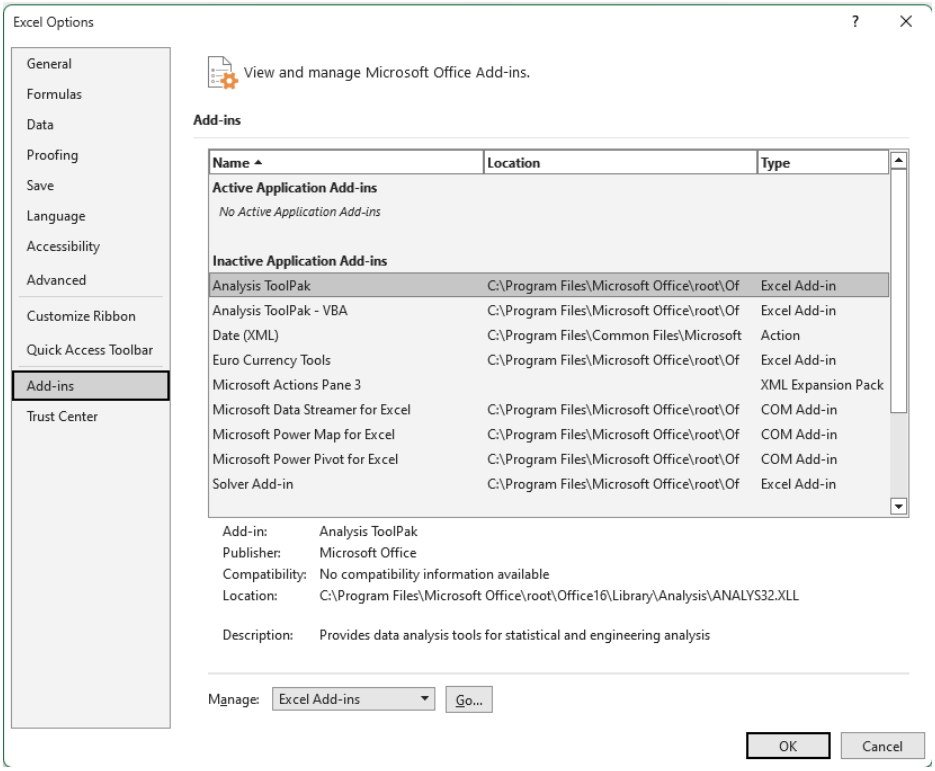
Статистика в Excel

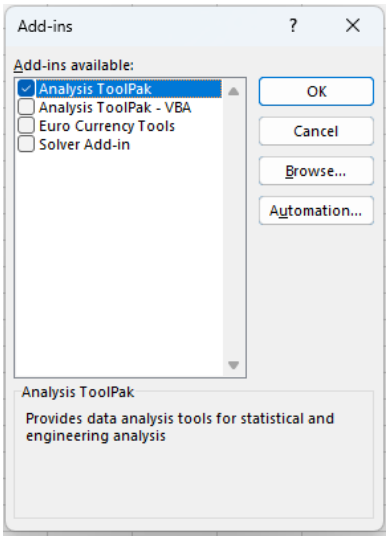
Excel разполага с допълнителен пакет (add-in) Analysis ToolPak, който преди да се използва трябва да се инсталира.

1. Инсталиране на Analysis ToolPak

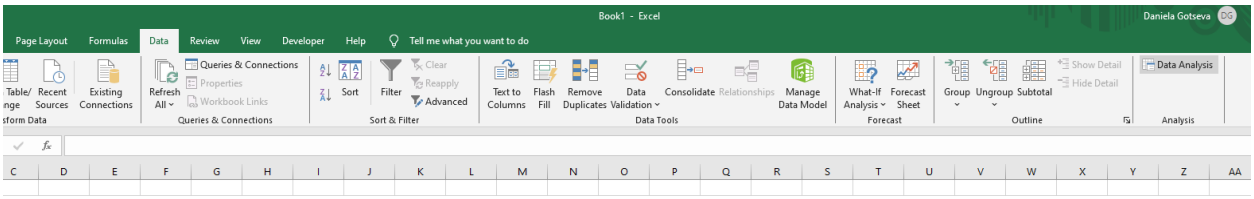
От менюто се избира File->Options->Add ins. Показва се следния диалогов прозорец:



Оттук се избира Analysis ToolPak и се натиска бутона Go. Следва избор на пакета, който да се инсталира:



И се избира OK. Новите функции се намират в меню Data, където има добавена нова група Analysis. В нея има Data Analysis:

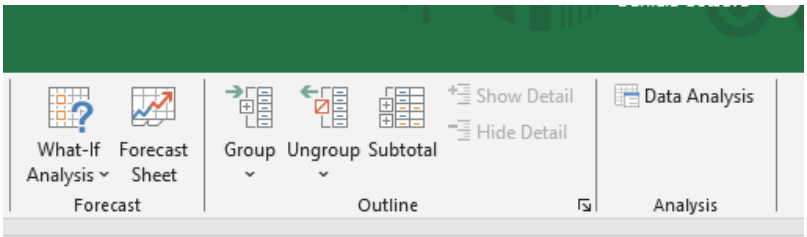


2. Изчисляване на Регресия (Regression)

