

Jedną z funkcji MS Excel jest przetwarzanie danych oraz analiza statystyczna. Pliki potrzebne do ćwiczenia znajdują się tutaj:

[http://jacenty.kis.p.lodz.pl/i2cs/msoffice/lab\\_04b.zip](http://jacenty.kis.p.lodz.pl/i2cs/msoffice/lab_04b.zip).

### Ćwiczenie 1

Otwórz *lab\_04b.xls*. Pierwszy arkusz (**database**) zawiera dane z “dużą” ilością pracowników

Używając funkcji **Filtruj** dostępnej w zakładce **Dane** znajdź:

- wszystkich ludzi mieszkających w kanadzie zarabiających mniej niż 15 zł na godzinę
- wszystkich ludzi których nazwisko zaczyna się na “cal”,
- wszystkich ludzi pracujących więcej niż 35 w tygodniu zatrudnionych od 1997 do 1999.

	A	B	C	D	E	F	G
1		name	family nam	country	employed sin	hours per we	hourly ra
2	1	Colleen	Abel	Canada	2000	42,0	15,00 zł
3	2	Karina	Abel	Germany	1997	42,0	15,00 zł
4	3	James	Abel	Great Britain	2001	35,0	15,00 zł
5	4	William	Abel	Great Britain	1999	40,0	15,00 zł
6	5	Peter	Allen	Australia	1996	40,0	15,00 zł
7	6	Mary	Altman	Germany	1997	29,5	15,00 zł
8	7	Bob	Ambrose	Germany	1995	35,5	15,00 zł
9	8	Barry	Bally	Germany	1993	40,0	15,00 zł
10	9	Jacqueline	Banks	Australia	1994	40,0	15,00 zł
11	10	Mary	Barber	Germany	1997	35,5	15,00 zł
12	11	Tim	Barthoff	Germany	1999	40,0	15,00 zł
13	12	Teri	Binga	Australia	1998	40,0	15,00 zł
14	13	Frieda	Binga	Australia	1992	40,0	15,00 zł

### Ćwiczenie 2

Użyj funkcji **Sortowanie** dostępnej w zakładce **Dane**. Posortuj tabelę rosnąco po kraju, malejąco po nazwisku malejąco po roku zatrudnienia.

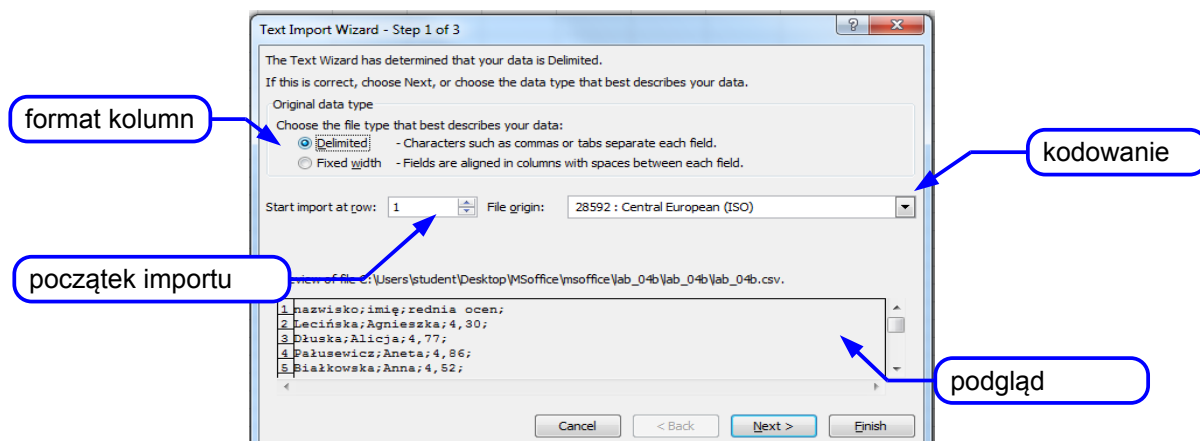
	A	B	C	D	E	F	G
1		name	family nam	country	employed sin	hours per we	hourly ra
2	93	Lindsey	Winger	Australia	1996	35,0	8,50 zł
3	91	Esther	Williams	Australia	1996	40,0	8,50 zł
4	87	Edward	Trelly	Australia	1996	40,0	8,50 zł
5	86	Amy	Tooley	Australia	2000	15,5	8,50 zł
6	85	Greg	Thomas	Australia	1997	40,0	8,50 zł
7	81	Lorrie	Sullivan	Australia	1991	15,5	8,50 zł
8	72	Grace	Sloan	Australia	1994	40,0	12,00 zł
9	71	Steve	Singer	Australia	1996	40,0	12,00 zł
10	69	Lynne	Simmons	Australia	1998	35,0	12,00 zł
11	59	Joanne	Parker	Australia	1994	40,0	18,00 zł
12	58	Alice	Owens	Australia	1997	15,5	18,00 zł
13	48	Jeri Lynn	MacFall	Australia	1994	40,0	18,00 zł
14	47	Joshua	Maccaluso	Australia	2001	40,0	18,00 zł

