RAPORT PROJEKTU Z EKSPLORACJI DANYCH

Autor: Edyta Margol, IAD I rok II stopień

Dane zawierają 7 zmiennych, z czego każda ma 2000 obserwacji (plik Projekt2 arkusz modele). Poniżej przedstawiono fragment zbioru.

X1	X2	X3	X4	X5	X6	X7
27,18496	46,15066	96,79169	-84,0821	17,64097	52,72836	106,1778
46,14038	60,68874	126,1776	-89,7854	27,75922	59,00527	118,2243
20,25693	67,91115	140,4745	-163,22	19,78014	45,61274	90,75389
57,33207	74,80017	154,5909	-109,736	15,47976	50,75431	100,7761
22,83415	24,54442	53,89242	-27,965	15,29557	57,19266	114,5106
46,21825	13,84118	31,96779	50,91296	19,56755	39,30269	78,29279
51,46097	45,13017	94,32177	-32,4686	25,45952	57,76816	114,6504
50,29862	75,76017	155,5802	-126,683	28,08678	50,45225	100,3351
51,22092	27,24745	59,37145	20,69948	16,35249	49,1375	97,36636

Każda z tych zmiennych zawiera braki danych. Poszczególną ich liczbę dla każdej zmiennej przedstawiono poniżej.

zmienna	X1	X2	X3	X4	X5	X6	X7
braki danych	1	3	2	1	2	1	3

Poniżej natomiast wyznaczono wartości podstawowych statystyk:

zmienna	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**
srednia	45,775	45,82771	96,48565	-45,7902	20,49328	50,54277	101,0415
odchylenie	47,19001	30,38817	63,12087	129,9103	23,17059	22,90702	56,00709
mediana	44,62674	45,19556	94,7569	-46,5614	19,90018	50,14758	100,2873
Q1	32,45499	28,54264	61,56408	-98,7921	16,64035	46,66999	93,34108
Q3	57,26558	62,62509	129,4954	5,982533	23,3805	53,35571	106,752
Q0	20,03558	10,02088	4,061507	-3124,51	3,797429	34,68974	-911,334
Q4	2044,89	1064,097	2132,922	3855,322	1031,371	1050,05	2099,544
dominanta	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D

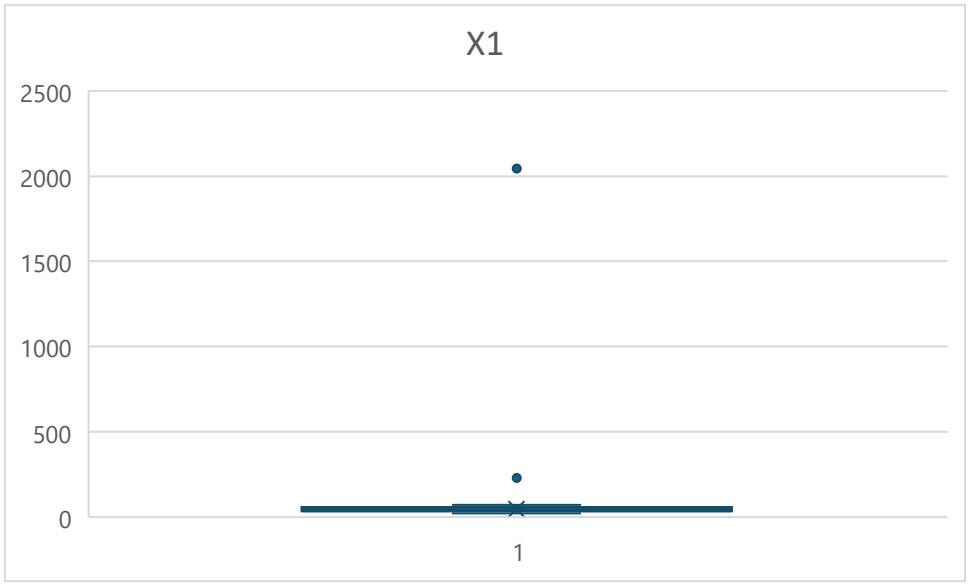
Dominanta oznaczona jako #N/D wskazuje, że nie występuje w analizowanym przedziale, co może być spowodowane brakami danych lub unikalnością wartości. Natomiast dla poszczególnych zmiennych są następujące wnioski:

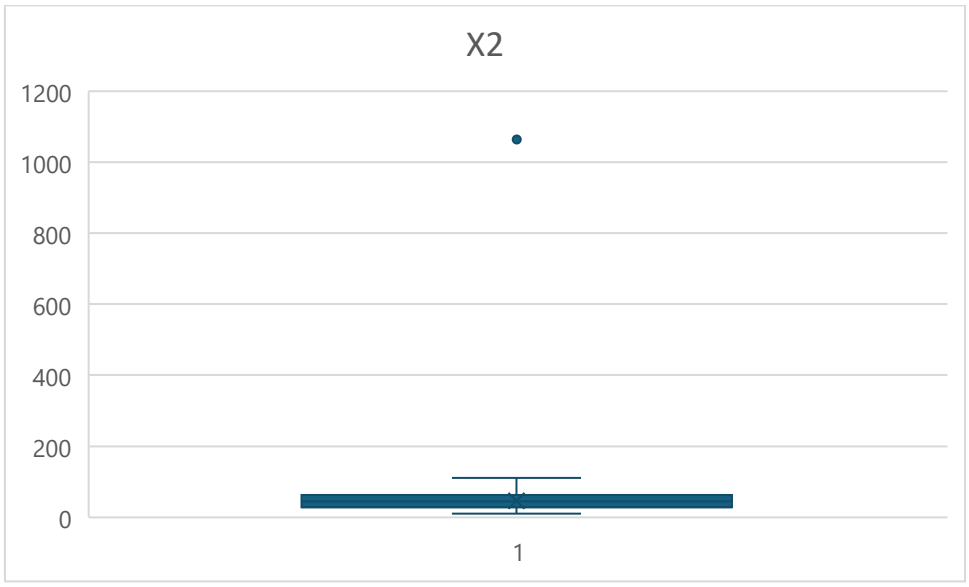
- X1: Średnia i mediana są zbliżone, co wskazuje na symetryczny rozkład większości danych. Wysokie wartości maksymalne znacząco podnoszą odchylenie standardowe, co może sugerować występowanie wartości odstających.
- X2: Rozkład jest symetryczny (średnia ≈ mediana). Wysoka wartość maksymalna (1064,10) sugeruje możliwą obecność skrajnych wartości, które mogą mieć wpływ na średnią i zwiększać odchylenie standardowe.

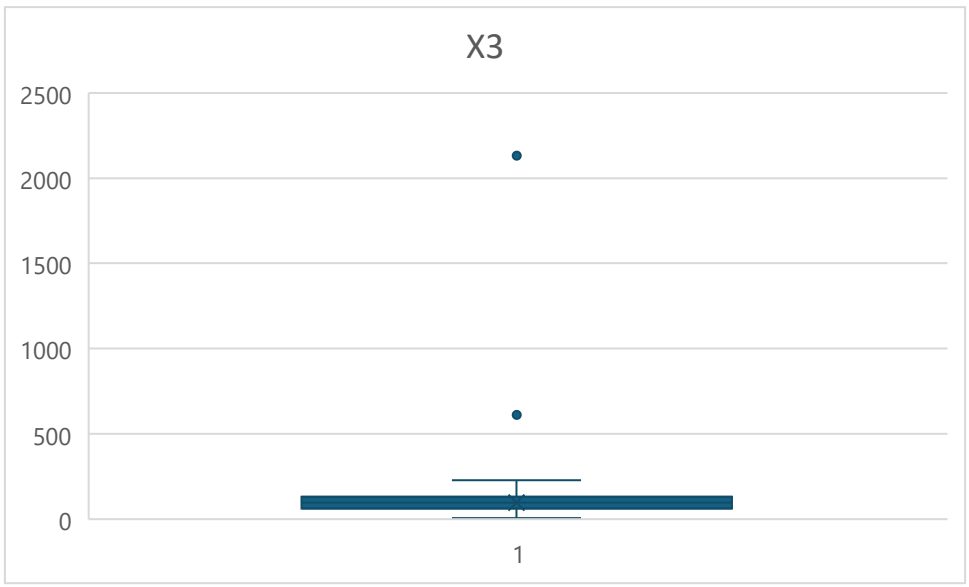
- X3: Średnia i mediana wskazują na względnie symetryczny rozkład, ale wysokie odchylenie standardowe i wartość maksymalna mogą sugerować duże rozproszenie danych oraz obecność odstających obserwacji.
- X4: Dane w X4 są najbardziej rozproszone (najwyższe odchylenie standardowe). Bardzo duże wartości minimalne i maksymalne wskazują na silną obecność wartości odstających, co może zniekształcać analizy tej zmiennej.
- X5: Średnia i mediana są do siebie zbliżone, a odchylenie standardowe jest relatywnie niskie, co wskazuje na stabilność danych w większości przypadków. Jednak wartość maksymalna jest bardzo wysoka, co może być odstającym punktem w danych.
- X6: Dane w X6 są symetryczne (średnia ≈ mediana), a rozkład jest stosunkowo skupiony (niskie odchylenie standardowe). Wysoka wartość maksymalna sugeruje obecność potencjalnych odstających obserwacji.
- X7: Średnia i mediana są zbliżone, co sugeruje symetrię większości danych. Jednak ekstremalnie niskie i wysokie wartości wskazują na obecność silnych odstających obserwacji, które wpływają na odchylenie standardowe i mogą zakłócać wyniki analiz.