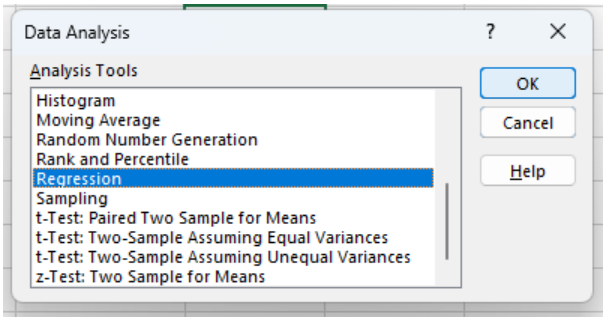
Нека е даден следния пример. Дадена е таблица с информация за цената на продукт и разхода за неговата реклама като входни данни и продаденото количество от него като изход. Трябва да се определи каква е връзката между входните и изходната данни. За целта се използва регресия от статистиката:

	A	B	C
1	Quantity Sold	Price	Advertising
2	8500	€ 2.00	€ 2,800.00
3	4700	€ 5.00	€ 200.00
4	5800	€ 3.00	€ 400.00
5	7400	€ 2.00	€ 500.00
6	6200	€ 5.00	€ 3,200.00
7	7300	€ 3.00	€ 1,800.00
8	5600	€ 4.00	€ 900.00

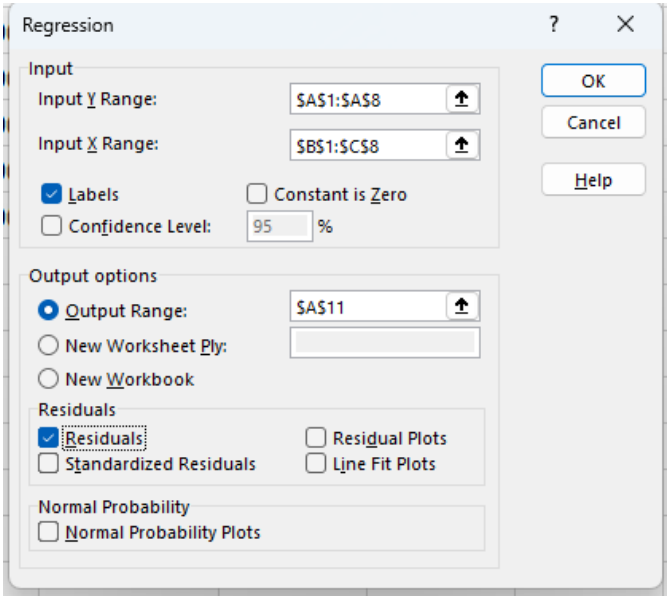
От Data таб в група Analysis се избира Data Analysis:



Следва избор на Regression и бутон OK:



В полето Input Y Range се посочва областта, която заема изходната данна: A1:A8 В полето Input X Range се посочва областта на входните данни: B1:C8. Отбелязва се, че данните имат имена на първия ред: Labels се чеква. Като Output Options се избира Output Range и се задава адрес на клетката, откъдето започва изхода: A11. От групата Residuals се чеква Residuals. Въвеждането приключва с бутона OK:



Excel генерира SUMMARY OUTPUT. Ето някои от по-важните показатели (оцветяването е от автора за по-голяма четливост):

- R Square – показва доколко изхода е зависим от входа. Колкото е по-близко до 1, толкова е по-голяма връзката между данните.

11	SUMMARY OUTPUT	
12		
13	*Regression Statistics*	
14	Multiple R	0.98068
15	R Square	0.96174
16	Adjusted R Squ	0.9426
17	Standard Error	310.524
18	Observations	7
19		

- F и P-values – F показва дали множеството входни данни е правилно избрано. Трябва да бъде <=0.05. В този случай трябва да се премахне данната с висока P-стойност (>0.05) и да се генерира нова регресия. Този процес се повтаря докато F стане под 0.05

20	ANOVA								
21		*df*	*SS*	*MS*	*F*	Significance F			
22	Regression	2	9694299.6	4847150	50.26854403	0.001464128			
23	Residual	4	385700.43	96425.1					
24	Total	6	10080000						
25									
26		Coefficients	*tandard Erro*	*t Stat*	P-value	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
27	Intercept	8536.213882	386.91175	22.0624	2.49812E-05	7461.974654	9610.45	7461.97	9610.45
28	Price	-835.7223514	99.653045	-8.38632	0.001106064	-1112.40356	-559.041	-1112.4	-559.041
29	Advertising	0.592228496	0.1043468	5.67558	0.004755309	0.302515325	0.88194	0.30252	0.88194
30									

- Коефициенти – линията на регресия се определя по формулата:

$$y = Quantity\ Sold = Coefficient\ Intercept - Price * Coefficient\ Price + Advertising * Coefficient\ Advertising$$

Т.е. при всяко увеличение на цената, броят на продадените продукти намалява с 836 (835.72235137828). За всяко увеличение на рекламата, броят на продадените продукти нараства с около 1 (0.59222849551644). Тези коефициенти могат да се използват за предсказване на бъдещото развитие. Например, ако цената е 4 евро, а рекламата в 3000 евро, очакваните продажби са:

$$8536.21388243109 - 835.72235137828 * 4 + 0.59222849551644 * 3000 = 6970$$

- Residuals (остатъци) – показват колко далеч от действителните данни са изчислените. Действителните данни се вземат от таблицата, а изчислените се получават от формулата за коефициентите и се намира тяхната разлика.

