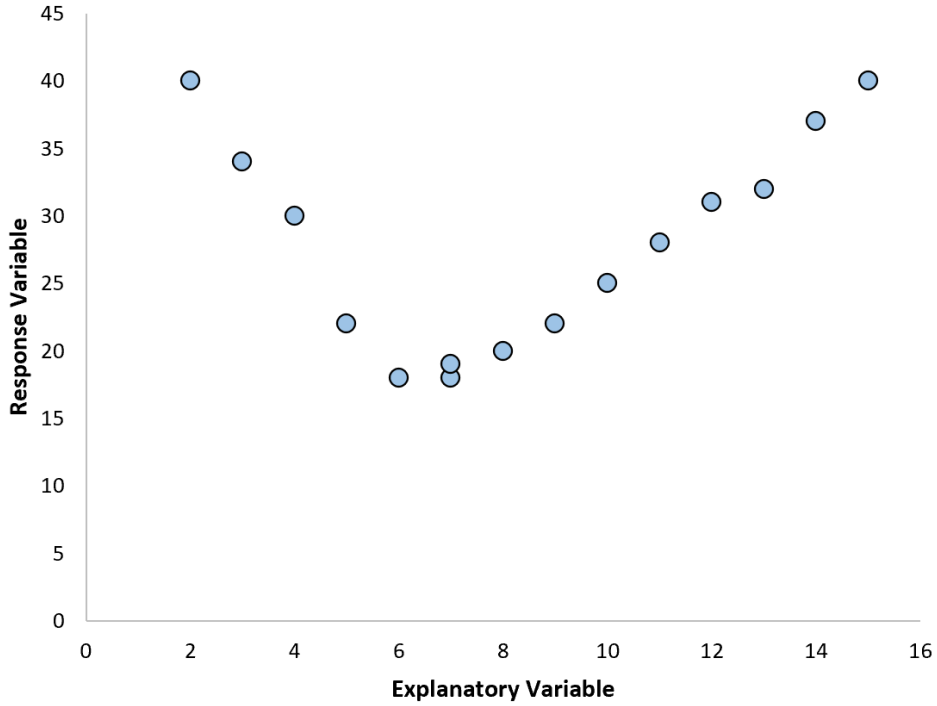
Practical no : 14 Perform Polynomial Regression in Excel

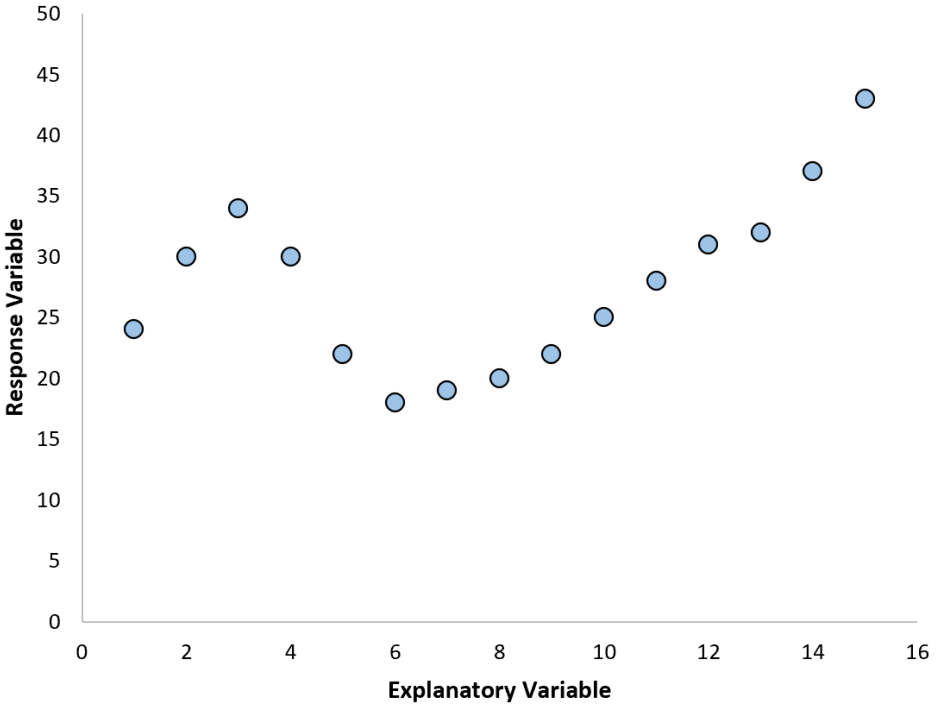
Regression analysis is used to quantify the relationship between one or more explanatory variables and a response variable.

The most common type of regression analysis is simple linear regression, which is used when an explanatory variable and a response variable have a linear relationship.



However, sometimes the relationship between an explanatory variable and a response variable is nonlinear.





In these cases it makes sense to use **polynomial regression**, which can account for the nonlinear relationship between the variables.

This tutorial explains how to perform polynomial regression in Excel.

Example: Polynomial Regression in Excel

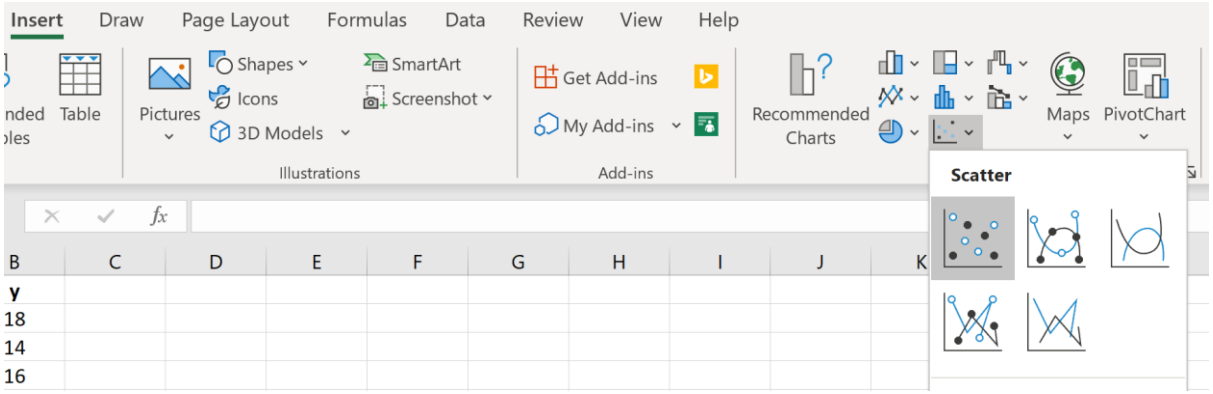
Suppose we have the following dataset in Excel:

	A	B	C	D	E	F
1	**x**	**y**				
2	2	18				
3	4	14				
4	4	16				
5	5	17				
6	6	18				
7	7	23				
8	7	25				
9	8	28				
10	9	32				
11	12	29				
12						
13						
14						
15						
16						
17						
18						
19						

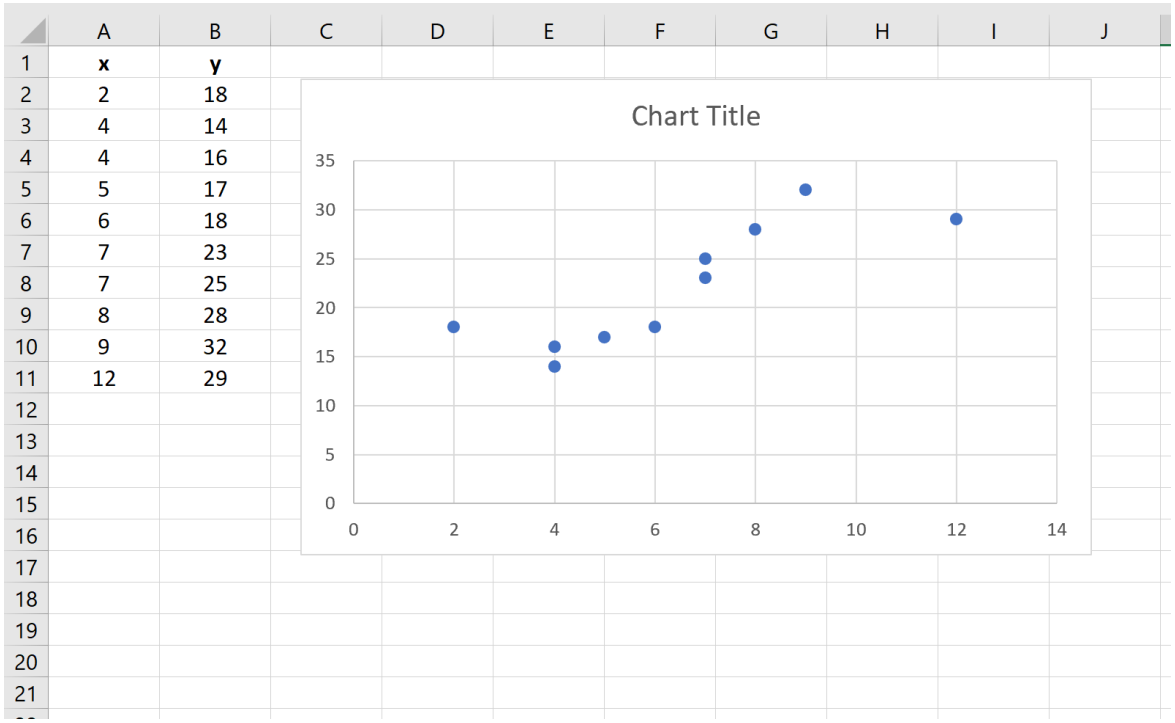
Use the following steps to fit a polynomial regression equation to this dataset:

Step 1: Create a scatterplot.

