POLITECHNIKA LUBELSKA

Wydział Matematyki i Informatyki Technicznej

Kierunek: Inżynieria i Analiza Danych

Projekt z zakresu analizy danych

Projekt indywidualny

Autor:
Szymon Chruśliński
nr albumu: 100961

Spis treści

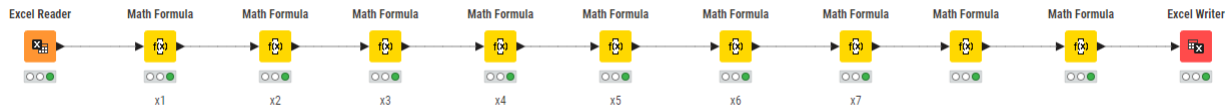
1 Usuwanie wartości odstających

Przed przystąpieniem do właściwej analizy znajdziemy i usuniemy wartości odstające, czyli obserwacje, które znacząco odbiegają od pozostałych danych w zbiorze. Warto to zrobić, ponieważ takie obserwacje mogą zniekształcać wyniki analiz statystycznych, co może prowadzić do błędnych wniosków. Dzięki ich usunięciu wyniki analizy są bardziej wiarygodne i dokładne.

1.1 Identyfikacja wartości odstających

Do znalezienia wartości odstających wykorzystamy program Knime.

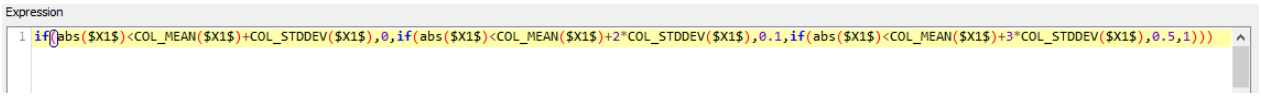
Tak przedstawia się układ nodów, dzięki którym zidentyfikujemy wartości odstające:



Rysunek 1: Układ nodów

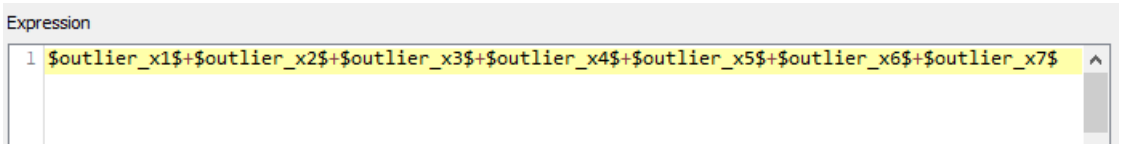
Opis nodów:

- **Excel Reader** - wczyta zbiór danych
- **Math Formula** - pozwoli na użycie formuł do szukania wartości odstających
- **Excel Writer** - zapisze zmodyfikowany zbiór danych



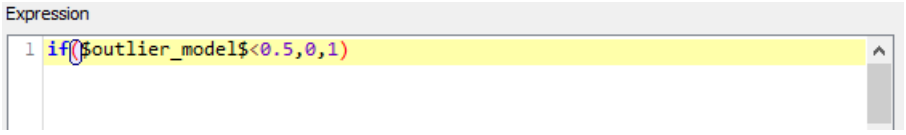
Rysunek 2: Formuła do sprawdzenia wartości odstającej

Za pomocą tej formuły znajdziemy wartości odstające dla poszczególnych zmiennych.
Pozostałe formuły (aż do x7) wyglądają analogicznie.



Rysunek 3: Sumowanie

Dzięki tej formule sumowana jest wartość odstawania dla każdej zmiennej. Suma posłuży nam do sprawdzenia warunku czy dana obserwacja jest wartością odstającą.



Rysunek 4: Warunek

Jeśli warunek jest spełniony, to obserwacja uznawana jest za wartość odstającą.

Do pliku stworzonego w KNIME dodajemy kolumnę `czy_outlier` z formułą: `=JEŻELI(P2=1;“outlier”;“”)`. Oprócz tego za pomocą formatowania warunkowego tło komórek spełniających warunek zostanie pokolorowane na czerwono, dzięki której łatwo odczytamy wynik.