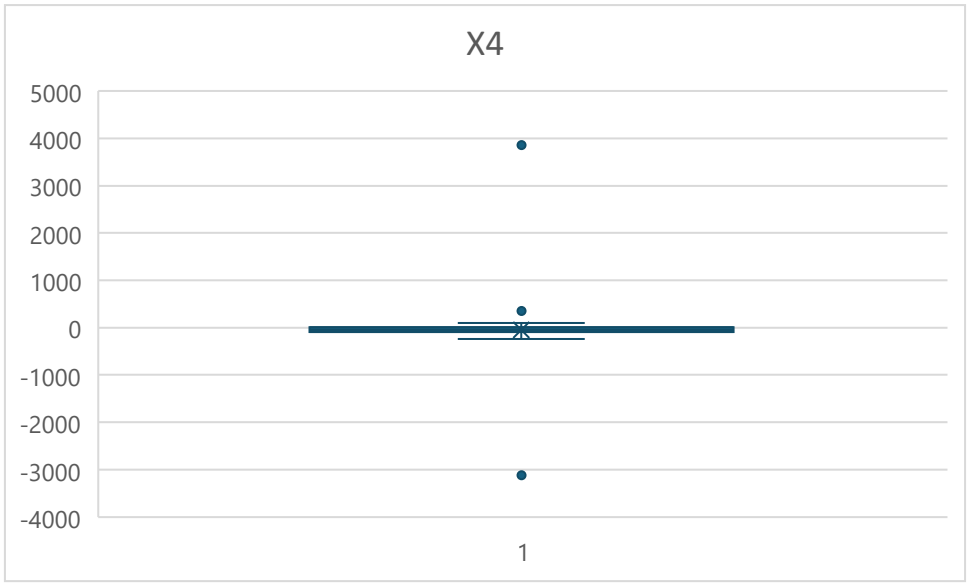
Podsumowując: zmienna X4 wyróżnia się największą zmiennością oraz ekstremalnymi wartościami, co może wymagać dodatkowego przetwarzania (np. usunięcia wartości odstających). Z kolei zmienne X5 i X6 są bardziej stabilne, z niskim rozproszeniem wokół średnich. Wysokie wartości maksymalne w X2, X3, X4, i X7 wskazują na możliwość występowania obserwacji odstających, które mogą znacząco wpłynąć na analizy i modele.