First, we need to create a scatterplot. Go to the **Charts** group in the **Insert** tab and click the first chart type in **Scatter**:



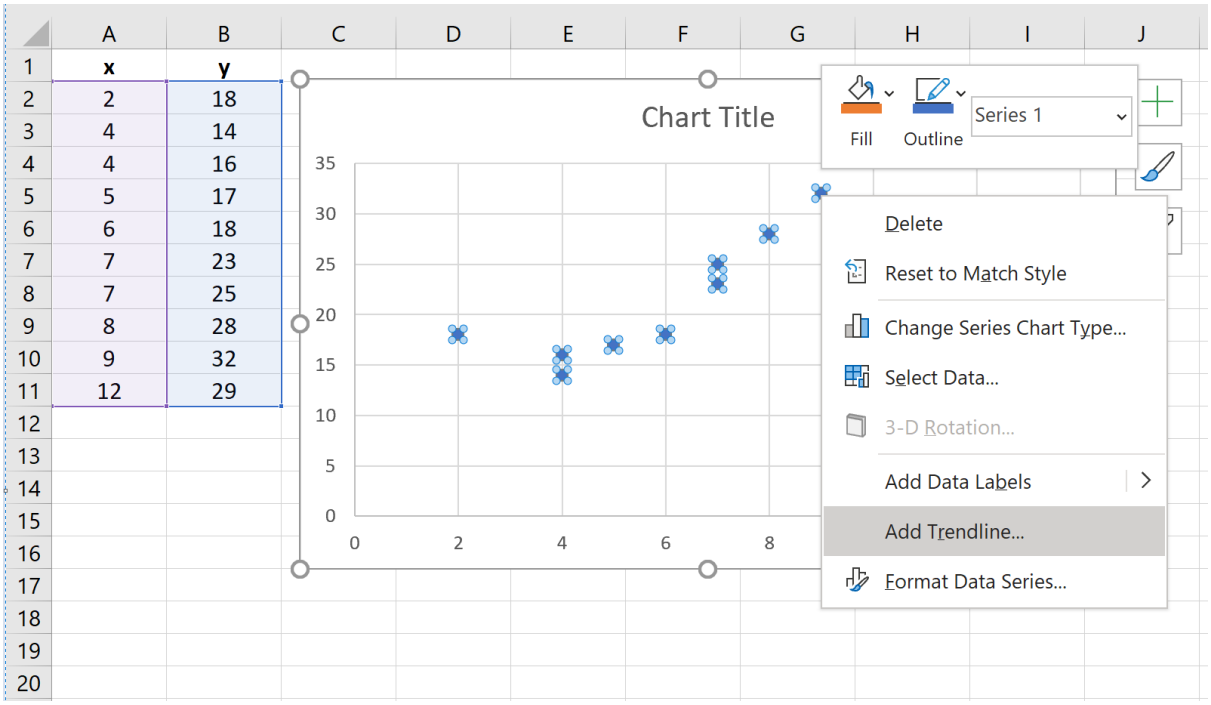
A scatterplot will automatically appear:



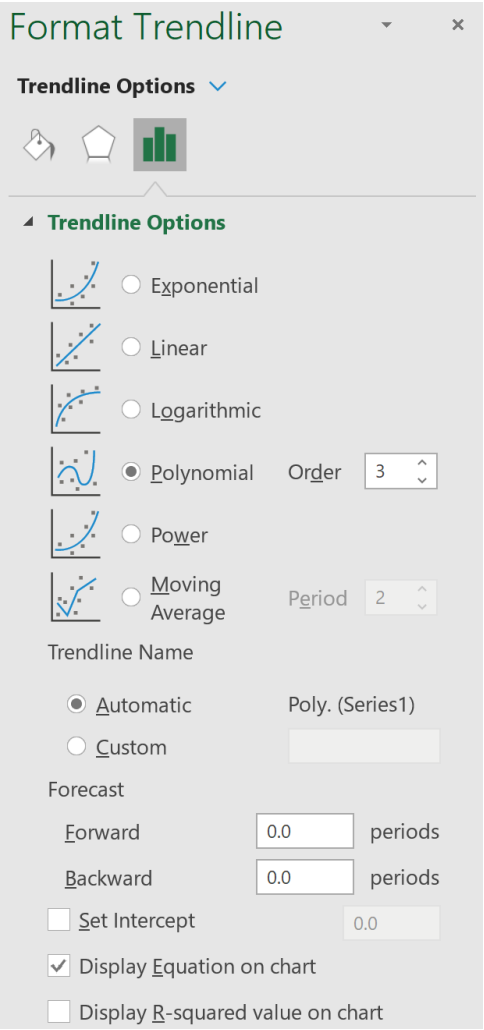
	A	B
1	**x**	**y**
2	2	18
3	4	14
4	4	16
5	5	17
6	6	18
7	7	23
8	7	25
9	8	28
10	9	32
11	12	29

Step 2: Add a trendline.

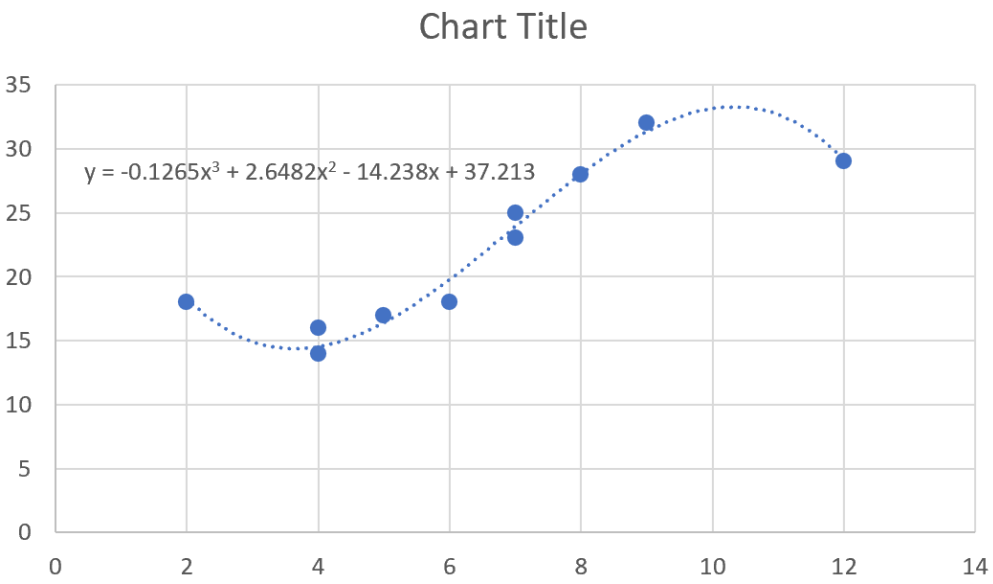
Next, we need to add a trendline to the scatterplot. To do so, click on any of the individual points in the scatterplot. Then, right click and select **Add Trendline…**



A new window will pop up with the option to specify a trendline. Choose **Polynomial** and choose the number you'd like to use for **Order**. We will use 3. Then, check the box near the bottom that says **Display Equation on chart**.



A trendline with a polynomial regression equation will automatically appear on the scatterplot:



Step 3: Interpret the regression equation.

For this particular example, our fitted polynomial regression equation is:

$y = -0.1265x^3 + 2.6482x^2 – 14.238x + 37.213$

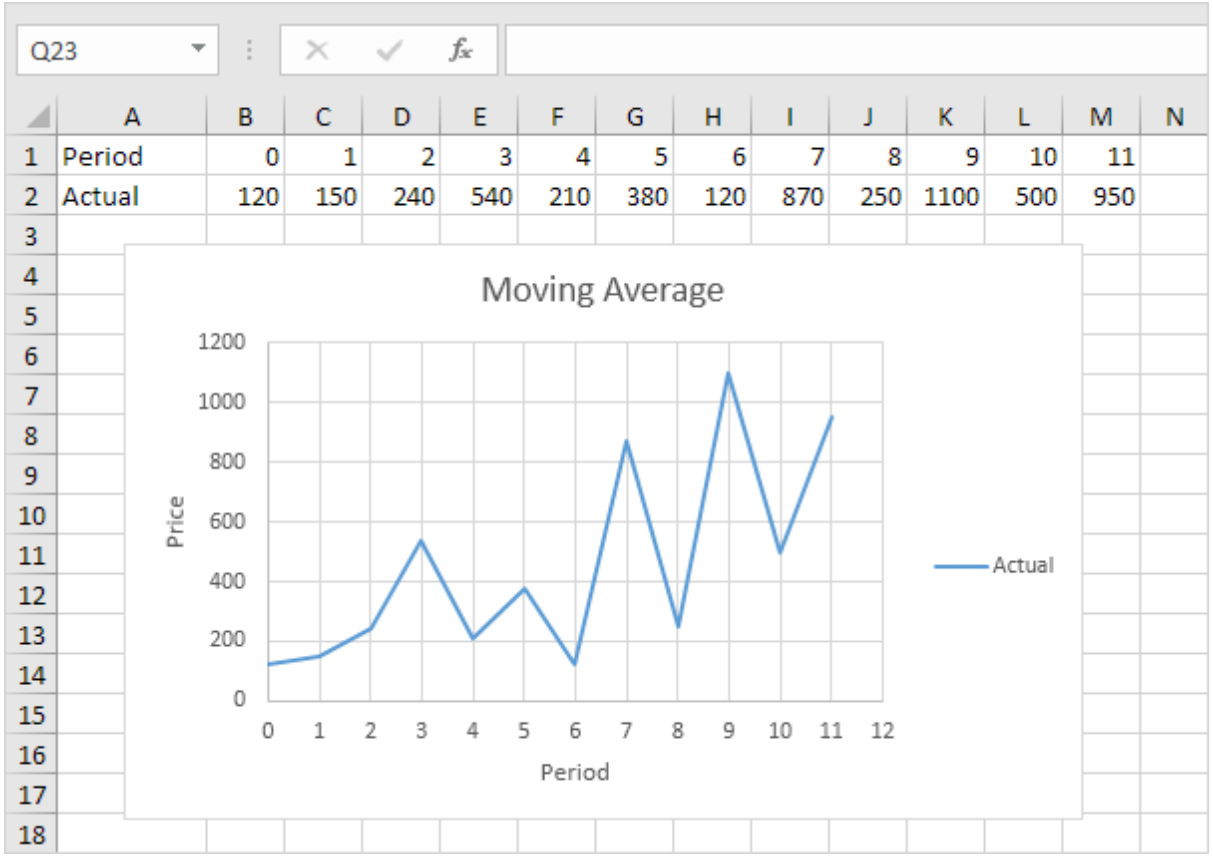
This equation can be used to find the expected value for the response variable based on a given value for the explanatory variable. For example, suppose x = 4. The expected value for the response variable, y, would be:

$y = -0.1265(4)^3 + 2.6482(4)^2 – 14.238(4) + 37.213 =$ **14.5362**.

Practical no : 15 Moving Average in Excel

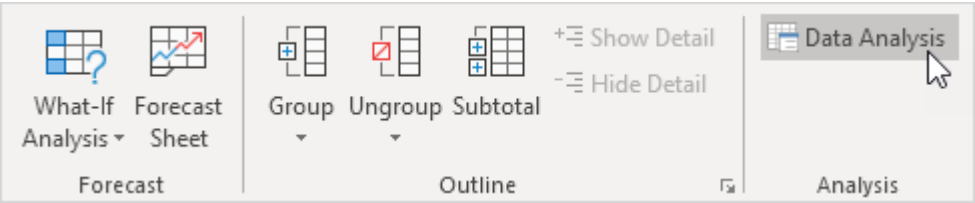
Steps to perform Moving Average in Excel :

1. First, let's take a look at our time series.



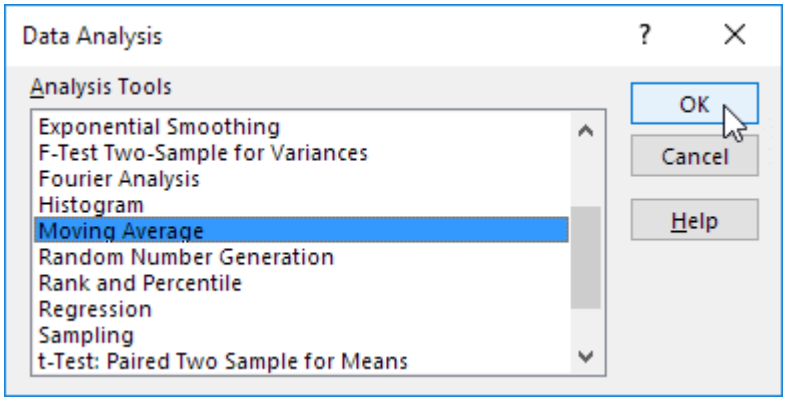
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Period	0	1	2	3	4	5	6	7	8	9	10	11	
2	Actual	120	150	240	540	210	380	120	870	250	1100	500	950	

2. On the Data tab, in the Analysis group, click Data Analysis.



Note: can't find the Data Analysis button? Click here to load the Analysis ToolPak add-in.