Zmodyfikowany plik wygląda następująco:

1	X1	X2	X3	X4	X5	X6	X7	outlier_x1	outlier_x2	outlier_x3	outlier_x4	outlier_x5	outlier_x6	outlier_x7	outlier_model	outlier_wynik	czy_outlier
426	44,1144	28,965	62,4583	1,33386	17,59	51,2901	102,748	0	0	0	0	0	0	0	0	0	
427	59,1672	68,3344	141,52	-86,6688	18,0089	38,0325	76,061	0	0	0	0	0	0	0	0	0	
428	40,5648	46,2032	96,4607	-57,48	21,235	54,0859		0	0	0	0	0	0				
429	43,9187	66,2626	136,818	-110,95	21,1085	49,0232	98,9332	0	0	0	0,1	0	0	0	0,1	0	
430	54,7437	48,8457	102,02	-37,0496	16,6969	52,7137	105,837	0	0	0	0	0	0	0	0	0	
431	57,2298	24,226	53,186	41,7816	21,8195	43,7293	88,1013	0	0	0	0	0	0	0	0	0	
432	24,1892	31,8797	68,4825	-47,2606	17,7892	50,6319	100,735	0	0	0	0	0	0	0	0	0	
433	22,8402	56,9493	118,782	-125,168	25,8667	53,244	105,894	0	0	0	0,1	0	0	0	0,1	0	
434	60,3736	71,9574	148,15	-95,1249	18,8326	47,7659	95,0907	0	0	0	0,1	0	0	0	0,1	0	
435	53,8239	49,0774	102,78		19,1685	51,8944	104,356	0	0	0		0	0	0			
436	58,3066	49,2428	102,764	-31,1153	17,8002	51,9445	-896,883	0	0	0	0	0	0	1	1	1	outlier
437	62,6911	26,9526	58,5883	44,5245	21,2649	48,9686	97,1411	0	0	0	0	0	0	0	0	0	
438	69,4794	79,3902	163,743	-99,2118	20,6175	55,2755	110,848	0	0,1	0,1	0,1	0	0	0	0,3	0	
439	69,4386	63,3481	131,321	-51,1669	17,6315	44,957	89,0824	0	0	0	0	0	0	0	0	0	
440	49,8847	30,1475	64,689	9,32682	29,6696	61,8641	123,877	0	0	0	0	0	0	0	0	0	
441	22,9404	69,9291	144,426	-163,906	18,6248	46,9979	93,3332	0	0	0	0,1	0	0	0	0,1	0	
442	32,4672	69,5809	143,679	-143,808	22,2203	52,0825	103,687	0	0	0	0,1	0	0	0	0,1	0	

Rysunek 5: Sprawdzenie outlierów

Na podstawie przeprowadzonych działań możemy stwierdzić, że w naszym zbiorze danych jest 10 wartości odstających, które należy usunąć.

1.2 Usuwanie wierszy

Usunięcia wierszy dokonamy za pomocą funkcji Filtruj dostępnej w pakiecie Microsoft Excel.

- Zaznaczamy całą tabelę i wybieramy narzędzie Filtruj
- Klikamy strzałkę przy kolumnie `czy_outlier`
- W menu filtrów wybieramy Filtruj według tekstu, a następnie Równa się
- Wpisujemy kryterium (u nas “outlier”)

Po wykonaniu tych czynności otrzymujemy:

1	X1	X2	X3	X4	X5	X6	X7	outlier_x1	outlier_x2	outlier_x3	outlier_x4	outlier_x5	outlier_x6	outlier_x7	outlier_model	outlier_wynik	czy_outlier
436	58,3066	49,2428	102,764	-31,115	17,8002	51,9445	-896,88	0	0	0	0	0	0	1	1	1	outlier
533	34,5691	36,9958	78,2837	-41,849	28,595	59,946	1119,54	0	0	0	0	0	0	1	1	1	outlier
711	44,7883	77,2344	158,745	-142,13	20,4062	1052,86	2105,82	0	0,1	0,1	0,1	0	1	1	2,3	1	outlier
1101	49,6071	45,7507	95,7887	-38,038	1018,11	49,1874	97,6306	0	0	0	0	1	0	0	1	1	outlier
1201	51,9899	46,2414	591,918	-34,744	22,3448	53,1798	106,703	0	0	1	0	0	0	0	1	1	outlier
1498	60,6181	28,1921	556,009	36,6598	21,9447	46,6285	94,0365	0	0	1	0	0	0	0	1	1	outlier
1603	35,9871	103,484	211,887	-238,48	22,6897	62,1966	123,99	0	0,1	0,1	0,5	0	0	0	0,7	1	outlier
1756	225,674	40,3057	85,1664	330,431	22,4187	45,9705	91,5811	1	0	0	0,5	0	0	0	1,5	1	outlier
1868	58,6085	1033,29	2070,69	-2982,7	21,6438	51,4684	103,828	0	1	1	1	0	0	0	3	1	outlier
1984	2035,49	27,1611	59,2997	3989,51	13,837	56,7352	113,522	1	0	0	1	0	0	0	2	1	outlier