### Ćwiczenie 3

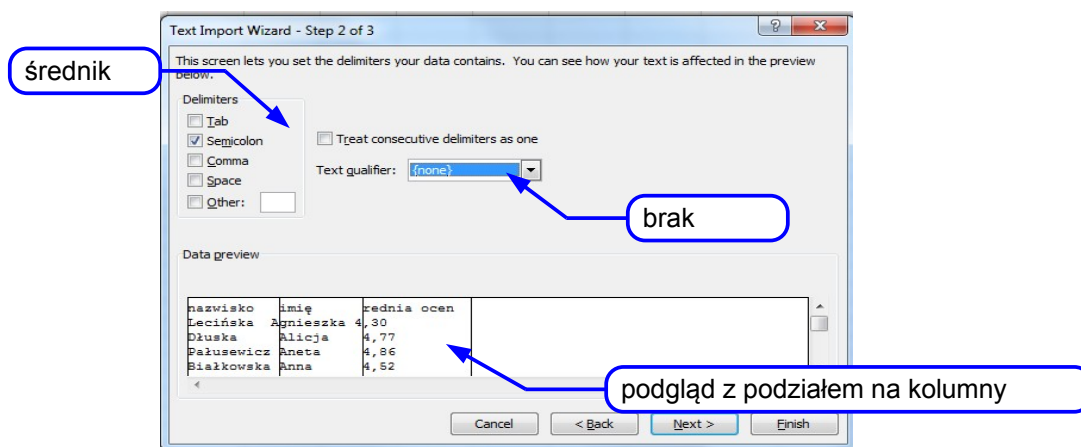
Zaimportuj dane z *lab\_04b.csv*. Ten plik to przykład pliku CSV (comma-separated values). CSV jest bardzo prosty i użyteczny do przenoszenia danych pomiędzy różnymi programami i systemami.

**Wskazówki:**

1. Kliknij na przycisk **Z tekstu** w sekcji **Dane zewnętrzne** zakładki **Dane**.
2. Ustaw kodowanie na Central European (ISO).



3. Wybierz znak do rozdziału kolumnami character (średnik)



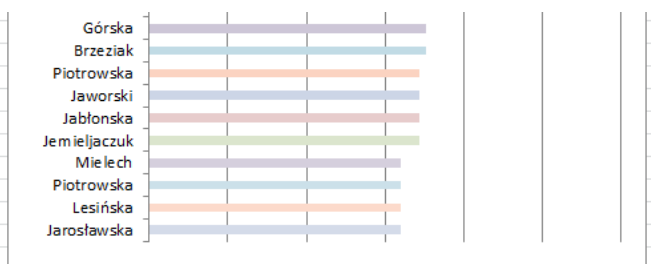
4. Arkusz powinien wyglądać następująco:

	A	B	C
1	nazwisko	imię	średnia ocen
2	Leścińska	Agnieszka	4,3
3	Dłuska	Alicja	4,77
4	Pałusewicz	Aneta	4,86
5	Białkowska	Anna	4,52
6	Czaplicka	Anna	4,47
7	Salik	Anna	4,36
8	Skorulska	Anna	4,07
9	Świętochowska	Barbara	4,7
10	Chmielewski	Cezary	4,49
11	Jakubowski	Cezary	4,05
12	Ławecki	Daniel	4,79
13	Ławecki	Dariusz	4,6
14	Adamowicz	Dominik	4,04
15	Cyrkun	Dominik	4,33
16	Minta	Dominika	4,39
17	Piotrowska	Dominika	3,43
18	Krasowska	Dorota	3,73
19	Świętochowska	Dorota	3,51
20	Mielech	Edgar	3,2
21	Roszek	Edyta	4,19
22	Filpkowska	Eliza	3,91
23	Grygoruk	Eliza	4,2
24	Jurowska	Eliza	4,3