3. Select Moving Average and click OK.

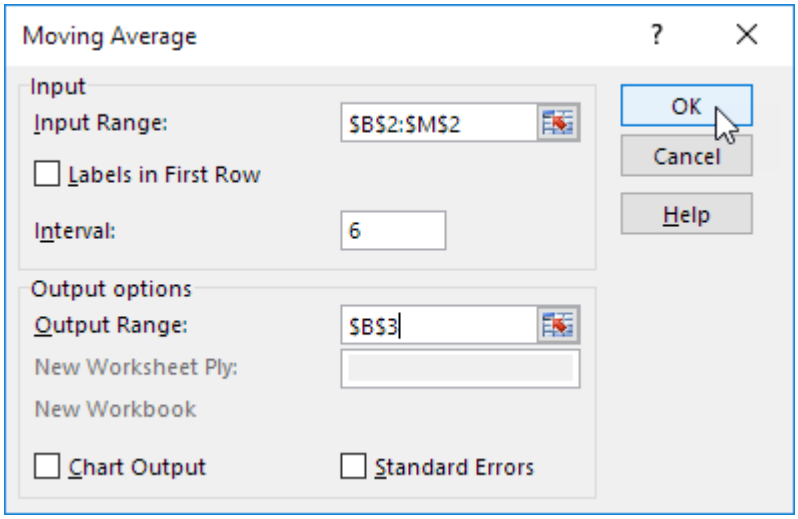


4. Click in the Input Range box and select the range B2:M2.

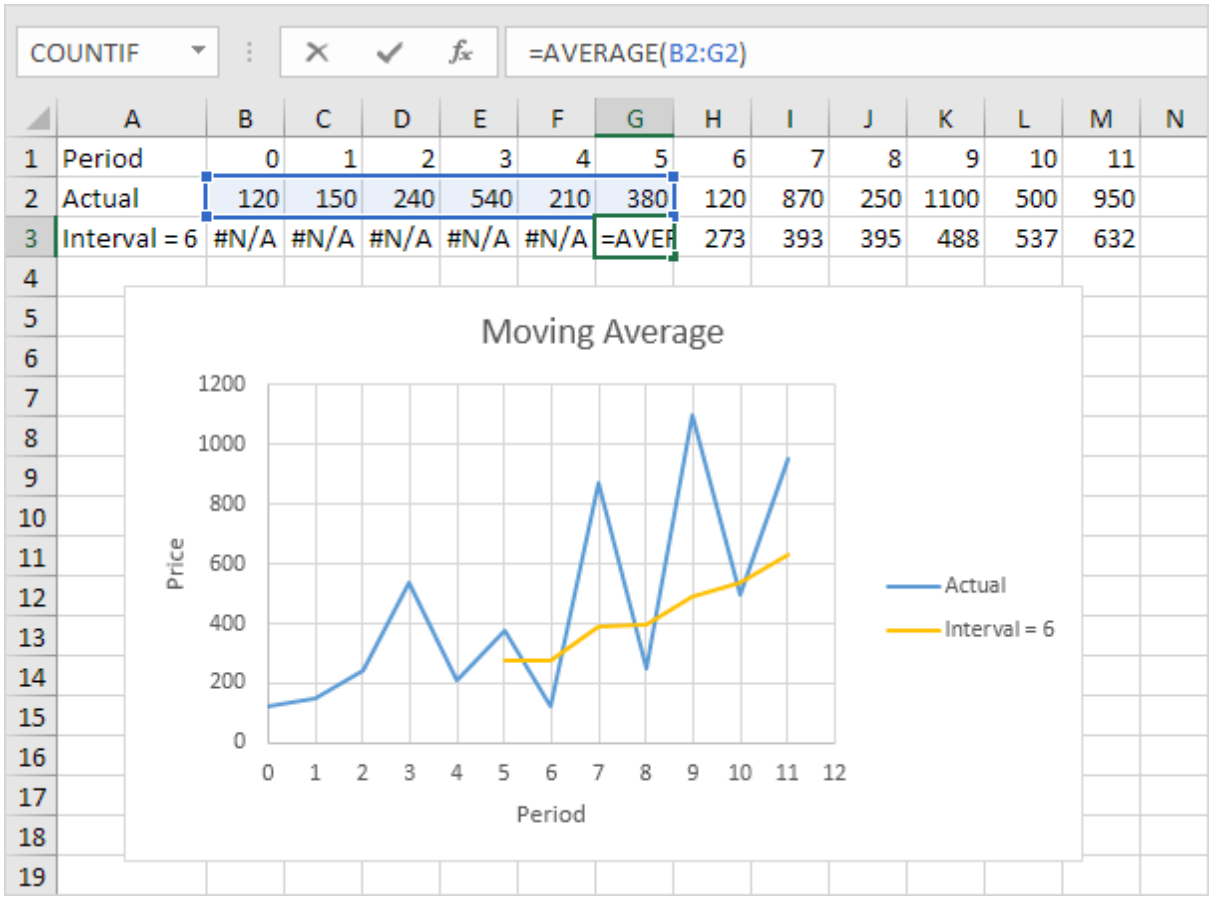
5. Click in the Interval box and type 6.

6. Click in the Output Range box and select cell B3.

7. Click OK.



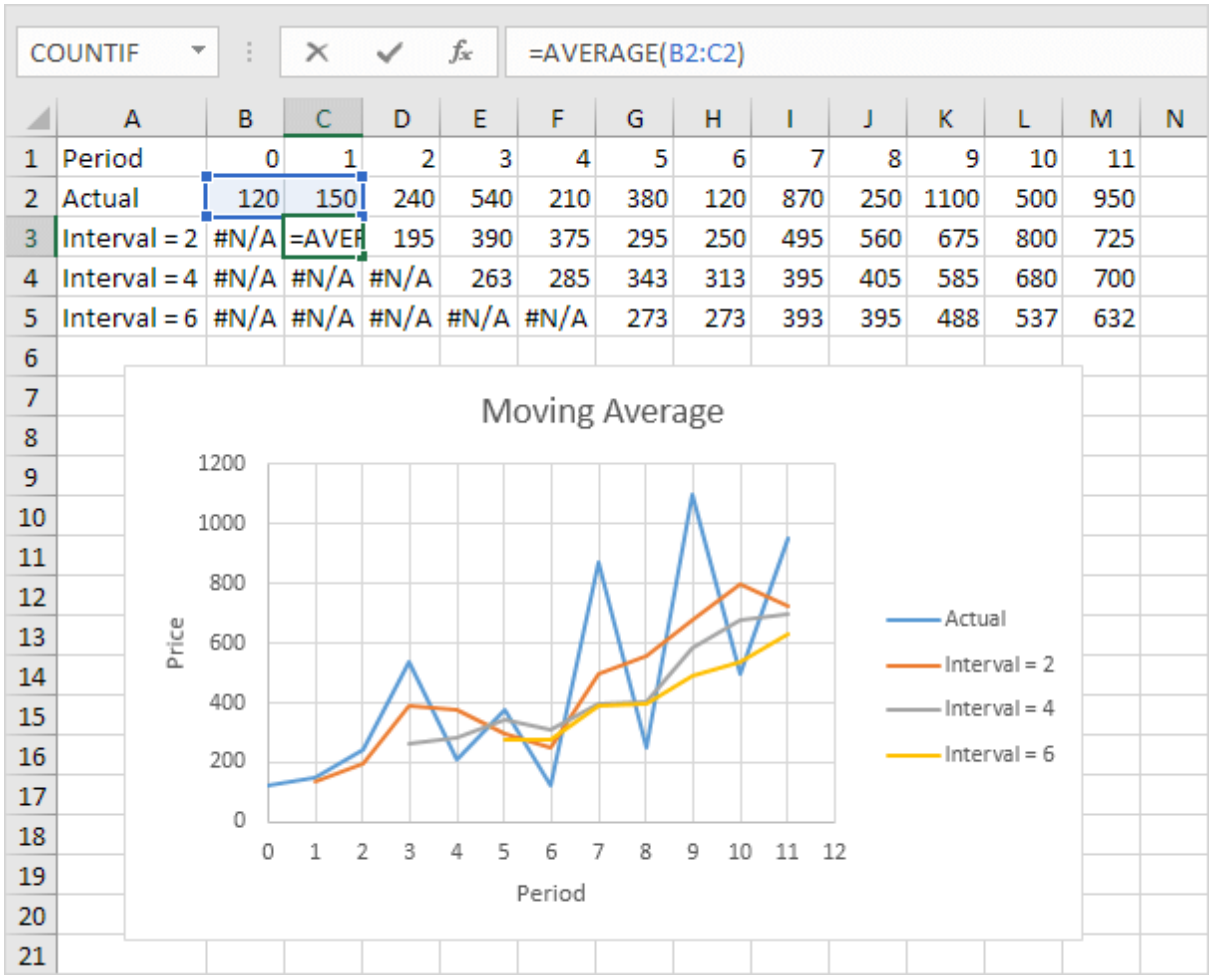
8. Plot a graph of these values.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Period	0	1	2	3	4	5	6	7	8	9	10	11
2	Actual	120	150	240	540	210	380	120	870	250	1100	500	950
3	Interval = 6	#N/A	#N/A	#N/A	#N/A	#N/A	=AVER	273	393	395	488	537	632

Explanation: because we set the interval to 6, the moving average is the average of the previous 5 data points and the current data point. As a result, peaks and valleys are smoothed out. The graph shows an increasing trend. Excel cannot calculate the moving average for the first 5 data points because there are not enough previous data points.