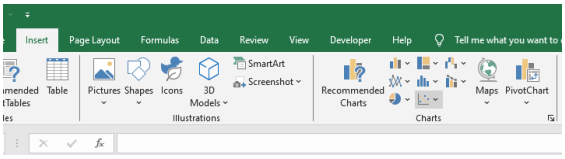
33	RESIDUAL OUTPUT		
34			
35	*Observation*	*Predicted Quantity Sold*	Residuals
36	1	8523.008967	-23.008967
37	2	4476.047825	223.95218
38	3	6265.938227	-465.93823
39	4	7160.883427	239.11657
40	5	6252.733311	-52.733311
41	6	7095.05812	204.94188
42	7	5726.330123	-126.33012
43			

Например, за Продадено количество 8500 се получава:

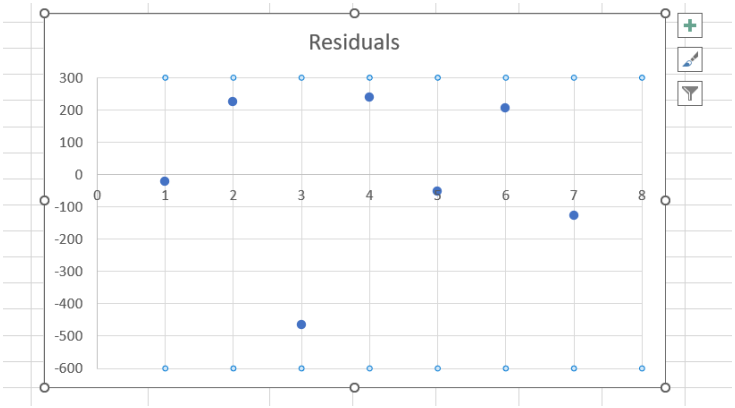
$$Predictive\ Quetity\ Sold = 8536.21388243109 - 835.72235137828 * 2 + 0.59222849551644 * 2800 = 8523.00896712056$$

$$Residual = 8500 - 8523.00896712056 = -23.0089671205569$$

- Scatter Plot (точкова графика) на остатъците. Избира се колоната на остатъците, след което се избира менюто Insert->Carts->Scatter:



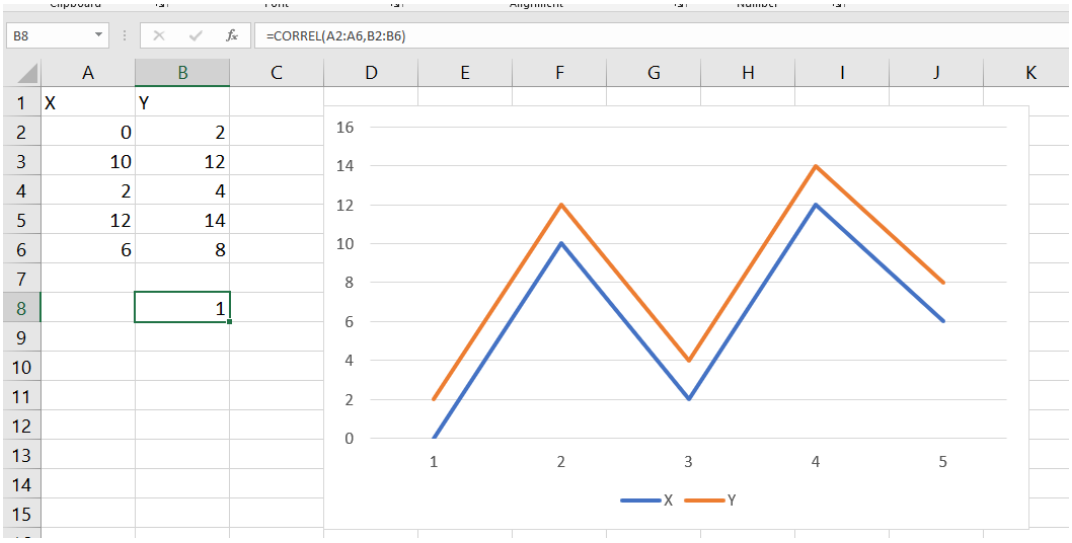
Резултатът е следния:



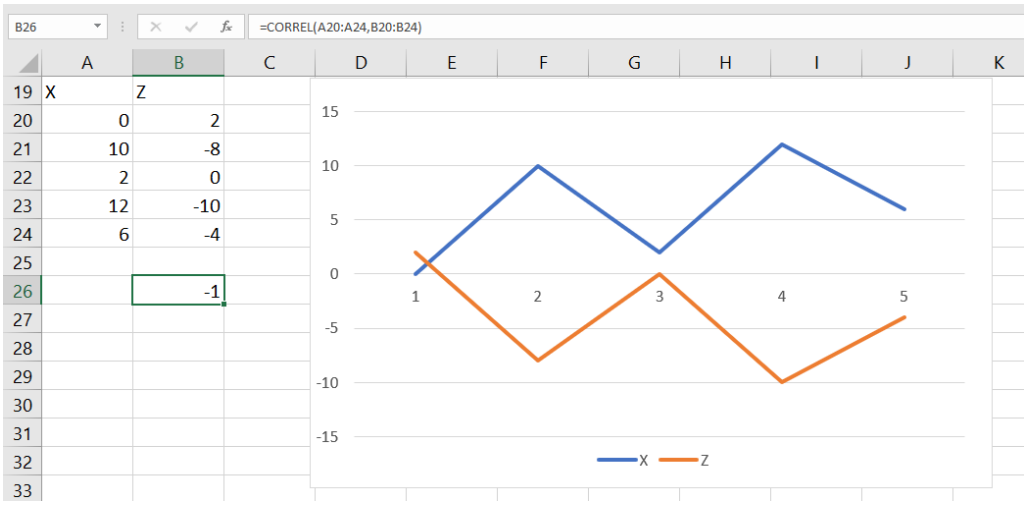
3. Определяне на Корелация (Correlation)

Коефициентът на корелация (стойност между -1 и +1) показва колко силно две данни си влияят една на друга. Тук може да се използва функцията CORREL от Excel или Analysis ToolPak.