Rysunek 6: Usuwanie outlierów

Zaznaczamy otrzymane wiersze i klikamy Usuń wiersz.

2 Statystyki podstawowe

2.1 Wyznaczanie

- **Średnia** – średnia arytmetyczna wartości w kolumnie

  Formuła: `=ŚREDNIA(B2:B1991)`

- **Mediana** – wartość, która dzieli uporządkowany zbiór danych na dwie równe części

  Formuła: `=MEDIANA(B2:B1991)`

- **1. Kwartyl** – wartość poniżej której znajduje się 25% danych

  Formuła: `=KWARTYL(B2:B1991;1)`

- **3. Kwartyl** – wartość poniżej której znajduje się 75% danych

  Formuła: `=KWARTYL(B2:B1991;3)`

- **Minimum** – najmniejsza wartość w zbiorze danych

  Formuła: `=MIN(B2:B1991)`

- **Maksimum** – największa wartość w zbiorze danych

  Formuła: `=MAX(B2:B1991)`

- **Dominanta** – wartość, która pojawia się w zbiorze danych najczęściej

  Formuła: `=WYST.NAJCZĘŚCIEJ(B2:B1991)`

- **Odchylenie standardowe** – informuje o tym, jak szeroko wartości danej wielkości są rozrzucone wokół jej średniej

  Formuła: `=ODCH.STANDARDOWE(B2:B1991)`

	X1	X2	X3	X4	X5	X6	X7
średnia	45,09136	44,32493	93,04785	-42,6385	19,87312	49,94918	99,86701
mediana	45,39823	44,27923	93,08049	-40,8312	19,73949	49,85257	99,76721
Q1	32,60088	27,26592	59,05714	-93,5127	16,43825	46,40542	92,91457
Q3	57,70692	60,88114	126,1336	9,769446	23,27937	53,28179	106,6864
min	20,01024	10,05796	4,564668	-195,213	3,820429	33,9551	0,902907
max	69,95664	79,99511	164,7075	118,4073	37,08842	68,93629	138,6079
dominanta	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D
odchylenie_std	14,52552	19,85467	39,82431	65,94839	4,935836	5,075819	10,43717

Rysunek 7: Statystyki podstawowe

2.2 Interpretacja

- Dla każdej zmiennej (X1 do X7) wartości średniej i mediany są bardzo zbliżone. Świadczy to o tym, że dane są względnie symetryczne
- Podobna odległość między medianą a kwartylami 1. i 3. dla większości zmiennych również wskazuje na symetrię rozkładu danych
- Największe wartości maksymalne występują w zmiennych X3 (164,7) i X7 (138,6), natomiast jeśli chodzi o wartości minimalne to najbardziej wyróżnia się zmienna X4, dla której minimum wynosi -195,2
- Brak dominanty (`#N/D`) dla wszystkich zmiennych oznacza, że każda wartość w zbiorze danych występuje tylko raz
- Zmienna X4 ma najwyższe odchylenie standardowe (65,94839), co wskazuje na największe zróżnicowanie danych w tej zmiennej, natomiast najmniejsze mają zmienne X5 i X6

2.3 Porównanie statystyk podstawowych dla danych z wartościami odstającymi i bez

	X1	X2	X3	X4	X5	X6	X7
średnia	46,1943	44,848	94,5894	-42,0012	20,3792	50,4648	100,899
mediana	45,4059	44,2825	93,1702	-40,6739	19,7526	49,8568	99,7712
Q1	32,6447	27,3034	59,1399	-93,5148	16,4479	46,4187	92,9204
Q3	57,7917	60,8924	126,582	10,0666	23,2764	53,2959	106,694
min	20,0102	10,058	4,56467	-2982,66	3,82043	33,9551	-896,883
max	2035,49	1033,29	2070,69	3989,51	1018,11	1052,86	2105,82
dominanta	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D
odchylenie_std	46,9926	29,7421	61,4497	129,952	22,8702	22,9991	56,0571