### Ćwiczenie 4

Używając formuł policz: średnią, maksimum i minimum, mediane i odchylenie standardowe ocen. Stwórz wykres przedstawiający rozkład ocen. Na wykresie studenci powinni być posortowani według ocen. Zformatuj wykres tak jakj jest to przedstawione poniżej:

113	Górska	Małgorzata	3,51
114	Brzeziak	Monika	3,51
115	Piotrowska	Dominika	3,43
116	Jaworski	Hubert	3,43
117	Jabłonska	Magdalena	3,43
118	Jemieljaczuk	Marta	3,43
119	Mielech	Edgar	3,2
120	Piotrowska	Iwona	3,2
121	Lesińska	Małgorzata	3,2
122	Jarostawska	Monika	3,2
123			
124		avg	4,251157
125		min	3,2
126		max	4,86
127		median	4,3
128		std. Dev.	0,4030988
129			

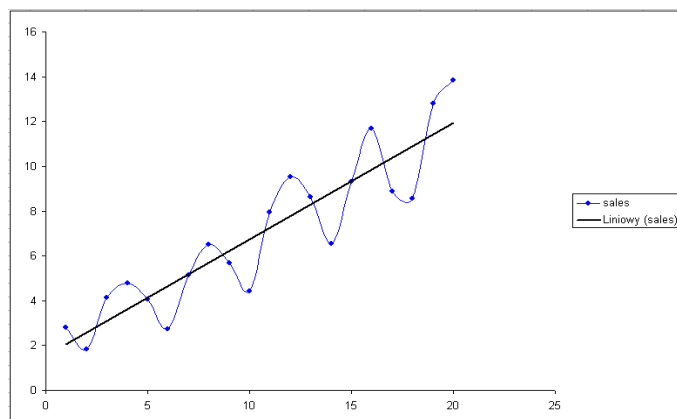


### Ćwiczenie 5

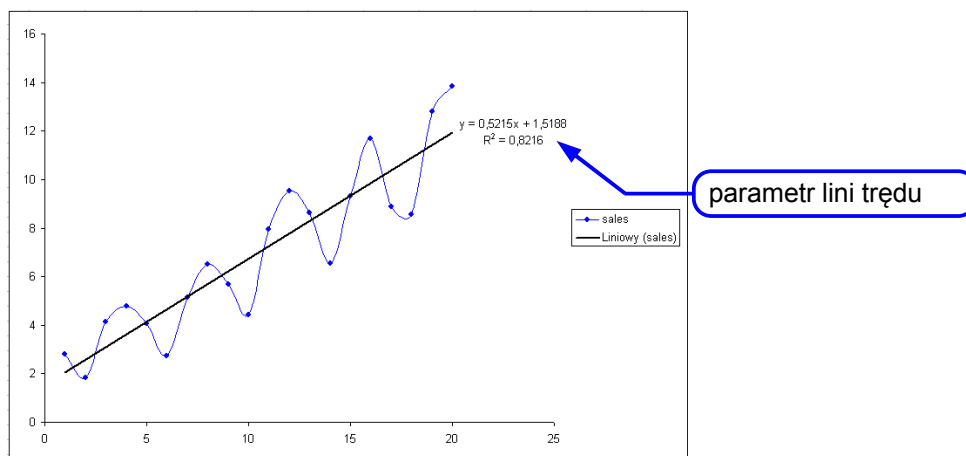
Kolejny arkusz (**sales**) w *lab\_04b.xls* zawiera kwartalne dane odnośnie sprzedaży jakiegoś produktu. Sprzedaż różni się pomiędzy kwartałami, co jest dość powęższne w przypadku produktów sezonowych. Kwartał do którego odnosi się dana sprzedarz oznaczony jest 1 , inne oznaczone są zerem. Kolumny **n** i **time** oznaczają to saom.