- Стойност +1 показва перфектна положителна корелация, т.е. при увеличение на X се увеличава и Y и обратно. Например:

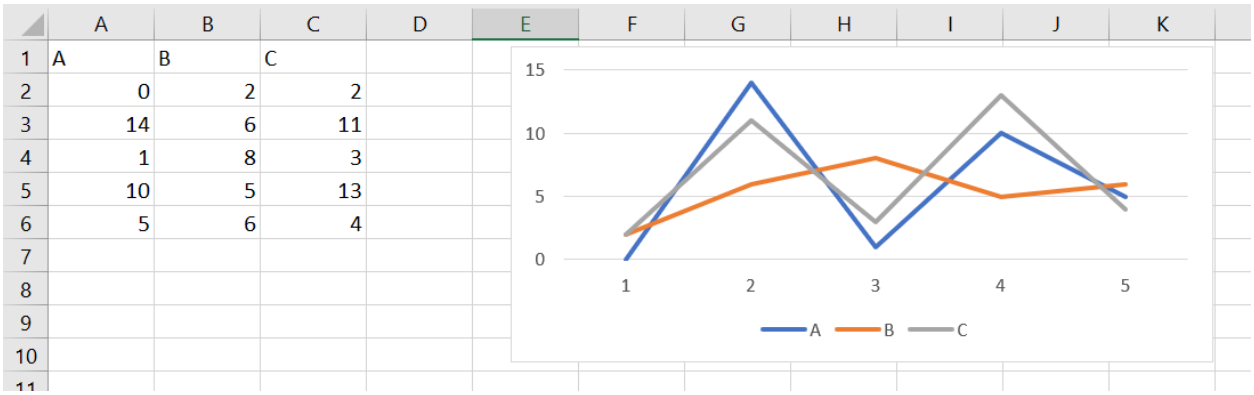


- Стойност на корелация -1 показва перфектна отрицателна връзка, т.е. ако данна X нараства, данна Z намалява и обратно.

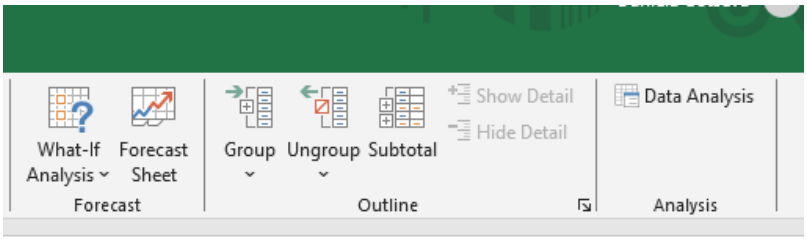


- Коефициент на корелация близък до нула показва, че между данните няма връзка

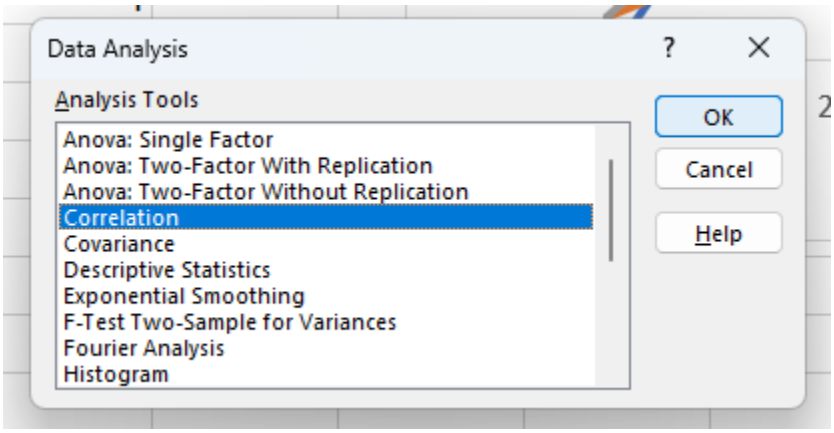
Ето как може да се намери корелацията с използване на Analysis ToolPak. Нека са дадени следните приемни данни:



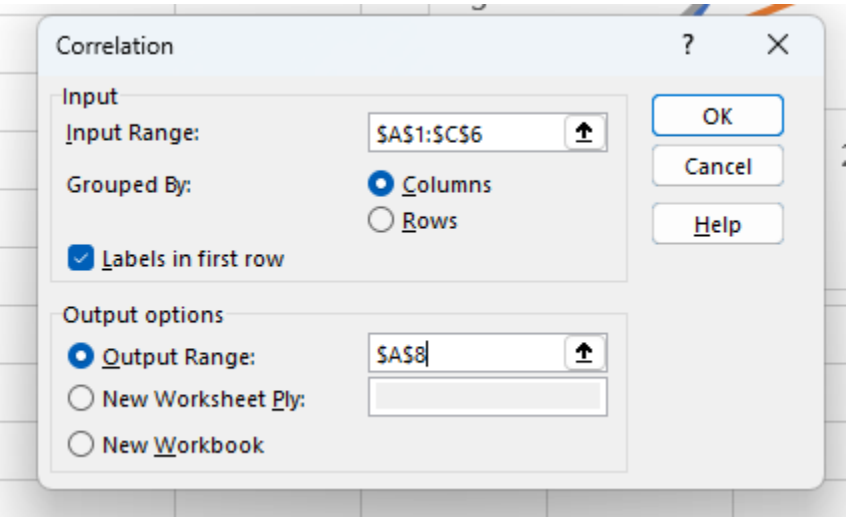
От Data таб в група Analysis се избира Data Analysis:



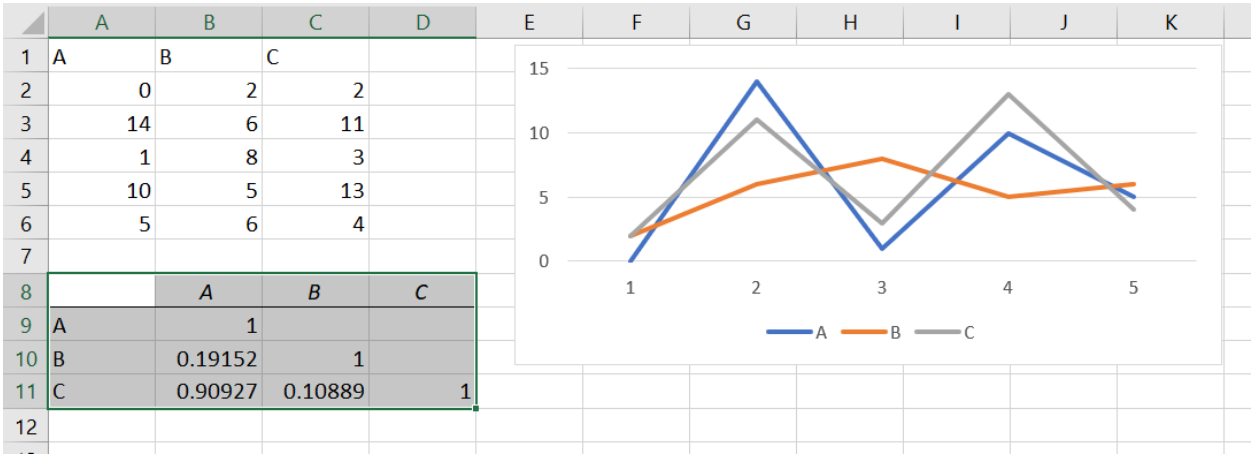
Следва избор на Correlation и бутон ОК:



Като Input Range се избират адресите A1:C6. Отбелязва се Labels in first row, а като Output Options се избира Output Range и се посочва клетка A8:



След избор на OK се получава:



	A	B	C
A	0	2	2
	14	6	11
	1	8	3
	10	5	13
	5	6	4

	A	B	C
A	1		
B	0.19152	1	
C	0.90927	0.10889	1

Изводът, който може да се направи, е че данни A и C са положително свързани (0.91). Данните A и B не са свързани (0.19). Данни B и C също нямат връзка помежду си (0.11)

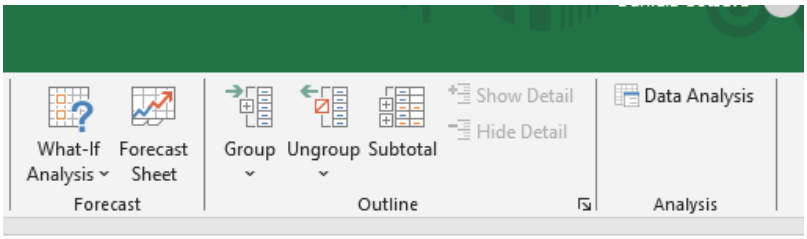
4. Създаване на Хистограма (Histogram)

Дадени са данните за броя на студентите (колона A) и номерата на контейнерите (C4:C8):

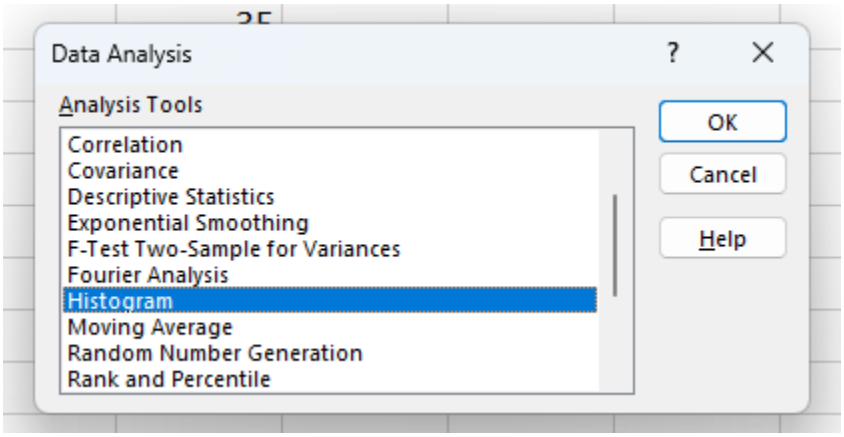
	A	B	C
1	Number of students		
2	22		
3	29		
4	40		20
5	30		25
6	48		30
7	24		35
8	21		40
9	19		
10	24		
11	22		
12	25		
13	52		
14	35		
15	40		
16	31		
17	37		
18	21		
19	23		

Да се построи хистограма.