Rysunek 8: Statystyki podstawowe dla danych z wartościami odstającymi

- Po usunięciu wartości odstających wartości minimalne i maksymalne znacząco się zmieniły:

  Dla X4: wartość minimalna zmieniła się z -2982,66 na -195,213, a maksymalna z 3989,51 na 118,4073

  Dla X7: wartość minimalna zmieniła się z -896,883 na 0,902907, a maksymalna z 2105,82 na 138,6079

  Wniosek: Usunięcie wartości odstających skutecznie ograniczyło skrajne wartości, co daje bardziej wiarygodny zakres danych

- Odchylenie standardowe w większości przypadków znacznie spadło:

  Dla X1: z 46,9926 do 14,52552

  Dla X4: z 129,952 do 65,94839

  Dla X7: z 56,0571 do 10,43717

  Wniosek: Zmniejszenie odchylenia standardowego wskazuje na zmniejszenie zróżnicowania danych

3 Szereg rozdzielczy dla zmiennej X1

3.1 Tworzenie szeregu rozdzielczego

Aby zbudować szereg rozdzielczy potrzebne nam będą:

- **Minimum** - Najniższa wartość w zbiorze danych
- **Maksimum** - Najwyższa wartość w zbiorze danych
- **Rozpiętość** - Różnica pomiędzy wartością maksymalną a minimalną

  Formuła: `=Maksimum-Minimum`
- **Ilość przedziałów** - u nas 10
- **Krok** - Szerokość każdego przedziału, określająca, jak duży zakres wartości przypada na pojedynczy przedział

  Formuła: `=Rozpiętość/Ilość przedziałów`

W naszym zbiorze wartości te są następujące:

xmin	20,01024
xmax	69,95664
rozpietosc	49,9464
ilosc przedzialow	10
krok	4,99464

Rysunek 9: Podstawowe własności szeregu

Tworzenie szeregu rozdzielczego rozpoczynamy od ustalenia wartości pierwszego lewego końca przedziału, będzie to wartość minimum. Prawy koniec przedziału obliczamy według wzoru: `=lewy koniec przedziału + krok`. Lewy koniec dla następnego przedziału jest równy prawemu końcowi dla poprzedniego. Czynności te powtarzamy dla wszystkich przedziałów.

Gdy mamy już wyznaczone końce przedziałów możemy przejść do wyznaczania innych wartości szeregu:

- **Liczba obserwacji w pojedynczym przedziale**

  Formuła: `=CZĘSTOŚĆ(kolumna X1;prawe końce)`
- **Środek przedziału**

  Formuła: `=(Lewy koniec+Prawy koniec)/2`

Oprócz tego do wyliczenia wartości statystyk będą nam potrzebne: iloczyn środka przedziału i liczby obserwacji `(=(x_i*n_i))` oraz iloczyn kwadratu środka przedziału i liczby obserwacji `(=(x_i)^2*n_i)`.

Zbudowany szereg rozdzielczy:

	lewykoniec	prawykoniec	ile w przedziale (n_i)	srodek przedzialu (x_i)	x_i*n_i	(x_i)^2*n_i
1	20,010235	25,00487544	209	22,50755542	4704,079	105877,32
2	25,004875	29,99951547	182	27,50219546	5005,4	137659,48
3	29,999515	34,99415551	211	32,49683549	6856,832	222825,35
4	34,994156	39,98879554	201	37,49147552	7535,787	282527,76
5	39,988796	44,98343557	179	42,48611556	7605,015	323107,53
6	44,983436	49,97807561	212	47,48075559	10065,92	477937,5
7	49,978076	54,97271564	167	52,47539563	8763,391	459862,41
8	54,972716	59,96735568	227	57,47003566	13045,7	749736,73
9	59,967356	64,96199571	196	62,46467569	12243,08	764759,8
10	64,961996	69,95663574	205	67,45931573	13829,16	932905,65
			0			
			suma (n_i)		suma	suma
			1989		89654,36	4457199,5

Rysunek 10: Szereg rozdzielczy dla danych bez wartości odstających

Rozkład obserwacji w przedziałach jest równomierny. Każdy przedział zawiera mniej więcej 200 obserwacji, co sugeruje, że dane są rozłożone w sposób zrównoważony. Takie rozmieszczenie umożliwia bardziej wiarygodną analizę rozkładu oraz innych miar statystycznych.