9. Repeat steps 2 to 8 for interval = 2 and interval = 4.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Period	0	1	2	3	4	5	6	7	8	9	10	11
2	Actual	120	150	240	540	210	380	120	870	250	1100	500	950
3	Interval = 2	#N/A	=AVER	195	390	375	295	250	495	560	675	800	725
4	Interval = 4	#N/A	#N/A	#N/A	263	285	343	313	395	405	585	680	700
5	Interval = 6	#N/A	#N/A	#N/A	#N/A	#N/A	273	273	393	395	488	537	632

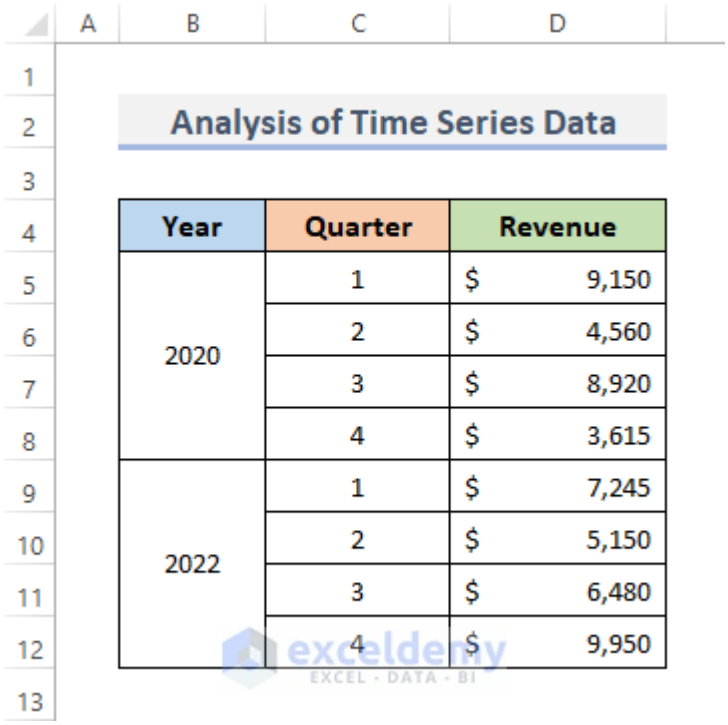
Conclusion: The larger the interval, the more the peaks and valleys are smoothed out. The smaller the interval, the closer the moving averages are to the actual data points.

Practical no : 16 Analyze Time Series Data in Excel (With Easy Steps)

Step 1: Input Time Series Data

To illustrate the time series analysis, we are going to use a company's quarterly revenue in two specific years. For instance, we need to input the time series data properly.

- Firstly, we put the year series data in column **B**. In our case, it has only been two years.
- Secondly, input the quarter of each year.
- Thirdly, just insert the total revenue in every quarter.



Analysis of Time Series Data

Year	Quarter	Revenue
2020	1	$ 9,150
	2	$ 4,560
	3	$ 8,920
	4	$ 3,615
2022	1	$ 7,245
	2	$ 5,150
	3	$ 6,480
	4	$ 9,950

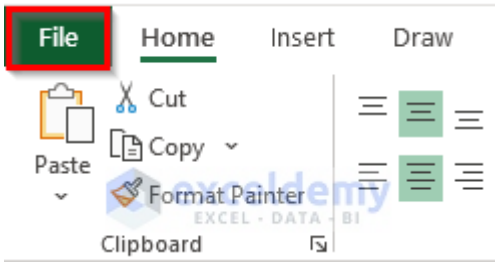
Read More: How to Analyze Raw Data in Excel

Step 2: Enable Data Analysis Feature

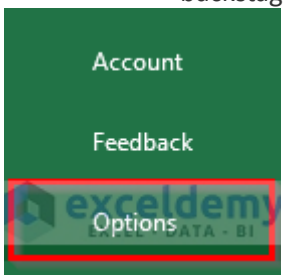
Excel Data Analysis feature gives us more ability to comprehend our data. Additionally, analyzing data offers superior visual insights, statistics, and structures. But this feature remains disabled by default.

To enable the tool, we have to follow the sub-steps below.

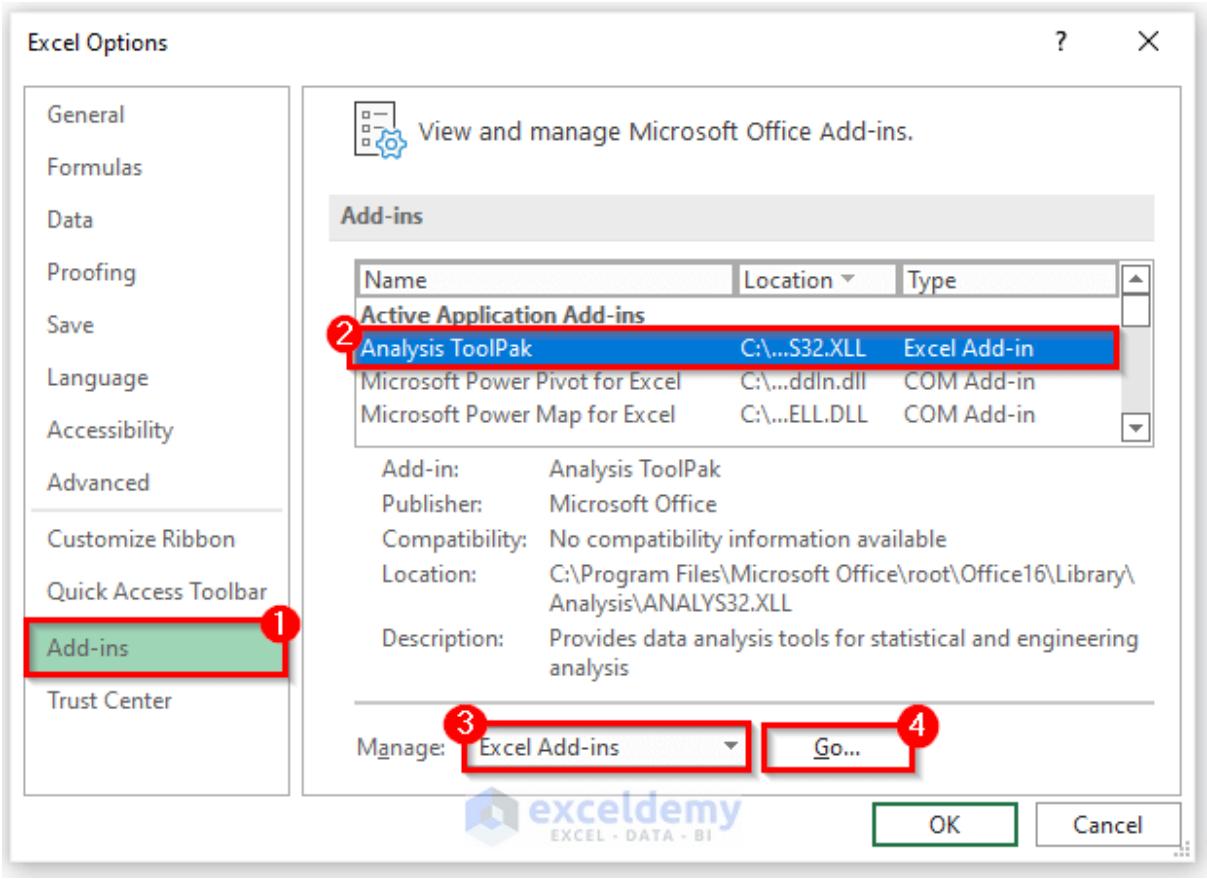
- In the first place, go to the **File** tab from the ribbon.



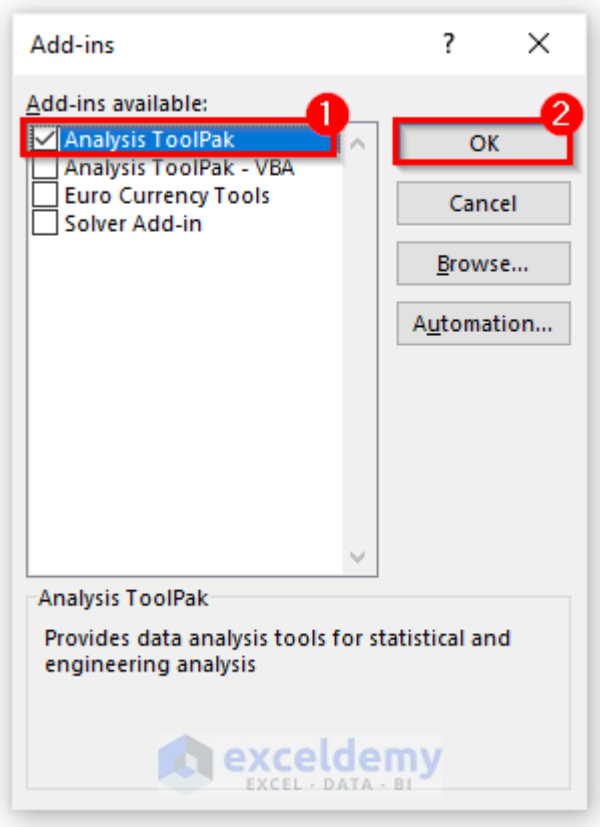
- This will take to the backstage of the excel menus.
- Then, go to the **Options** menu which you found in the bottom left corner of the excel backstage menu.



- Further, the **Excel Options** dialog box will appear.
- Afterward, go to **Add-ins**, and under the **Add-ins** option select the **Analysis Toolpak**.
- Subsequently, choose the **Excel Add-ins** from the **Manage** drop-down menu.
- Furthermore, click on the **Go** button.



- The **Add-ins** window will come up.
- Consequently, checkmark the **Analysis Toolpak**.
- Finally, click on the **Ok** button to complete the procedures.
- In this way, we will get the **Data Analysis** button under the **Data** tab.

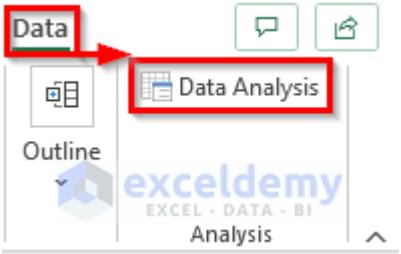


Read More: How to Analyze Large Data Sets in Excel

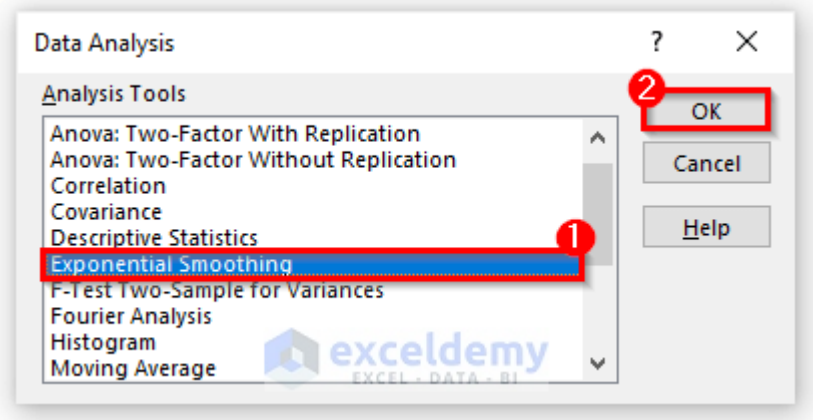
Step 3: Execute Statistical Analysis

Statistical analysis is the gathering and evaluation of data to find relationships and correlations. It belongs to the data analytics feature. Now, we will prepare the essential information to carry out **Statistical Analysis**. Follow the procedure below as a result.