Na początku stwórz wykres liniowy sprzedarzy w zależności od czasu. Dodaj do niego linię trendu. Generalnie sprzedaż wzrasta ale są sezonowe wahania.

Zadanie polega na przewidzeniu sprzedarzy w przyszłości uwzględniając ogólny trend oraz wahania.

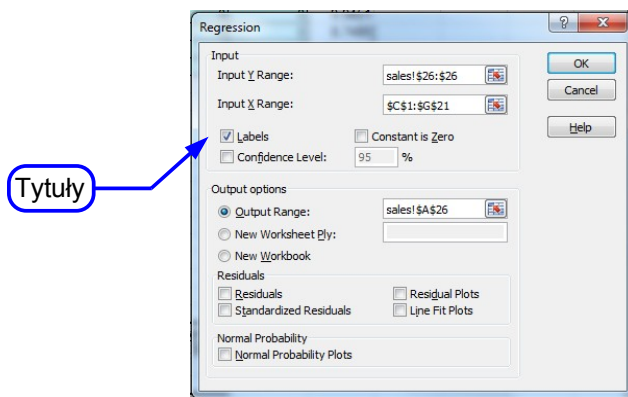


W opcjach lini trendu zaznacz aby  $R^2$  było widoczne. Linia trendu ma postać  $y = ax + b$ , parametr  $R^2$  pokazuje jak dobrze linia trendu dopasowana jest do danych. Maksymalna wartość  $R^2$  to 1, byłoby tak gdyby dane dokładnie odpowiadały linii trendu. W rzeczywistej sytuacji można uznać że linia trendu jest dość dobra kiedy  $R^2$  jest większa od 0.8.



Wybierz **Analizę statystyczną** w grupie **Analiza** zakładki **Dane** (jeżeli grupa Analiza nie jest dostępna trzeba zainstalować dodatek: **Analysis ToolPak**). Wybierz **Regresję** z listy i naciśnij OK.

Zaznacz zakresy X i Y (Y to sprzedaz, X to czas i kwartały) i wskaż komórkę w której ma pojawić się wynik (dowolna komórk po za tabelą). Zaznacz opcję **Tytuły**, ułatwi to orientację w wynikach regresji.



Zatwierdź przyciskiem OK. Tak powinny wyglądać wyniki regresji (bez zaznaczenia na żółto)

SUMMARY OUTPUT									
<b>Regression Statistics</b>									
Multiple R		0.984208							
R Square		0.968666							
Adjusted R Square		0.893644							
Standard Error		0.678157							
Observations		20							
<b>ANOVA</b>									
	df	SS	MS	F	Significance F				
Regression	5	213.2611	42.652229	115.9287252	7.0013E-11				
Residual	15	6.898457	0.4598971						
Total	20	220.1596							
		Coefficient	Standard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept		1.586838	0.387537	4.0946705	0.000956539	0.76082132	2.41285368	0.760821	2.412854
time		0.491663	0.026807	18.341157	1.10202E-11	0.43452577	0.54879923	0.434526	0.548799
1st quarter		0	0	65535	#NUM!	0	0	0	0
2nd quarter		-1.68826	0.429741	-3.928557	#NUM!	-2.604234	-0.772291	-2.60423	-0.77229
3rd quarter		0.877475	0.432242	2.0300546	0.060474898	-0.0438272	1.79877715	-0.04383	1.798777
4th quarter		1.793413	0.436378	4.1097642	0.000927715	0.86329386	2.72353114	0.863294	2.723531

przecięcie [i]

[tr]

[q1r]

[q2r]

Komórki potrzebne do przewidywania trendu sprzedaży zostały zaznaczone na żółto. Analiza regresji

wielokrotnej przyjmują postać:  $y = b_0 + a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n + [\text{losowy kompenet}]$ , gdzie  $y$  to jest sprzedaż  $x_1$  - czas,  $x_n$  - jedynki i zera (kwartały).  $b_0$  to przecięcie,  $a_1$  odpowiedni czas i tak dalej. Nie ma czynnika losowego (różnica pomiędzy rzeczywistymi danymi a przewidywanymi)