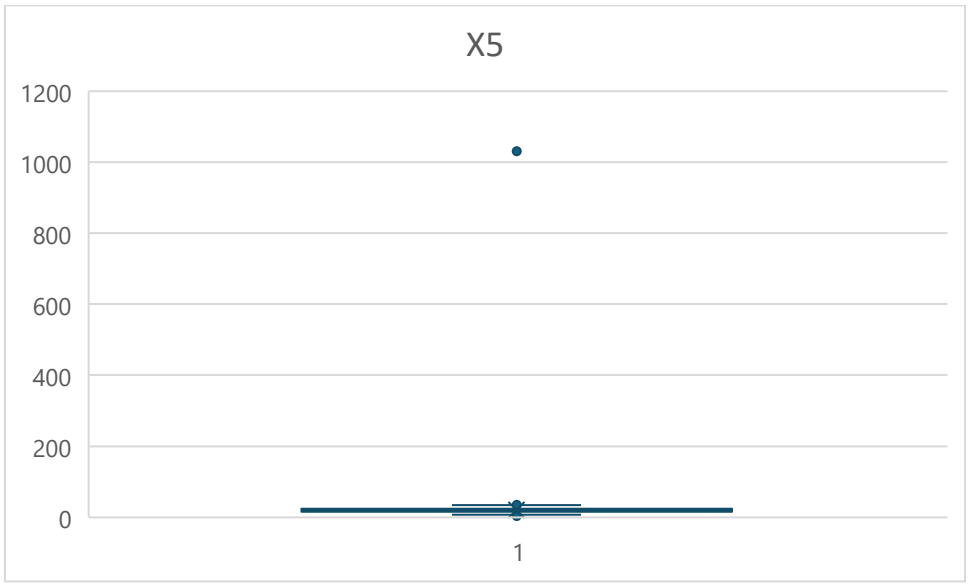
Aby zweryfikować obecność wartości odstających w danych, przygotujemy wykresy pudełkowe dla każdej zmiennej (plik „Projekt2”, arkusz „dane początkowe”). Wartości znajdujące się poza górnym i dolnym wąsem na wykresach będą uznane za wartości odstające.