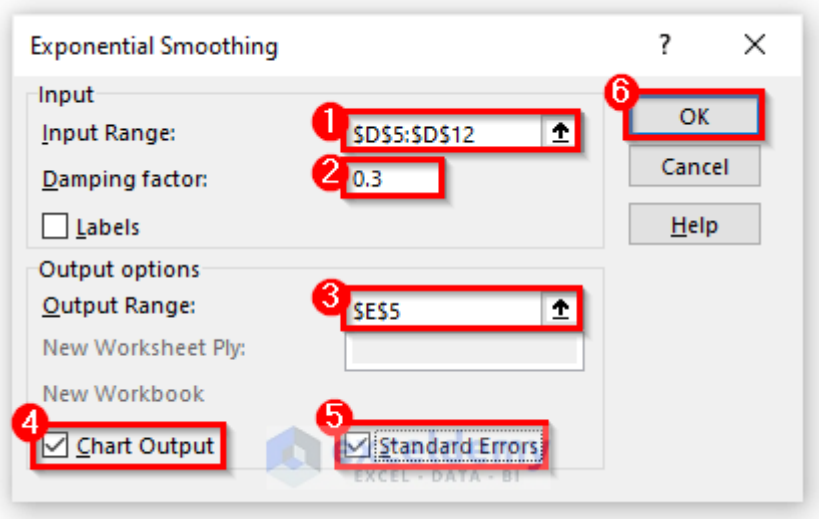
- To begin with, go to the **Data** tab from the ribbon.
- Then, click on the **Data Analysis** tool under the **Analysis** group.



- As a result, the **Data Analysis** dialog box will pop up.
- Now, scroll down a little bit and select **Exponential Smoothing**.
- Subsequently, click on the **OK** button.



- Accordingly, this will display the **Exponential Smoothing** dialog box.
- Next, select the cell range in the **Input Range** field. In this case, we select the range **$D$5:$D$12** which is the **Revenue** column.
- Further, specify the **Damping factor** as per requirement.
- Then, select the range **$E$5** in the **Output Range** field.
- Furthermore, checkmark the **Chart Output** and **Standard Errors** boxes.
- Finally, click **OK** to finish the procedures.



Read More: How to Analyze Text Data in Excel

Final Output to Analyze Time Series Data in Excel

After clicking the **OK** button on the **Exponential Smoothing**, this will return us to the excel workbook with two new columns and a graph chart.

- The **Smoothed Level** and **Standard Error** columns represent the outcomes of the statistical analysis.
- If we take a closer look at the smoothed levels, we can see that the column contains the following formula.

```
=0.7*D6+0.3*E6
```

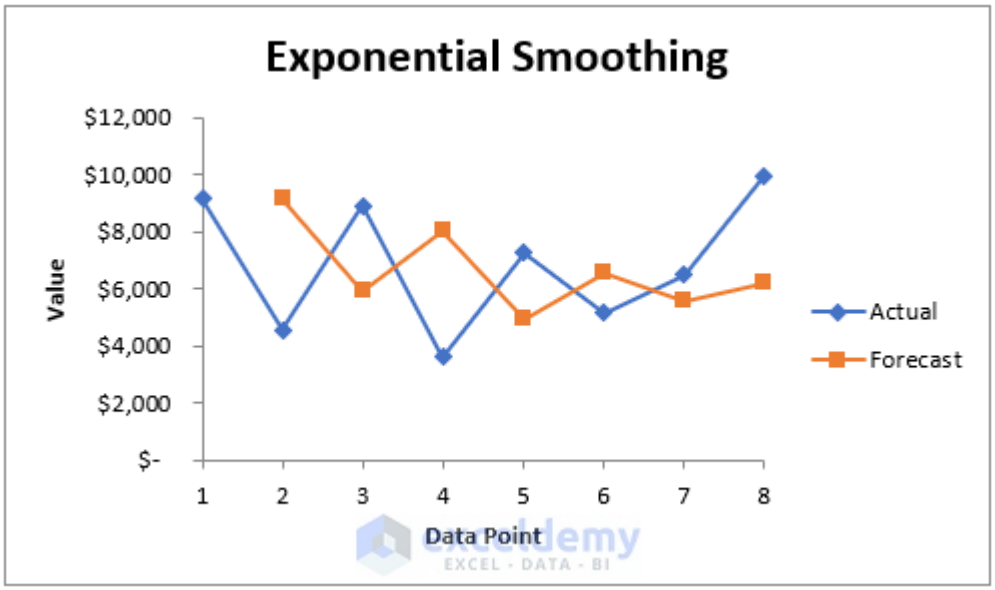
- If we examine the standard errors, we can see that the formula of the combination of the **SQRT** and **SUMXMY2** functions, in the column is as follows.

```
=SQRT(SUMXMY2(D6:D8,E6:E8)/3)
```

Analysis of Time Series Data

Year	Quarter	Revenue	Smoothed Levels	Standard Errors
2020	1	$ 9,150	#N/A	#N/A
	2	$ 4,560	$ 9,150	#N/A
	3	$ 8,920	5937	#N/A
	4	$ 3,615	8025.1	#N/A
2022	1	$ 7,245	4938.03	4058.545347
	2	$ 5,150	6552.909	3350.09361
	3	$ 6,480	5570.8727	2985.478535
	4	$ 9,950	6207.26181	1644.868454

- And we will also get a graphical representation of the **Revenue** and a forecast.



Read More: How to Analyze Sales Data in Excel

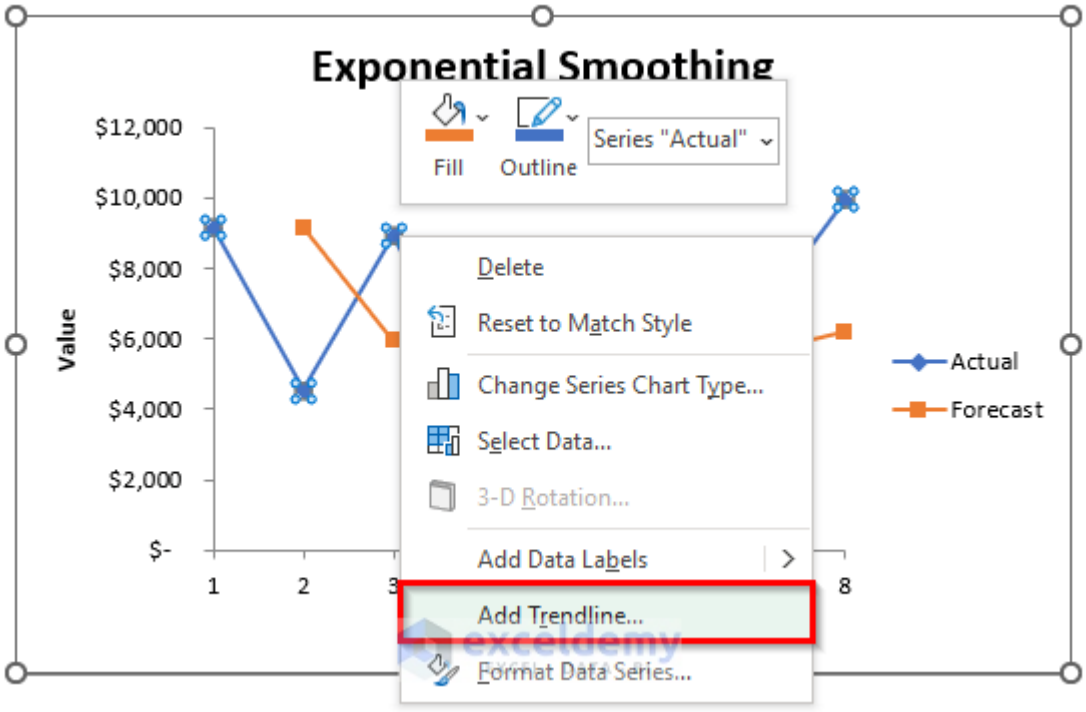
Time Series Forecasting in Excel

In order to identify recurrent temporal relationships and correlations, we frequently utilize Excel. We use it to examine time-series data. Such as sales, server use, or inventory data. We must first ensure that our time-based series data collection is complete before we can generate a forecast sheet. A chart

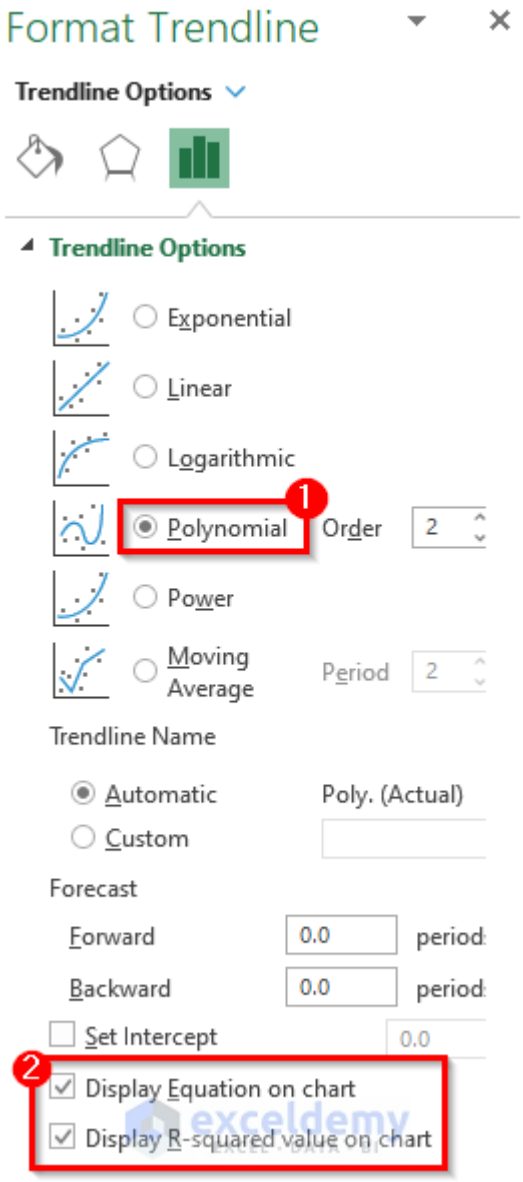
plots a time series over time. A time-series analysis may include three elements. And the elements are **level**, **trend**, and **seasonality**.

STEPS:

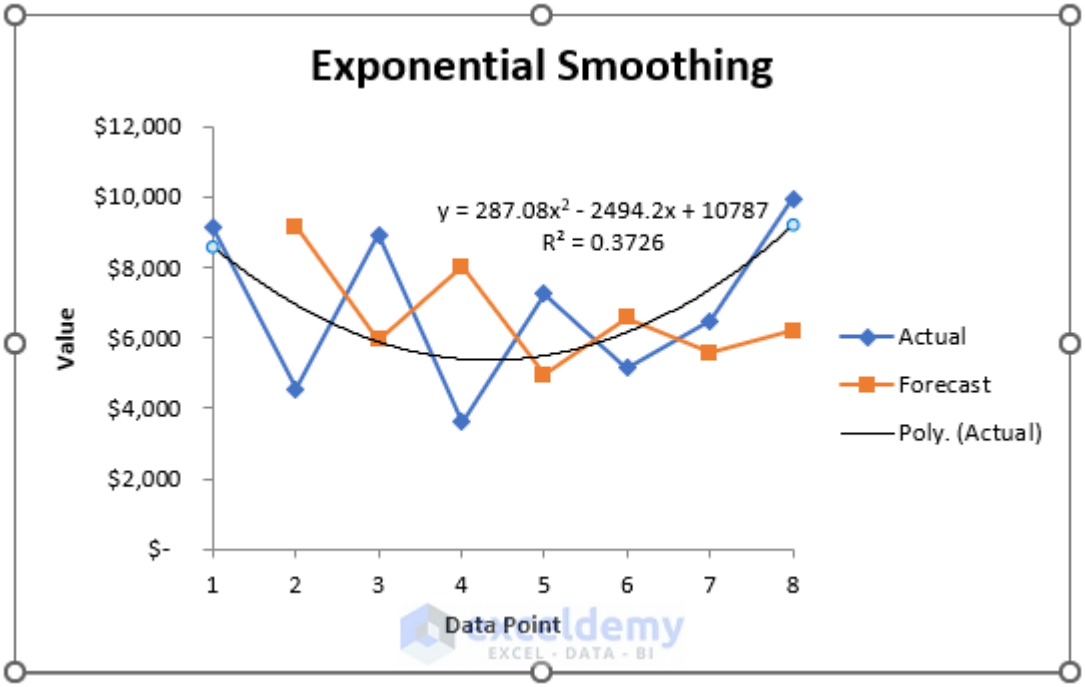
- Firstly, select the **Actual** revenue curve line.
- Secondly, right-click on the mouse, this will open the context menu.
- Thirdly, select **Add Trendline**.