W kolumnie po prawo tabeli wpisz formułę regresji:

$$[i] + [tr] * [t] + [q1r] * [q1] + [q2r] * [q2] + [q3r] * [q3] + [q4r] * [q4]$$

gdzie:

- [i] **bezwzględne** odwołanie do komórki z wartością przecięcia w tablicy regresji
- [t] odwołanie do komórki z czasem
- [q1], [q2], [q3], [q4] odwołanie do komórek z kwartałami
- [tr] **bezwzględne** odwołanie do współczynnika czasu w tablicy regresji
- [q1r], [q2r], [q3r], [q4r] **bezwzględne** odwołania do współczynników poszczególnych kwartałów w tablicy regresji.

Rezultaty powinny wyglądać tak jak poniżej:

n	sales	time	1st quarter	2nd quarter	3rd quarter	4th quarter	estimated
1	2,798	1	1	0	0	0	2,0785
2	1,85	2	0	1	0	0	0,8819
3	4,15	3	0	0	1	0	3,9393
4	4,8	4	0	0	0	1	5,3469
5	4,06	5	1	0	0	0	4,04515
6	2,726	6	0	1	0	0	2,84855
7	5,149	7	0	0	1	0	5,90595
8	6,521	8	0	0	0	1	7,31355
9	5,679	9	1	0	0	0	6,0118
10	4,42	10	0	1	0	0	4,8152
11	7,962	11	0	0	1	0	7,8726
12	9,54	12	0	0	0	1	9,2802
13	8,638	13	1	0	0	0	7,97845
14	6,54	14	0	1	0	0	6,78185
15	9,302	15	0	0	1	0	9,83925
16	11,69	16	0	0	0	1	11,24685
17	8,884	17	1	0	0	0	9,9451
18	8,54	18	0	1	0	0	8,7485
19	12,8	19	0	0	1	0	11,8059
20	13,85	20	0	0	0	1	13,2135

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,984208
R Square	0,968666
Adjusted R Square	0,893644
Standard Error	0,678157
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	5	213,2611	42,652229	115,9287252	7,0013E-11
Residual	15	6,898457	0,4598971		
Total	20	220,1596			

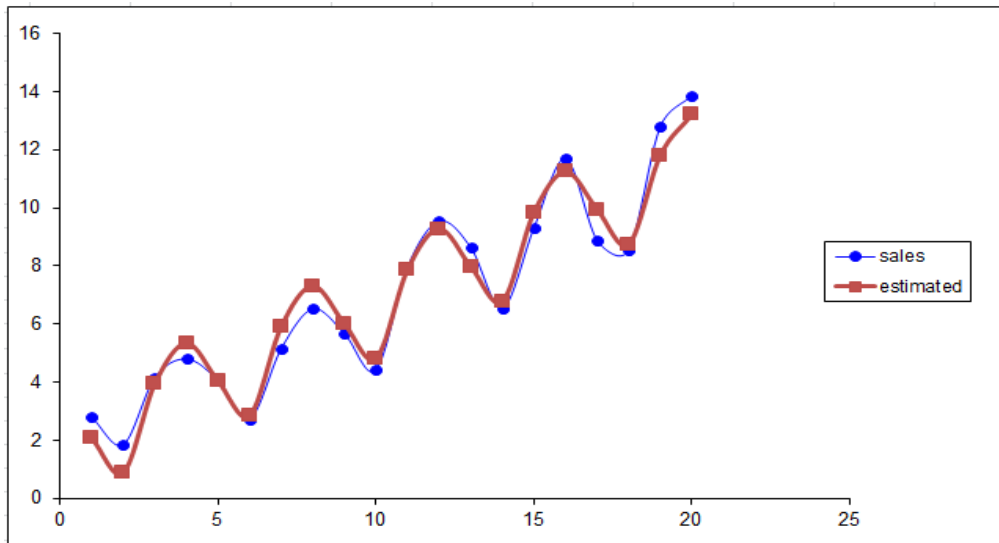
	Coefficients	Standard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 95,0%	Upper 95,0%
Intercept	1,586838	0,387537	4,0946705	0,000956539	0,76082132	2,41285368	0,760821	2,412854
time	0,491663	0,026807	18,341157	1,10202E-11	0,43452577	0,54879923	0,434526	0,548799
1st quarter	0	0	65535	#NUM!	0	0	0	0
2nd quarter	-1,68826	0,429741	-3,928557	#NUM!	-2,604234	-0,772291	-2,60423	-0,77229
3rd quarter	0,877475	0,432242	2,0300546	0,060474898	-0,0438272	1,79877715	-0,04383	1,798777
4th quarter	1,793413	0,436378	4,1097642	0,000927715	0,86329386	2,72353114	0,863294	2,723531