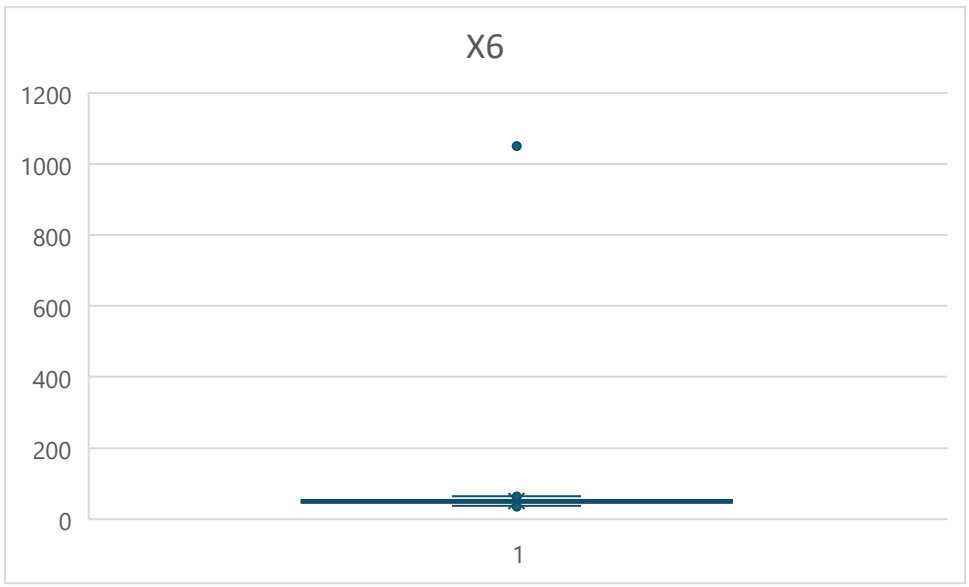

X1
2500
2000
1500
1000
500
0
1

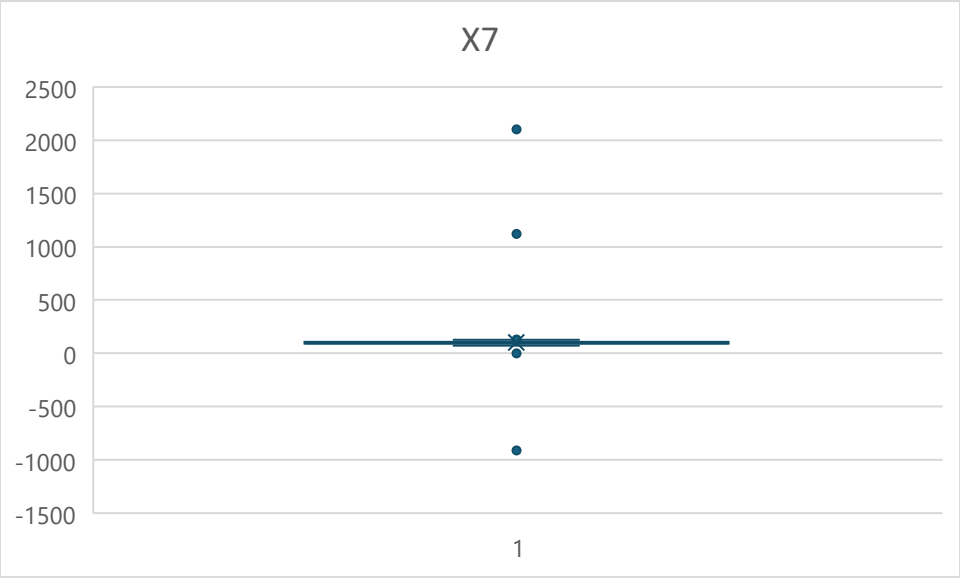

X2
1200
1000
800
600
400
200
0
1


X3
2500
2000
1500
1000
500
0
1


X4
5000
4000
3000
2000
1000
0
-1000
-2000
-3000
-4000
1


X5
1200
1000
800
600
400
200
0
1


X6
1200
1000
800
600
400
200
0
1



Można zauważyć, że każda ze zmiennych zawiera obserwacje odstające.

Dalsze etapy projektu będą realizowane dwiema metodami.

PODEJŚCIE 1 – bez usuwania braków danych oraz wartości odstających

Następnym etapem projektu jest skonstruowanie szeregu rozdzielczego przedziałowego dla zmiennej X1 (plik Projekt2 arkusz szereg).

Na początku wyznaczono podstawowe wartości.

liczba obserwacji n=	2000
x_min=	20,03558
x_max=	2044,89
rozpietosc R=	2024,854
dlugosc jednego przedzialu h=	1,012427

Następnie skokonstruowano szereg.

szereg						
nr przedziału	lewy koniec	prawy koniec	x_i	n_i	x_i*n_i	x_i^2*n_i
1	20,03558053	222,5209815	121,2783	1997	242192,7	29372717,6
2	222,5209815	425,0063825	323,7637	1	323,7637	104822,922
3	425,0063825	627,4917835	526,2491	0	0	0
4	627,4917835	829,9771845	728,7345	0	0	0
5	829,9771845	1032,462585	931,2199	0	0	0
6	1032,462585	1234,947986	1133,705	0	0	0
7	1234,947986	1437,433387	1336,191	0	0	0
8	1437,433387	1639,918788	1538,676	0	0	0
9	1639,918788	1842,404189	1741,161	0	0	0
10	1842,404189	2044,88959	1943,647	1	1943,647	3777763,23
				0		
				1999	244460,1	33255303,8
	srednia=	122,2912145	srednia z szeregu przedzialowego (przyblizenie powstale w wyniku przeksztalcen)			
	srednia=	1032,462585	srednia bezposrednio z funkcji srednia (dokladna)			
	wariancja=	1680,828735				
	odchylenie st=	40,99791135				

Pierwszy lewy koniec wyznaczono jako wartość minimalną. Prawy koniec określono jako lewy koniec powiększony o długość jednego przedziału. Kolejne lewe końce są równe poprzednim prawym końcom, a prawy koniec każdego następnego przedziału wyznacza się, dodając długość przedziału do lewego końca. Kolumna x_i stanowi środek przedziału, n_i – liczbę obserwacji w danym przedziale, x_i*n_i wartość oczekiwaną (liczbę obserwacji w każdym z przedziałów), x_i^2*n_i - drugi moment

Wartości te zostały obliczone w celu wyliczenia średniej, wariancji i odchylenia standardowego z szeregu przedziałowego.