3.2 Wyznaczenie wartości statystyk

- Aby wyliczyć **średnią** skorzystamy z następującej formuły: `=Suma_n_i/Suma_(n_i * x_i)`
- Do policzenia **wariancji** użyjemy formuły: `=(Suma_((x_i)^2 * n_i)/Suma_n_i)-średnia^2`
- Natomiast **odchylenie standardowe** wyliczymy za pomocą: `=PIERWIASTEK(Wariancja)`

U nas wartości statystyk przedstawiają się tak:

średnia	45,07509
wariancja	209,1609
odchylenie_std	14,4624

Rysunek 11: Statystyki podstawowe dla szeregu

3.3 Szereg rozdzielczy dla danych z wartościami odstającymi

	lewykoniec	prawykoniec	ile_w_przedziale (n_i)	srodek_przedzialu (x_i)	x_i*n_i	(x_i)^2*n_i
1	20,010235	221,5586578	1997	120,7844466	241206,5	29133998,4
2	221,55866	423,1070802	1	322,332869	322,3329	103898,478
3	423,10708	624,6555025	0	523,8812913	0	0
4	624,6555	826,2039249	0	725,4297137	0	0
5	826,20392	1027,752347	0	926,9781361	0	0
6	1027,7523	1229,30077	0	1128,526558	0	0
7	1229,3008	1430,849192	0	1330,074981	0	0
8	1430,8492	1632,397614	0	1531,623403	0	0
9	1632,3976	1833,946037	0	1733,171826	0	0
10	1833,946	2035,494459	1	1934,720248	1934,72	3743142,44
			0			
			suma (n_i)		suma	suma
			1999		243463,6	32981039,3

Rysunek 12: Szereg rozdzielczy dla danych z wartościami odstającymi

Rozmieszczenie obserwacji jest nierównomierne – jeden przedział zawiera aż 1997 obserwacji, podczas gdy pozostałe przedziały prawie żadnych. Taki rozkład wskazuje na bardzo silne skupienie danych w jednym przedziale, może to prowadzić do problemów w analizie, ponieważ nie oddaje pełnej różnorodności danych.

4 Uzupełnianie danych

4.1 Wyznaczenie modeli za pomocą regresji liniowej

Aby wyznaczyć za pomocą regresji liniowej, niezbędne będzie obliczenie macierzy korelacji, która umożliwi identyfikację zmiennych o istotnych zależnościach. Na podstawie tej macierzy będziemy w stanie określić, które zmienne są ze sobą silnie powiązane, co pomoże w doborze odpowiednich zestawów zmiennych do budowy modeli.

Do stworzenia macierzy korelacji wykorzystamy funkcję Korelacja dostępną w zakładce Analiza Danych w programie Microsoft Excel. Po wybraniu odpowiedniej funkcji, określamy zakres danych, który obejmuje wszystkie zmienne, dla których ma zostać obliczona korelacja.

Otrzymujemy:

	X1	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*
X1	1	0,01979402	0,021411211	0,422318484	0,04042	-0,022537	-0,018186642
X2	0,019794	1	0,999973835	-0,896806702	0,04472	-0,00583	-0,002661366
X3	0,0214112	0,99997384	1	-0,897140551	0,04356	-0,006375	-0,003126308
X4	0,4223185	-0,8968067	-0,897140551	1	-0,0207	-0,004075	-0,005178747
X5	0,040424	0,0447177	0,043563779	-0,02069202	1	-0,011766	-0,011361085
X6	-0,022537	-0,0058301	-0,006375081	-0,004074861	-0,0118	1	0,99843844
X7	-0,018187	-0,0026614	-0,003126308	-0,005178747	-0,0114	0,9984384	1

Rysunek 13: Macierz korelacji

Z macierzy odczytujemy, że istnieje silna korelacja między X2 i X3 (korelacja 0,99997). Oprócz tego X6 i X7 są bardzo silnie skorelowane (0,9984). Wykorzystamy te zmienne do budowy modeli.