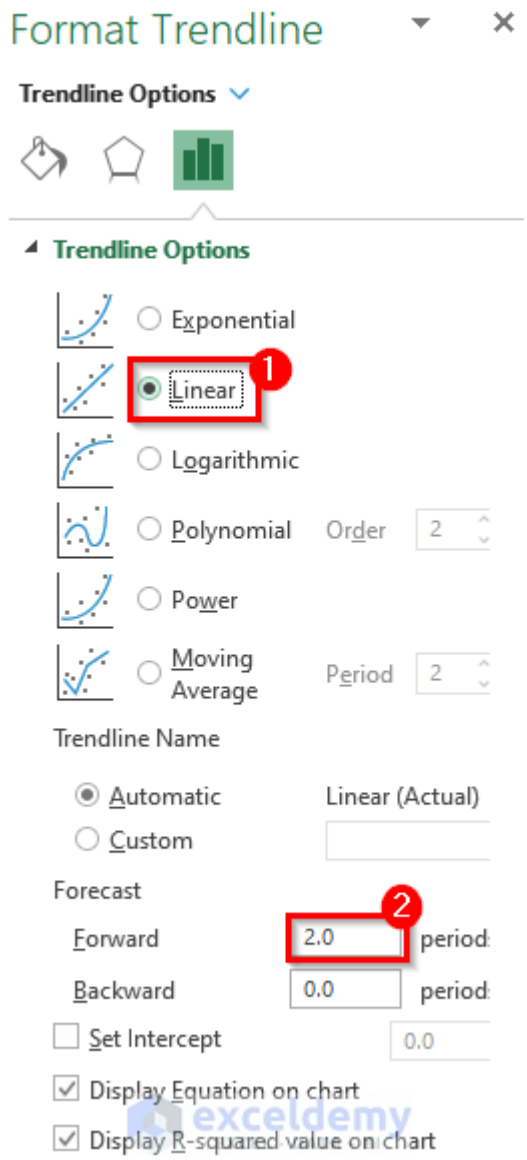
- The **Format Trendline** window will therefore show up on the right side of the spreadsheets.
- Further, check for **Polynomial** options from the **Trendline Options**.
- Checkmark the **Display Equation on chart** and **Display R-squared value on chart** boxes after that.



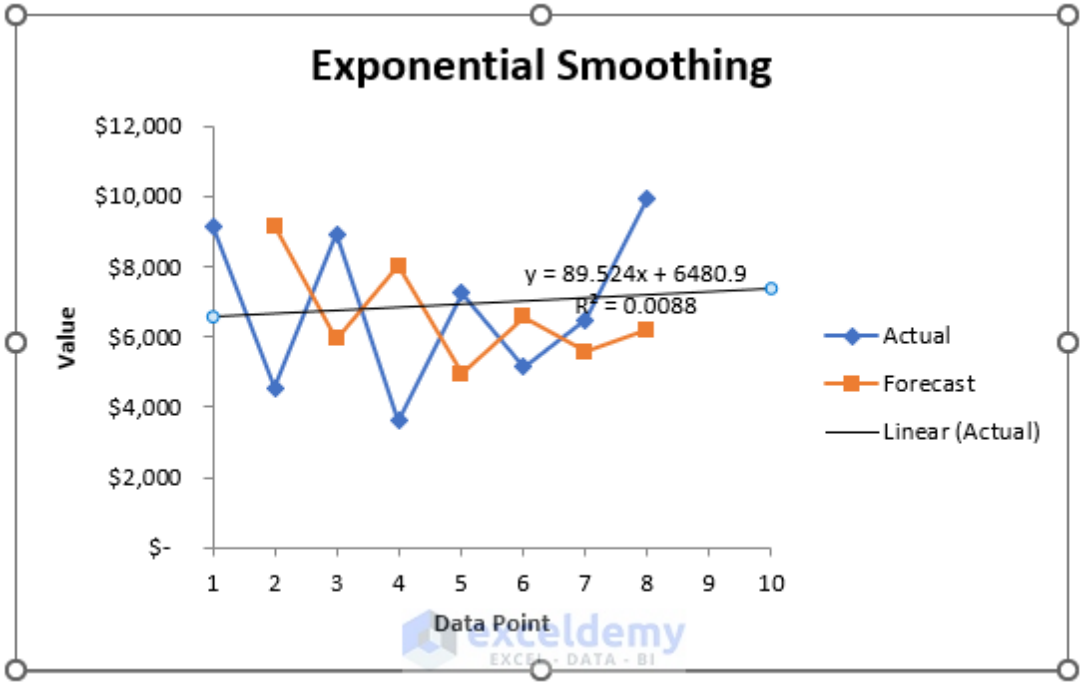
- Furthermore, view the illustration below for a better understanding.
- In the forecast models, the polynomial trend line has a lower error rate.
- As a result, the required trend line will be returned in the graph.



- Once more, choose **Linear** if you like to have a linear trend line.
- Then, mention the periods, in our case we mention the **Forward** period under the **Forcast** option.



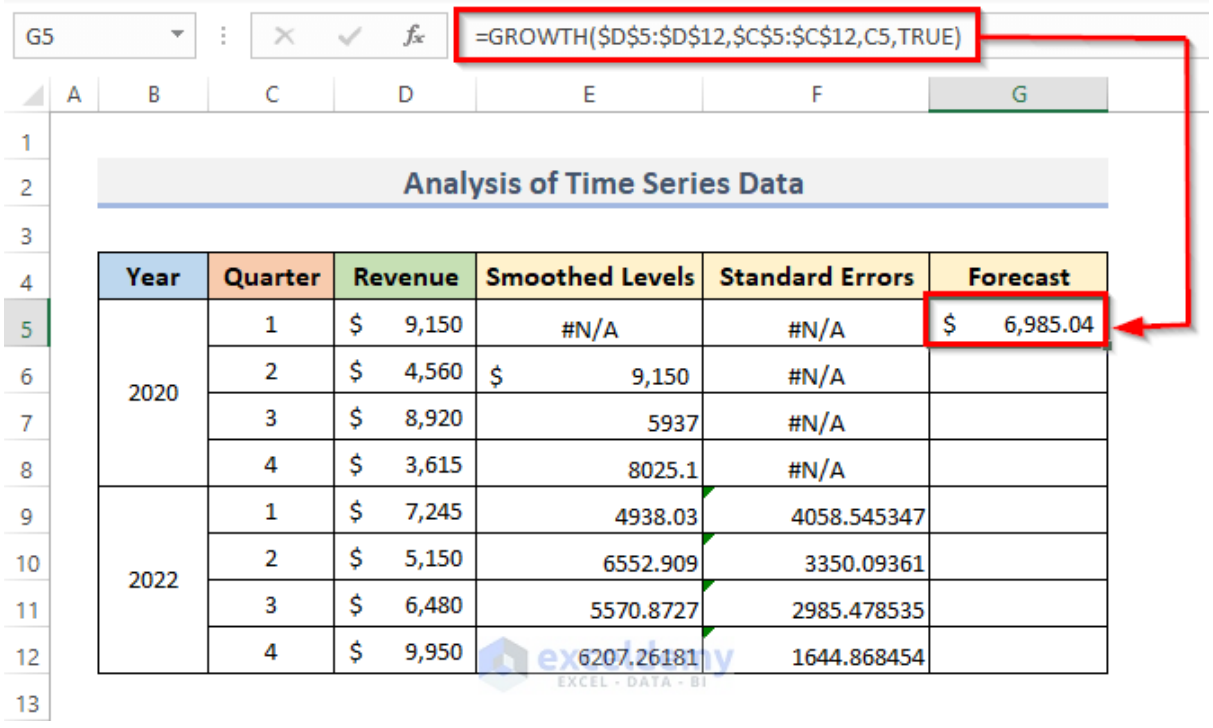
- As a result, it will display a linear trend line next to the actual data on the graph.



- Moreover, suppose we want to forecast exponential dependence. For this, we are using **the GROWTH function**. **GROWTH** delivers the y-values for a set of new x-values. And, that you specify by leveraging pre-existing x-values and y-values. We can also use this function to fit an exponential curve to already-existing x- and y-values.
- So, we insert a new column named **Forecast**.
- Thus, select the cell where you want the result of the forecast value using the **GROWTH** function.
- Then, put the formula into that selected cell.

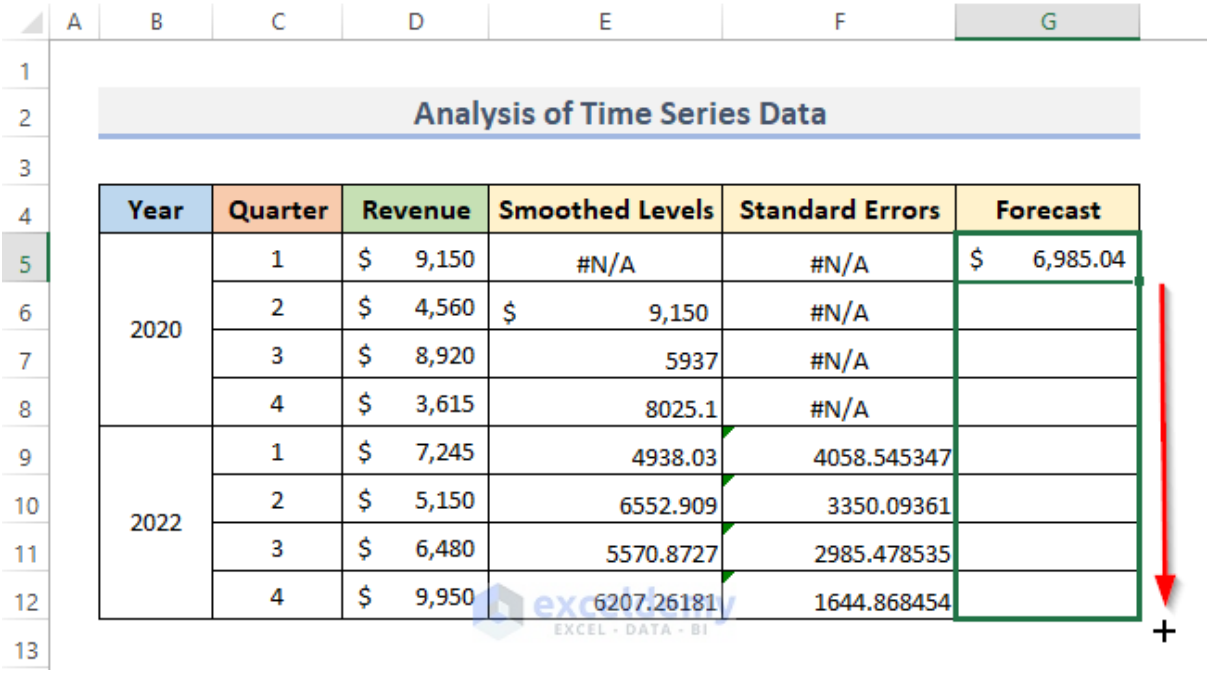
```
=GROWTH($D$5:$D$12,$C$5:$C$12,C5,TRUE)
```

- Lastly, to finish the computation, hit the **Enter** key.