przewidywane wartość

formuła w komórce H2:

$$=B\$42+B\$43*C2+B\$44*D2+B\$45*E2+B\$46*F2+B\$47*G2$$

Tak jak widać współczynnik  $R^2$  wynosi około 0,969 czyli bardzo blisko 1. Stwórz wykres na którym będą prawdziwe dane i przewidywane:



Posiadając dobre równanie linii trendu można przewidzieć jak będzie zachowywać się sprzedaż w przyszłości zakładając, że trend będzie stały.

## Ćwiczenie 6

źródło: [http://www-zo.iinf.polsl.gliwice.pl/~kadam/pimfet\\_std/excel/excel.htm](http://www-zo.iinf.polsl.gliwice.pl/~kadam/pimfet_std/excel/excel.htm)

Ostatni skoroszyt *lab\_04b.xls* (**babies**) zawiera dane dotyczące noworodków. Ćwiczenie pokazuje w jaki sposób można z wykorzystaniem MS Excel stworzyć histogram.

Dla ułatwienia pracy nazwiemy zakres danych których będziemy używać w dalszej części ćwiczenia. Np. zamiast pisać H3:H181 będziemy mogli użyć nazwy **weight**.

1. Zaznacz komórki, które chcesz nazwać
2. Kliknij okienko nazwy z lewej strony paska formuły.



3. Wpisz nazwę (np. **weight**).

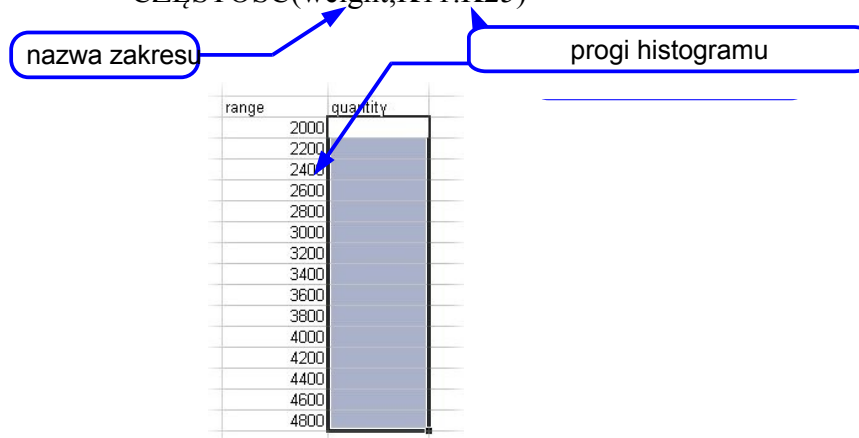
W ten sam sposób nazwij wzrost jako: **height**. Następnie policz wartości minimalne i maksymalne wzrostu. Znając dolne i górne wartości stwórz histogram. (np dla wagi zakres będzie wyglądał następująco: [1800; 2000[, [2000; 2200[, etc.). Aby to osiągnąć stwórz tabelę zawierającą dane potrzebne do histogramu. Użyj funkcji **CZĘSTOŚĆ** której argumentem będzie zakres komórek nazwany **weight** oraz zakresy wartości.

<b>max(weight)</b>	<b>min(weight)</b>	<b>max(height)</b>	<b>min(height)</b>
4640	1880	62	44

Formuła która zostanie wykorzystana do stworzenia tabeli potrzebnej do histogramu nazywa się “**formuła tablicowa**”. Jej wartości wykorzystują zakresy komórek.