Nasze modele przyjmą postać `a*X + b`. Do wyznaczenia współczynników a i b zastosujemy następujące metody:

- **Współczynnik a** - Zostanie obliczony przy pomocy funkcji, której argumentami są kolumna zawierająca zmienną zależną (wartość do przewidzenia) oraz kolumna zawierająca zmienną niezależną (wykorzystywaną jako część modelu)

  Formuła: `=NACHYLENIE(C1:C1991;D1:D1991)`

- **Współczynnik b** - Zostanie obliczony za pomocą funkcji, dla której zakresy danych są takie same jak w przypadku współczynnika a

  Formuła: `=ODCIĘTA(C1:C1991;D1:D1991)`

Wyznaczone współczynniki wyglądają następująco:

pomiedzy x2 i x3 istnieje bardzo silna korelacja				
X2=a_2*X3+b_2	a_2=	0,49999	b_2=	-2,24958
X3=a_3*X2+b_3	a_3=	1,99994	b_3=	4,5039
pomiedzy X6 i X7 istnieje bardzo silna korelacja				
X6=a_6*X7+b_6	a_6=	0,4971	b_6=	0,27759
X7=a_7*X6_b_7	a_7=	2,00538	b_7=	-0,24486

Rysunek 14: Wybór zmiennych i wyznaczenie modeli

Tak przedstawiają się zbudowane przez nas modele:

X2_model	X3_model	X6_model	X7_model	X2_błąd	X3_błąd	X6_błąd	X7_błąd
30,7704	65,95307	48,83811	98,28009	-0,04497	0,088514	0,292117	-0,5927638
10,9036	26,42554	46,21	92,04638	0,057574	-0,11864	-0,188266	0,35408953
53,0557	110,9899	45,71059	92,24979	0,188846	-0,37677	0,412577	-0,8539674
36,1109	76,43452	53,09077	106,2965	-0,14457	0,288261	0,036904	-0,0542677
25,7484	55,61546	45,80439	90,9441	-0,19186	0,38176	-0,332319	0,64042395
16,0069	36,8743	61,28174	122,6305	0,178786	-0,36053	-0,009016	0,0892419
78,6366	161,5141	50,23808	100,9477	-0,12917	0,26192	0,222357	-0,4440799
19,9696	44,90995	54,51842	109,2842	0,234027	-0,47059	0,099107	-0,170046
77,7381	160,3482	46,71299	93,51478	0,186327	-0,36915	0,040974	-0,1024664
24,6292	53,27191	46,02426	92,53167	-0,24446	0,48685	0,239469	-0,5048502
48,567	101,1469	50,42246	100,9432	-0,24415	0,488729	0,035748	-0,0687004
50,2776	105,1192	50,49832	101,4027	0,031493	-0,06236	0,189015	-0,3755829
55,2922	114,6489	56,4039	112,876	-0,21813	0,437386	0,004699	0,03111596
37,5473	79,49382	55,72979	110,648	-0,05124	0,101764	-0,432214	0,9030625
43,8074	92,58345	48,56283	96,65321	0,233659	-0,46736	-0,24386	0,48034764
34,1167	73,21283	46,07324	91,49208	0,238726	-0,47851	-0,327912	0,63327426
55,5359	116,012	47,15568	94,74265	0,219732	-0,43828	0,210575	-0,4398035

Rysunek 15: Modele wyznaczone regresją liniową

Jak wynika z przedstawionych danych, wartości błędów są niewielkie, co świadczy o tym, że nasze modele dobrze estymują zmienne.

	X2_model	X3_model	X6_model	X7_model
błąd	0,031602	-0,06321	0,0772876	-0,15561671

Rysunek 16: Średnie błędy modeli

Średnie błędy modelu potwierdzają wysoką dokładność estymacji, co zgadza się z wcześniejszymi wnioskami o dobrej jakości modeli.