Year	Quarter	Revenue	Smoothed Levels	Standard Errors	Forecast
2020	1	$ 9,150	#N/A	#N/A	$ 6,985.04
	2	$ 4,560	$ 9,150	#N/A	
	3	$ 8,920	5937	#N/A	
	4	$ 3,615	8025.1	#N/A	
2022	1	$ 7,245	4938.03	4058.545347	
	2	$ 5,150	6552.909	3350.09361	
	3	$ 6,480	5570.8727	2985.478535	
	4	$ 9,950	6207.26181	1644.868454	

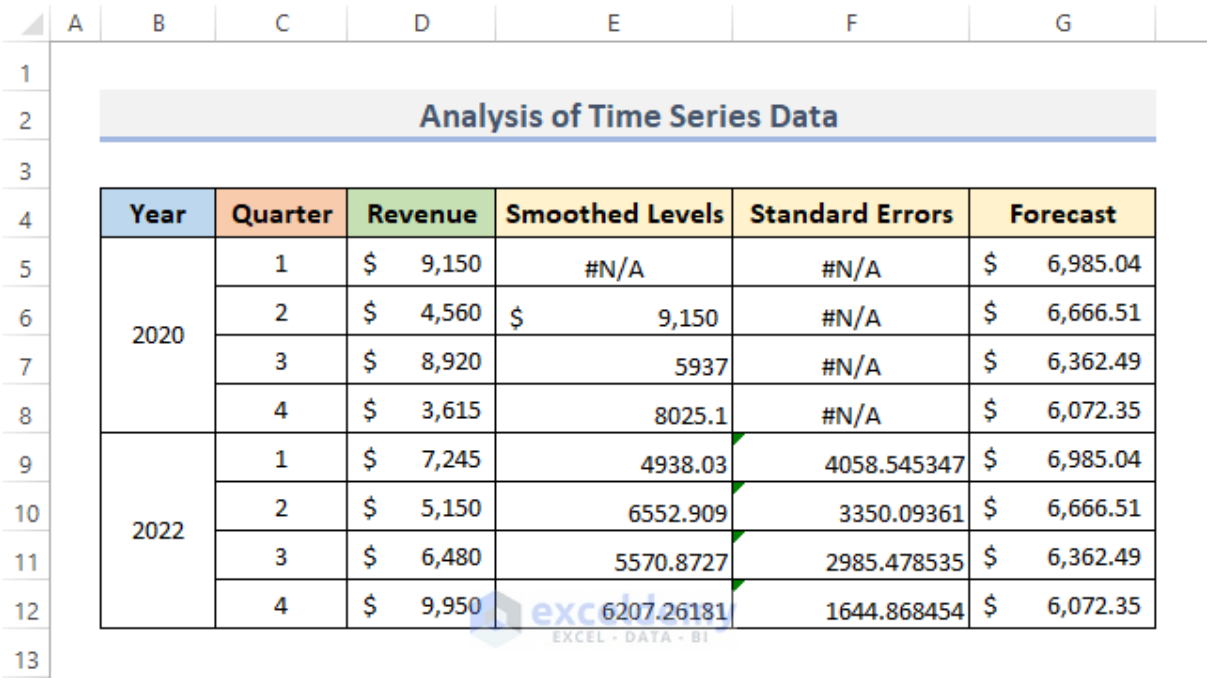
- Now, drag the **Fill Handle** down to duplicate the formula over the range. Or, to **AutoFill** the range, double-click on the plus (**+**) symbol.



Analysis of Time Series Data

Year	Quarter	Revenue	Smoothed Levels	Standard Errors	Forecast
2020	1	$ 9,150	#N/A	#N/A	$ 6,985.04
	2	$ 4,560	$ 9,150	#N/A	
	3	$ 8,920	5937	#N/A	
	4	$ 3,615	8025.1	#N/A	
2022	1	$ 7,245	4938.03	4058.545347	
	2	$ 5,150	6552.909	3350.09361	
	3	$ 6,480	5570.8727	2985.478535	
	4	$ 9,950	6207.26181	1644.868454	

- Finally, you can see the prediction for the revenue.



Analysis of Time Series Data

Year	Quarter	Revenue	Smoothed Levels	Standard Errors	Forecast
2020	1	$ 9,150	#N/A	#N/A	$ 6,985.04
	2	$ 4,560	$ 9,150	#N/A	$ 6,666.51
	3	$ 8,920	5937	#N/A	$ 6,362.49
	4	$ 3,615	8025.1	#N/A	$ 6,072.35
2022	1	$ 7,245	4938.03	4058.545347	$ 6,985.04
	2	$ 5,150	6552.909	3350.09361	$ 6,666.51
	3	$ 6,480	5570.8727	2985.478535	$ 6,362.49
	4	$ 9,950	6207.26181	1644.868454	$ 6,072.35

Practical no : 17 TIME SERIES ANALYSIS AND FORECASTING IN EXCEL WITH EXAMPLES

TIME SERIES IN EXCEL

If you capture the values of some process at certain intervals, you get the elements of the time series. Their variability is divided into regular and random components. As a rule, regular changes in the members of the series are predictable.

We will analyze time series in Excel. Example: a sales network analyzes data on sales of goods by stores located in cities with a population of fewer than 50,000 people. The period is for 2012-2015. The task is to identify the main development trend.

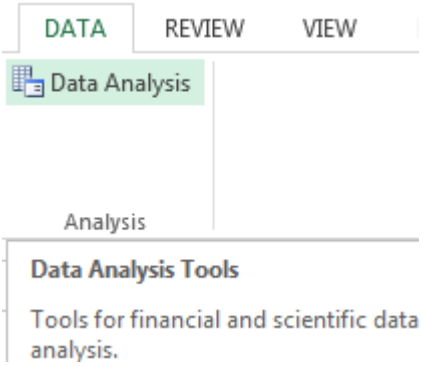
Enter the sales data in the Excel spreadsheet:

	A	B	C
1	**Year**	**Quarter**	**Sales**
2	2012	1	$165,000.00
3		2	$253,000.00
4		3	$316,000.00
5		4	$287,000.00
6	2013	1	$257,000.00
7		2	$308,000.00
8		3	$376,000.00
9		4	$351,000.00

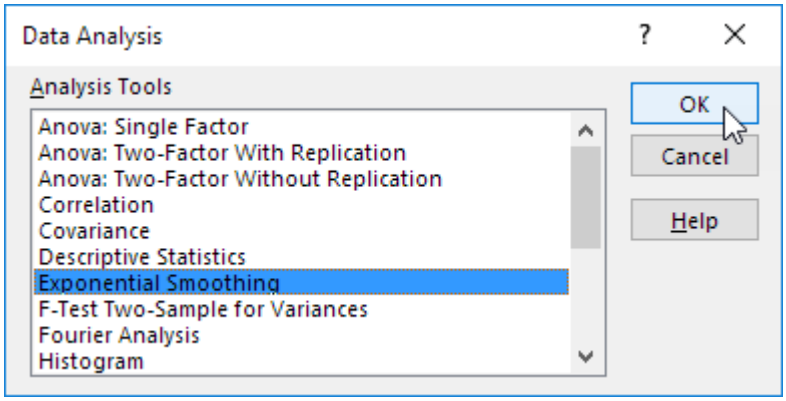
On the «DATA» tab click the «Data Analysis» button. Go to the menu if it is not visible. «Excel Options» – «Add-Ins». Click at the bottom «Go» to «Add-Ins Excel» and select « Data Analysis ».

The connection of the « Data Analysis » add-in is described here **in detail**.

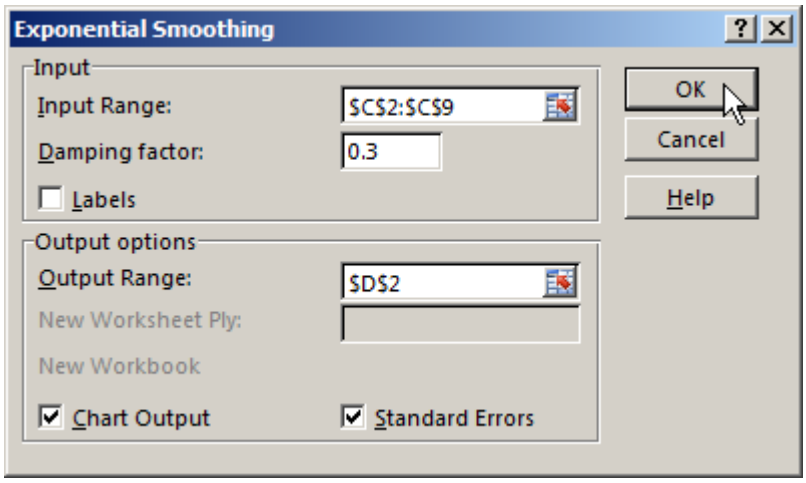
The button we need appears on the band.



Select «Exponential Smoothing» from the proposed list of tools for statistical analysis. This alignment method is suitable for our dynamic series, the values of which fluctuate strongly.

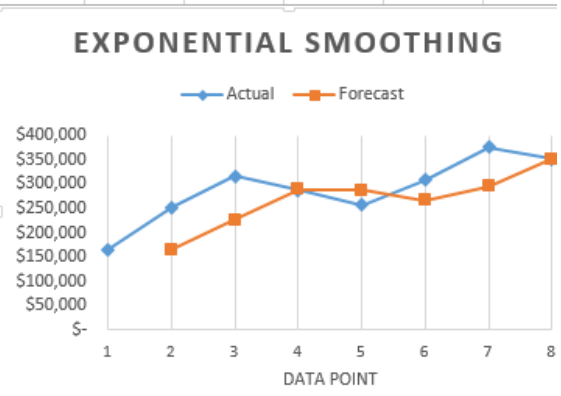


We fill the dialog box. The input interval is the range of sales values. The damping factor is the coefficient of exponential smoothing (default is 0.3). Output interval –is a reference to the upper left cell of the output range. The program will place the smoothed levels here and the will define size independently. We tick the «Chart Output», «Standard Errors».