Po skonstruowaniu szeregu zauważono, że obserwacje w całym zbiorze danych są nierównomiernie rozłożone pomiędzy przedziały, przy czym większość obserwacji znajduje się w pierwszym przedziale. Może to sugerować obecność wartości odstających.

Dodatkowo, średnia obliczona z szeregu przedziałowego znacznie różni się od średniej obliczonej bezpośrednio za pomocą funkcji średniej. Taki rozrzut może wskazywać, że wartości odstające mają istotny wpływ na analizę danych.

Kolejnym etapem projektu jest uzupełnienie braków danych.

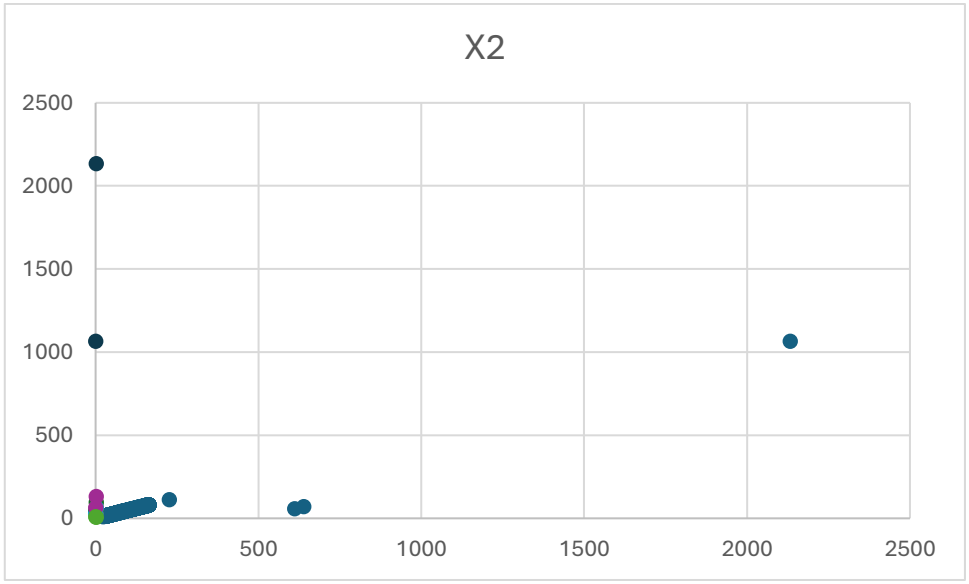
1 sposób uzupełniania danych

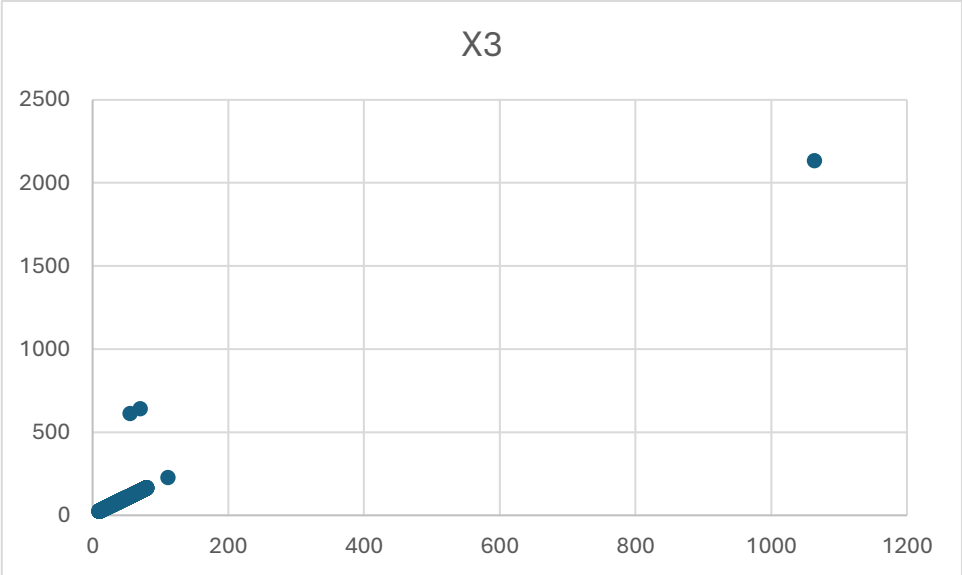
W pierwszej kolejności zostanie przeanalizowany najprostszy przypadek – korelacja liniowa pomiędzy zmiennymi. W przypadku występowania takiej zależności możliwe będzie oszacowanie brakujących wartości za pomocą równania regresji liniowej. W