Zaznacz cały zakres komórek w tabeli i wpisz funkcję CZĘSTOŚĆ:

=CZĘSTOŚĆ(weight;K11:K25)

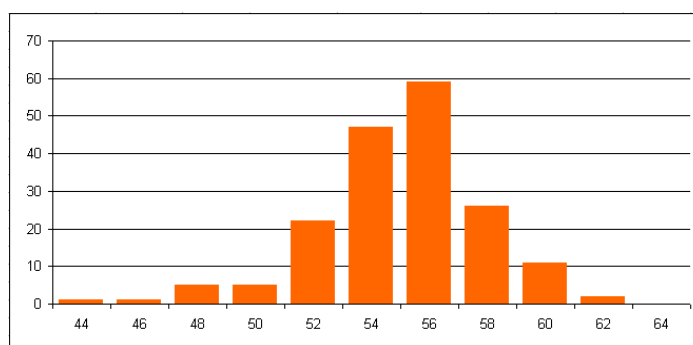
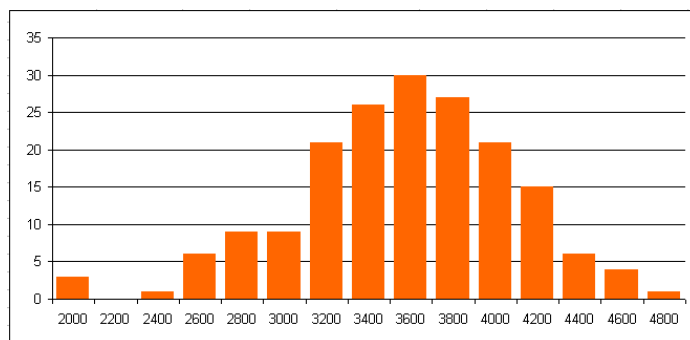


Zaakceptuj formułę poprzez jednoczesne wciśnięcie klawiszy **Ctrl+Shift+Enter**. Jest to jedyny sposób na wpisanie formuły tablicowej. Formuła pojawi się w nawiasach: {=CZĘSTOŚĆ(weight;K11:K25)}  
Dane pokazujące jak często występuje dany przedział pojawią się w odpowiednich komórkach:

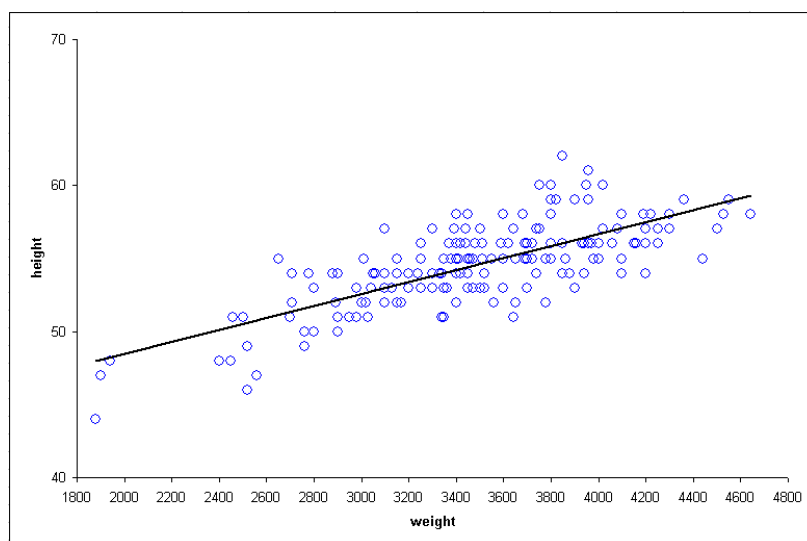
range	quantity
2000	3
2200	0
2400	1
2600	6
2800	9
3000	9
3200	21
3400	26
3600	30
3800	27
4000	21
4200	15
4400	6
4600	4
4800	1



Teraz można stworzyć histogramy wagi i wzrostu:



Na koniec stwórz wykres pokazujący zależność wzrostu od wagi. Dostosuj osie oraz zakresy i dodaj linię trendu:



Opracowanie: Jacek Wiślicki, [jacency@kis.p.lodz.pl](mailto:jacency@kis.p.lodz.pl) Laurent Babout, [lbabout@kis.p.lodz.pl](mailto:lbabout@kis.p.lodz.pl)

Przystosowanie do zajęć Technologie Informacyjne: Joanna Jójczyk [jojczyk@kis.p.lodz.pl](mailto:jojczyk@kis.p.lodz.pl)