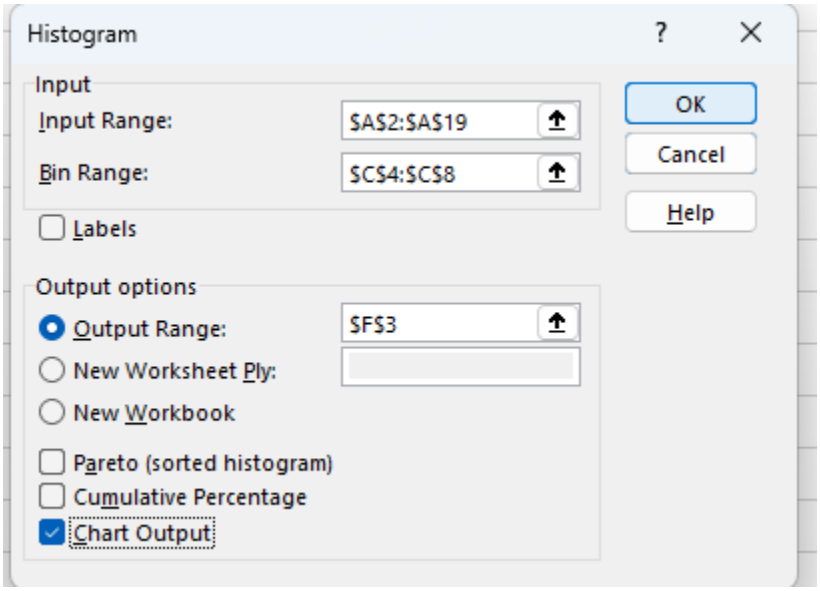
За целта от Data таб в група Analysis се избира Data Analysis:



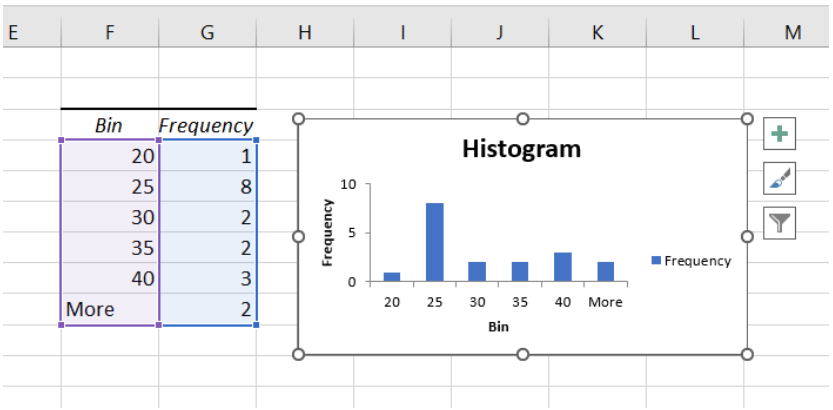
Следва избор на Histogram и бутон OK:



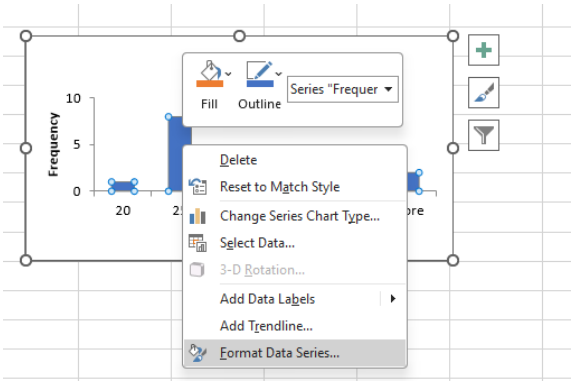
Като Input Range се избират A2:A19. За Bin Range се посочва C4:C8. В Output Options се маркира Output Range като се избира клетка F3. Следва избор на опция Chart Output.



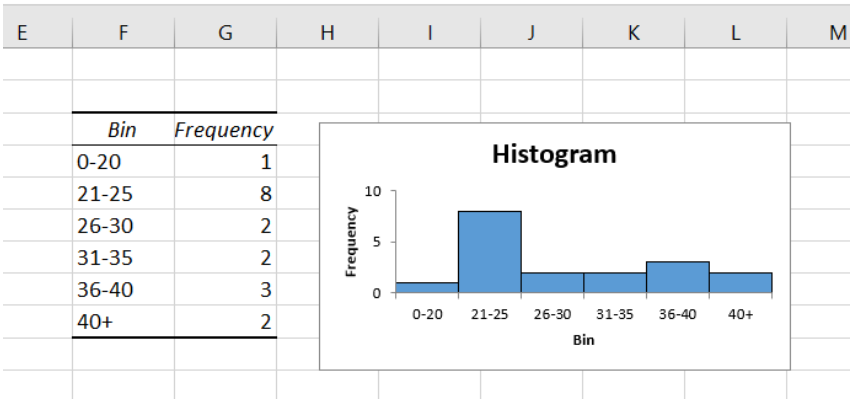
Следва избор на ОК и се получава:



Легендата може да се изтрие като се избере и се ползва Delete. Променят се надписите на контейнерите, за да показват диапазон. За да се премахнат празните позиции между правоъгълниците, се ползва десен бутон на мишката->Format Data Series, след което Gap width се задава 0%. За да се добавят линии на правоъгълниците, се избира десен бутон на мишката->Format Data Series, след което иконата Fill & Line->Border и се избира цвят:



Ето финалния резултат:

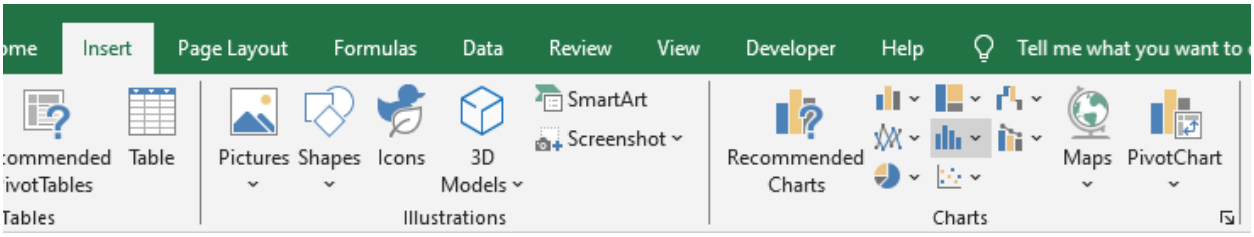


Bin	Frequency
0-20	1
21-25	8
26-30	2
31-35	2
36-40	3
40+	2

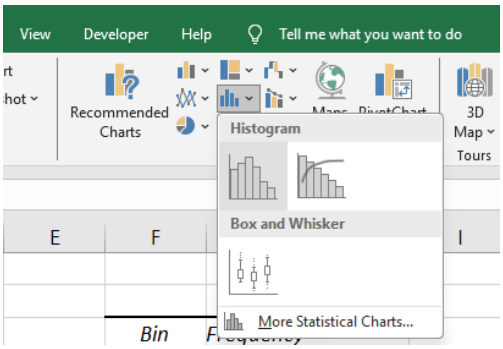
За версии на Excel 2016+ може да се използва Histogram Chart Type. За целта се избира диапазона A1:A19:

	A	B
1	Number of students	
2	22	
3	29	
4	40	
5	30	
6	48	
7	24	
8	21	
9	19	
10	24	
11	22	
12	25	
13	52	
14	35	
15	40	
16	31	
17	37	
18	21	
19	23	
20		

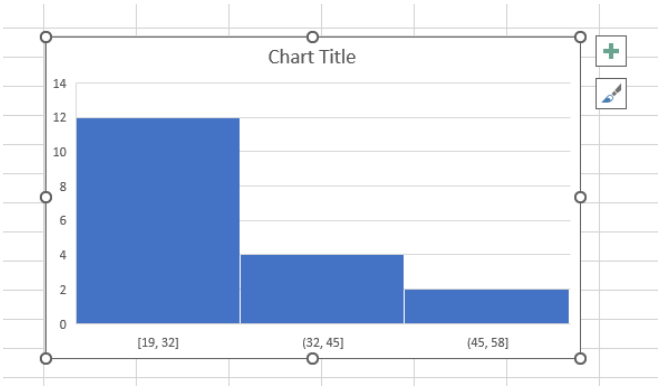
В таба Insert се посочва групата Charts и оттам иконата Histogram:



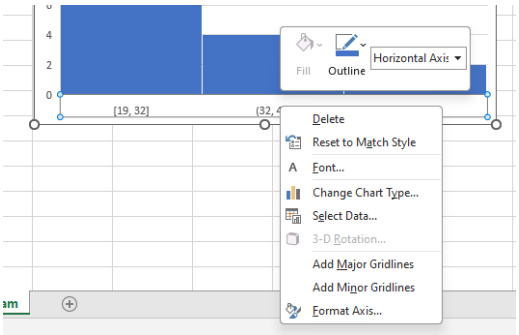
Следва избор на Histogram:



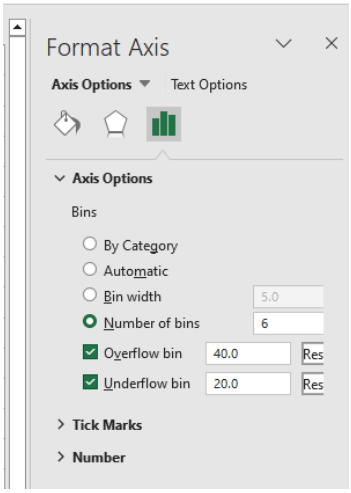
Резултатът е създаване на хистограма с три контейнера от Excel, който използва правилото на Scott за изчисление на броя на контейнерите и ширината им.



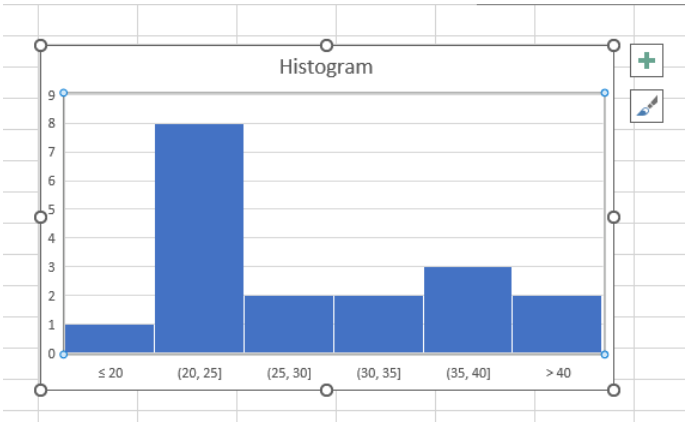
Десен бутон на мишката върху хоризонталната ос, откъдето се избира Format Axis:



Оттук може да се зададе брой на контейнерите (Number of bins), начало на последния контейнер (Overflow bin) и горна граница на първия (Underflow bin):



Ето и резултата:



5. sПроверка на хипотези (t-Test)