Close the dialog box by clicking OK. Results of the analysis:

	A	B	C	D	E
1	Year	Quarter	Sales	Smoothed levels	Standard errors
2	2012	1	$165,000.00	#N/A	#N/A
3		2	$253,000.00	$ 165,000.00	#N/A
4		3	$316,000.00	226600	#N/A
5		4	$287,000.00	289180	#N/A
6	2013	1	$257,000.00	287654	72436.43742
7		2	$308,000.00	266196.2	54579.54475
8		3	$376,000.00	295458.86	29955.39913
9		4	$351,000.00	351837.658	55299.48958
10					
11					
12					

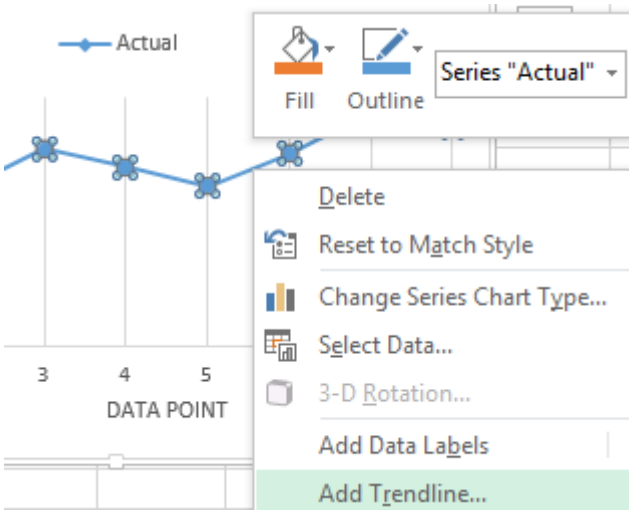


Excel uses next formula to calculate the standard errors: = SQRT(SUMXMY2('Actual value range'; 'range of forecast values') / 'size of the smoothing window'). For example, = SQRT(SUMXMY2:(C3:C5;D3:D5)/3).

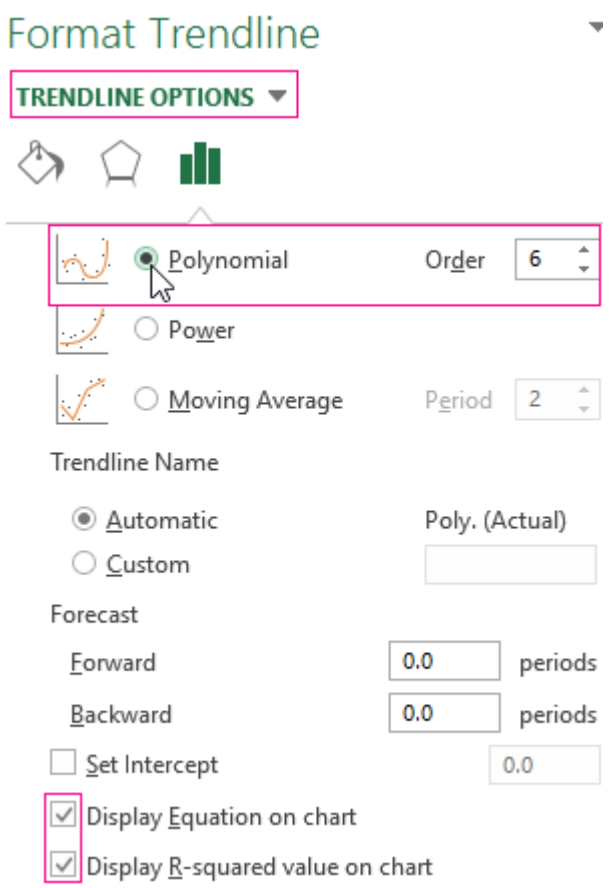
FORECASTING THE TIME SERIES IN EXCEL

We will compose the forecast of sales using the data from the previous example.

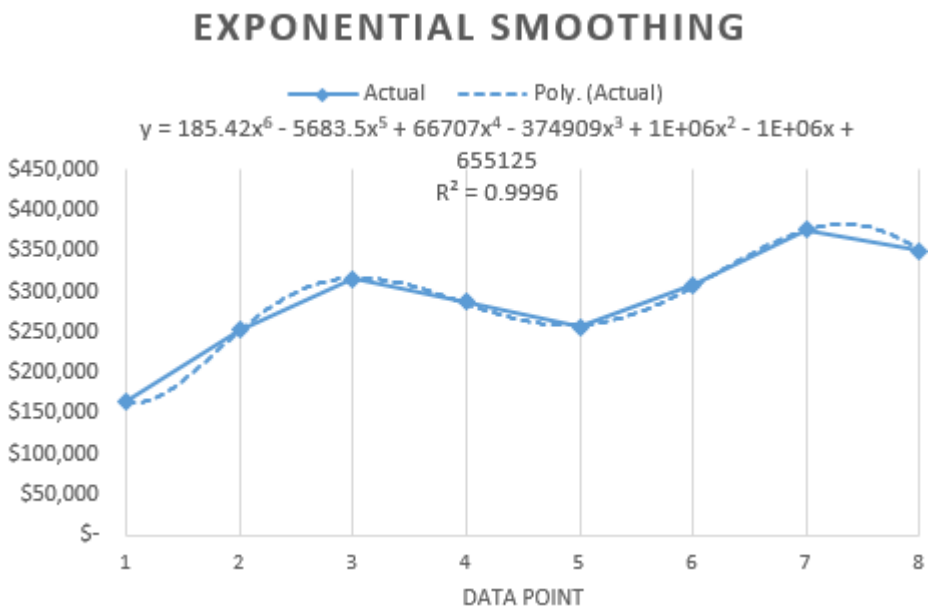
We will add a trend line (the right button on the chart - «Add Trend line») on the chart which shows the actual product sales volume.



Configure the parameters of the trend line:



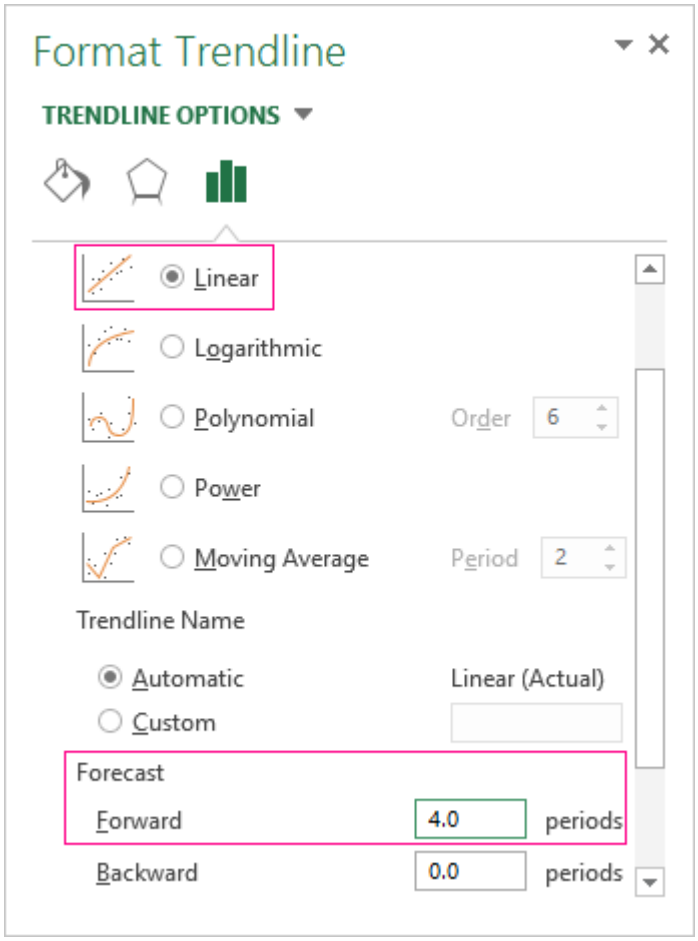
We choose a polynomial trend that minimizes the error of the forecast model.



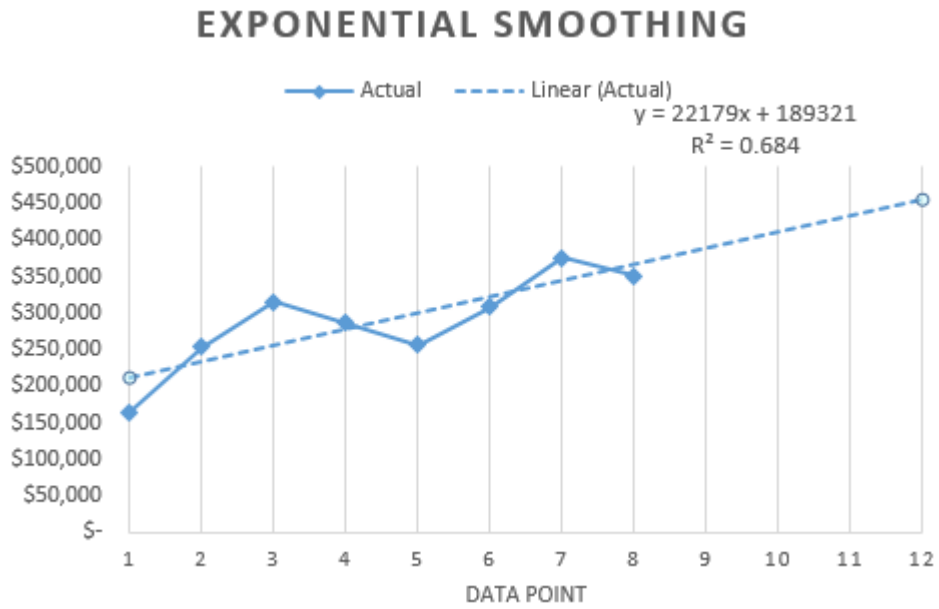
R2 = 0.9567 which means that this ratio explains 95.67% of changes in sales in process of time.

The trend equation is a model of the formula for calculating the forecast values.

Most authors recommend using a linear trend line for forecasting sales. You need to set the number of periods in the parameters to see the forecast on the chart.



We get a fairly optimistic result:



After all, there is the exponential dependence in our example. Therefore, there are more errors and inaccuracies when building a linear trend.

You can also use the function GROWTH to predict the exponential dependence in Excel.

F2 =GROWTH($C$2:$C$9,$B$2:$B$9,B2,TRUE)

	A	B	C	D	E	F
1	Year	Quarter	Sales	Smoothed levels	Standard errors	Forecast
2	2012	1	$165,000.00	#N/A	#N/A	$224,571.72
3		2	$253,000.00	$ 165,000.00	#N/A	$261,144.60
4		3	$316,000.00	226600	#N/A	$303,673.59
5		4	$287,000.00	289180	#N/A	$353,128.68
6	2013	1	$257,000.00	287654	72436.43742	$224,571.72
7		2	$308,000.00	266196.2	54579.54475	$261,144.60
8		3	$376,000.00	295458.86	29955.39913	$303,673.59
9		4	$351,000.00	351837.658	55299.48958	$353,128.68

For linear dependence, use the TREND function.

You cannot use any one method when making forecasts: the probability of large deviations and inaccuracies is large.