4.2 Mediana

W przypadku pozostałych zmiennych (X1, X4, X5) obserwujemy słabsze korelacje, dlatego braki w danych zostaną uzupełnione za pomocą wartości mediany. Wybór mediany jako metody imputacji wynika z jej większej odporności na wartości odstające w porównaniu do średniej. Metodę tę można również zastosować w przypadku dużych braków danych, gdzie wypełnianie średnią może być mniej reprezentatywne.

4.3 Uzupełnianie oraz przedstawienie danych

Aby uzupełnić brakujące dane, wykonujemy następujące kroki:

- Zaznaczamy odpowiednie kolumny.
- Używamy skrótu klawiszowego CTRL+G, aby otworzyć okno „Idź do".
- Wybieramy opcję „Specjalnie", a następnie zaznaczamy „Puste".
- Program automatycznie zaznaczy wszystkie puste komórki w wybranych kolumnach.
- Uzupełniamy zaznaczone komórki odpowiednimi formułami.

Dane po uzupełnieniu:

	X1_uzupełnione	X2_uzupełnione	X3_uzupełnione	X4_uzupełnione	X5_uzupełnione	X6_uzupełnione	X7_uzupełnione
	31,16561695	23,89887901	52,36552481	-9,365403111	26,8057329	45,19879841	90,33612896
	48,31296091	39,92567654	84,33797807	-23,15110782	12,53037309	41,36179206	82,68834385
	65,77523323	70,95427284	146,3077844	-81,31235205	17,27772974	58,36442016	115,9881256
	62,86274826	23,50197449	51,84665695	55,21957304	19,33510354	51,26752976	101,838847
	65,7848253	28,67987252	61,49648619	45,53003304	12,20595609	46,11870094	93,07339366
	52,92671348	53,62848223	112,0364745	-55,03201974	25,43720676	56,25647529	111,9263594
	66,70340326	76,06186856	156,3210884	-94,77879917	24,32425616	57,459776	114,1092111
	54,99759663	51,48774749	107,148937	-44,46804921	22,16770921	51,27473102	101,8472312
	52,5548352	44,82642654	94,30535145	-29,36960923	16,0761676	56,60286526	112,288285
	54,8865073	18,23478484	41,28742974	55,06866007	27,89984348	39,79322885	79,25242549
ile pustych	0	0	0	0	0	0	0

Rysunek 17: Uzupełnione dane

Dane zostały uzupełnione i są teraz kompletne, potwierdza to wynik funkcji =LICZ.PUSTE, która dla wszystkich komórek zwraca wartość 0. Oznacza to, że nie ma pustych komórek w danych, a wszystkie brakujące wartości zostały skutecznie uzupełnione.

Aby przeprowadzić pełną analizę uzupełnionych danych, obliczymy podstawowe statystyki, podobnie jak zrobiliśmy to poprzednio 2.1.

	X1_uzupełnione	X2_uzupełnione	X3_uzupełnione	X4_uzupełnione	X5_uzupełnione	X6_uzupełnione	X7_uzupełnione
średnia	45,09150992	44,25824989	93,01783506	-42,63755002	19,87298624	49,9244457	99,87265402
mediana	45,39823278	44,25823774	93,08049254	-40,83122027	19,7394857	49,84644824	99,76921906
Q1	32,60174299	27,17518471	59,04354154	-93,46428903	16,44070698	46,40149545	92,92263646
Q3	57,70201638	60,84723543	126,0837878	9,758288925	23,27640428	53,2785118	106,6867156
min	20,01023541	0,03270134	4,564668446	-195,2125718	3,820428911	0,726424754	0,902906521
max	69,95663574	79,99510734	164,7075314	118,4073135	37,08842221	68,93629281	138,6079201
dominanta	45,39823278	#N/D	#N/D	-40,83122027	19,7394857	#N/D	#N/D
odchylenie_std	14,52186649	19,91380103	39,82749761	65,93182192	4,93335533	5,193121346	10,43122

Rysunek 18: Statystyki podstawowe dla danych po uzupełnieniu

Po uzupełnieniu braków dane zachowały swoją charakterystykę statystyczną, a jedyną wyraźną zmianą jest pojawienie się dominanty w niektórych kolumnach. Jest to logiczna konsekwencja zastosowania mediany jako wartości uzupełniającej brakujące dane.

Tak przygotowane, uzupełnione dane są kompletne i zachowują kluczowe właściwości statystyczne, co umożliwia ich dalszą analizę.