związku z tym konieczne jest sprawdzenie, czy między zmiennymi istnieje liniowa zależność (plik: *Projekt2*, arkusz: dane początkowe).

	macierz korelacji						
	X1	X2	X3	X4	X5	X6	X7
X1	1	0,020652	0,019762	0,711768	0,001903	0,009155	0,009216
X2	0,020652	1	0,968607	-0,68669	-0,01305	0,013141	0,009597
X3	0,019762	0,968607	1	-0,66605	-0,01245	0,01476	0,011209
X4	0,711768	-0,68669	-0,66605	1	0,010781	-0,00227	0,000188
X5	0,001903	-0,01305	-0,01245	0,010781	1	-0,0108	-0,00458
X6	0,009155	0,013141	0,01476	-0,00227	-0,0108	1	0,824783
X7	0,009216	0,009597	0,011209	0,000188	-0,00458	0,824783	1

W żadnej parze różnych zmiennych nie zaobserwowano idealnie liniowej korelacji (tj. równej 1). Niemniej jednak bardzo wysoka korelacja występuje, na przykład, między zmiennymi X2 a X3. Analiza wykresu zależności tych dwóch zmiennych sugeruje możliwość istnienia liniowej zależności, choć widoczne są również wartości odstające, które mogą wpływać na wyniki analizy.





Mimo wszystko spróbujemy wyznaczyć równania regresji (plik Projekt2 arkusz Wyznaczanie X2 i X3).

Otrzymujemy, że:

$$X2 = 0{,}466878292 * X3 + 0{,}705809108$$

oraz

$$X3 = 2{,}009514229 * X2 + 4{,}553166778$$

Zastosowane rozwiązanie nie jest w pełni zadowalające, ponieważ nadal występują przypadki ekstremalnie wysokich błędów. Co istotne, błędy te pojawiają się w obserwacjach, gdzie nie występowały braki danych, a wartości zmiennych X2 i X3 nie były obliczane na podstawie pustych komórek.

Do identyfikacji tych wartości wykorzystano skalę kolorów, w której najwyższe błędy oznaczono kolorem czerwonym, a najniższe kolorem zielonym (szczegóły znajdują się w skoroszycie *Projekt2*, arkusz: *modele*).

Wniosek: W pierwszej kolejności należy usunąć wartości odstające.

2 sposób uzupełniania danych

Jako drugi sposób uzupełniania braków danych użyto interpolacji liniowej.

Interpolacja liniowa to metoda szacowania brakujących danych poprzez wyznaczanie wartości na podstawie linii prostej między sąsiadującymi znanymi wartościami. Sprawdza się szczególnie dobrze w danych, które zmieniają się w sposób liniowy lub prawie liniowy w krótkich przedziałach. Chociaż charakterystyka podanych danych nie

jest znana, warto podjąć próbę zastosowania tej techniki, ponieważ może przynieść zadowalające rezultaty.

Zasada działania interpolacji liniowej

Gdy brakuje wartości y dla pewnego punktu x, interpolacja liniowa wyznacza tę wartość na podstawie równania prostej pomiędzy dwoma sąsiadującymi punktami (x_1, y_1) i (x_2, y_2):

$$y = y_1 + \frac{y_2 - y_1}{x_2 - x_1} \cdot (x - x_1)y$$

Interpolacja liniowa może być używana do uzupełniania braków danych, zwłaszcza gdy dane mają naturalną tendencję do zmienności liniowej w krótkich odstępach czasu lub przestrzeni.