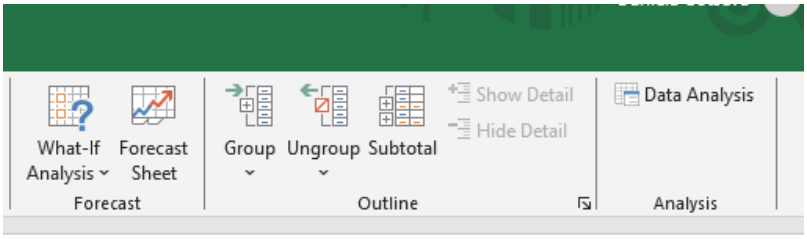
Показаният по-долу пример представя как се проверяват хипотези в Excel. За целта се използва t-Test, който проверява за нулева хипотеза (null hypothesis). Нулевата хипотеза означава еднаквост на две популации. В таблица са въведени данните от часовете, прекарани е обучение на 6 жени и 5 мъде студенти. Тестваме две хипотези:

$$H_0: \mu_1 - \mu_2 = 0$$

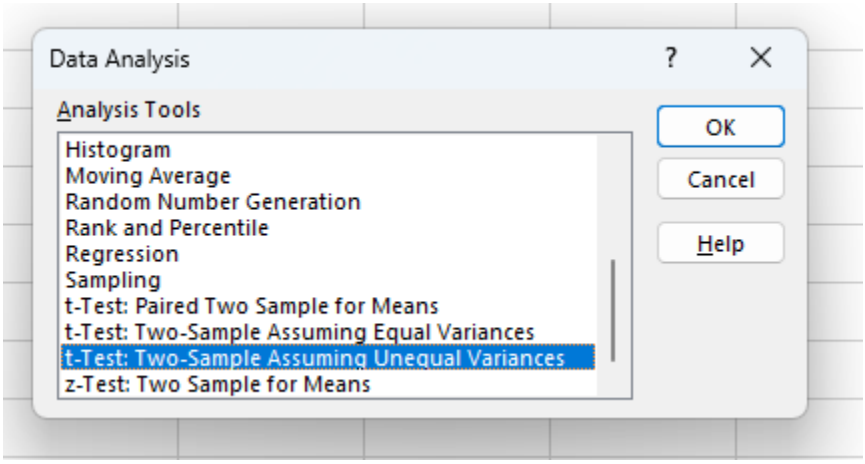
$$H_1: \mu_1 - \mu_2 \neq 0$$

	A	B
1	Female	Male
2	26	23
3	25	30
4	43	18
5	34	25
6	18	28
7	52	
8		

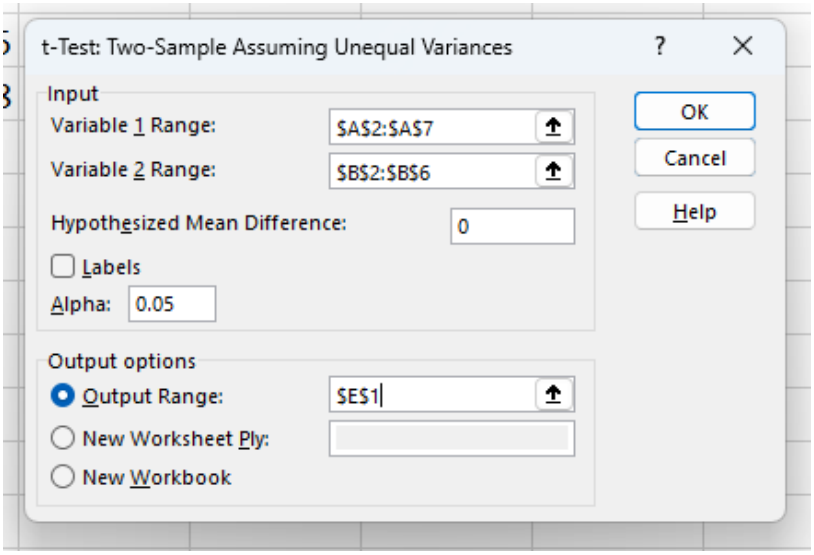
Като начало трябва да се изпълни F-Test, за да се провери дали вариациите на двете популации са еднакви, ако това не е ясно. В този случай това не е така, поради което може да се премине към изпълнение на f-Test. За целта от Data таб в група Analysis се избира Data Analysis:



Следва избор на t-Test: Two Sample Assuming Unequal Variances и бутон OK:



В полето Variable 1 Range се избира A2:A7. В полето Variable 2 Range се посочва B2:B6. За проверка на първата хипотеза в полето Hypothesized Mean Difference се въвежда нула. Следва избор на опцията Output Range в полето Output Options и се въвежда адреса E1:



Получава се следния резултат:

D	E	F	G	H
	t-Test: Two-Sample Assuming Unequal Variances			
		Variable 1	*Variable 2*	
	Mean	33	24.8	
	Variance	160	21.7	
	Observations	6	5	
	Hypothesized Mean Difference	0		
	df	7		
	t Stat	1.47261		
	P(T<=t) one-tail	0.09217		
	t Critical one-tail	1.89458		
	P(T<=t) two-tail	0.18434		
	t Critical two-tail	2.36462		

Накрая се проверява two-tail неравенство. Ако е изпълнено:

$$t\ Stat < -t\ Critical\ two - tail$$

или

$$t\ Stat > t\ Critical\ two - tail$$

Нулевата хипотеза се отхвърля. В посочения пример това не е така, защото:

$$-2.365 < 1.473 < 2.365$$

Следователно, нулевата хипотеза не се отхвърля. Едновременно с това , очакваната разлика между двете популации (33 – 24.8) не е достатъчно убедителна, за да се направи заключение, че средния брой часове, прекарани в учене, между студентите жени и мъже се различават значително.