Zalety:

1. **Prostota i efektywność:** jest to szybka i łatwa metoda obliczeniowa.
2. **Zachowanie trendów:** w przypadku małych luk interpolacja liniowa często dobrze odtwarza lokalny trend danych.
3. **Unikanie przesunięć:** interpolacja liniowa nie przesuwa globalnych statystyk danych.

<u>Zastosowanie interpolacji w KNIME</u>

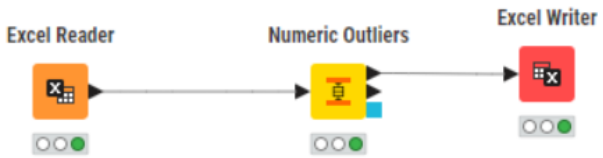
W KNIME do uzupełniania braków danych można użyć node'a Missing Value, który oferuje różne strategie, w tym interpolację liniową (folder z workspace'em w knime - projekt_missing_values). Sposób działania tej metody jest następujący:

1. KNIME analizuje dane kolumnowo, szukając brakujących wartości w danej kolumnie.
2. Dla każdej brakującej wartości:
    - Znajduje najbliższe dostępne wartości powyżej i poniżej brakującego miejsca.
    - Wylicza wartość zgodnie z równaniem interpolacji liniowej.
3. Jeśli brakująca wartość znajduje się na początku lub końcu danych (nie ma jednego z sąsiadów), stosuje inną wybraną strategię (np. wartość domyślną).

Wynikowe zmienne, uzyskane za pomocą interpolacji liniowej, zostały zapisane w pliku odstające_interpolacja i następnie skopiowane do skoroszytu Projekt2, arkusz: modele. Obliczono różnice (błędy) pomiędzy oryginalnymi zmiennymi a zmiennymi uzyskanymi z modelu. Dla każdej obserwacji, poza tymi, w których występowały braki danych, błędy te wyniosły dokładnie 0.

PODEJŚCIE 2 – identyfikacja wartości odstających i zastąpienie ich poprzez braki danych

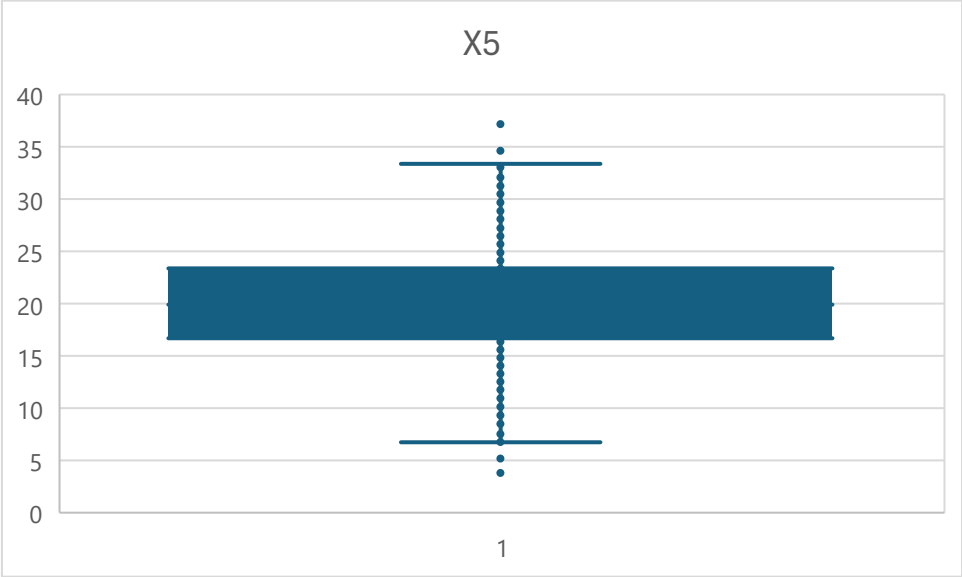
Wartości odstające zostaną usunięte w KNIME poprzez ich zastąpienie brakami danych. Następnie dane te zostaną zapisane do nowego pliku, aby umożliwić dalszą analizę (folder do wglądu – projekt_odstajace).

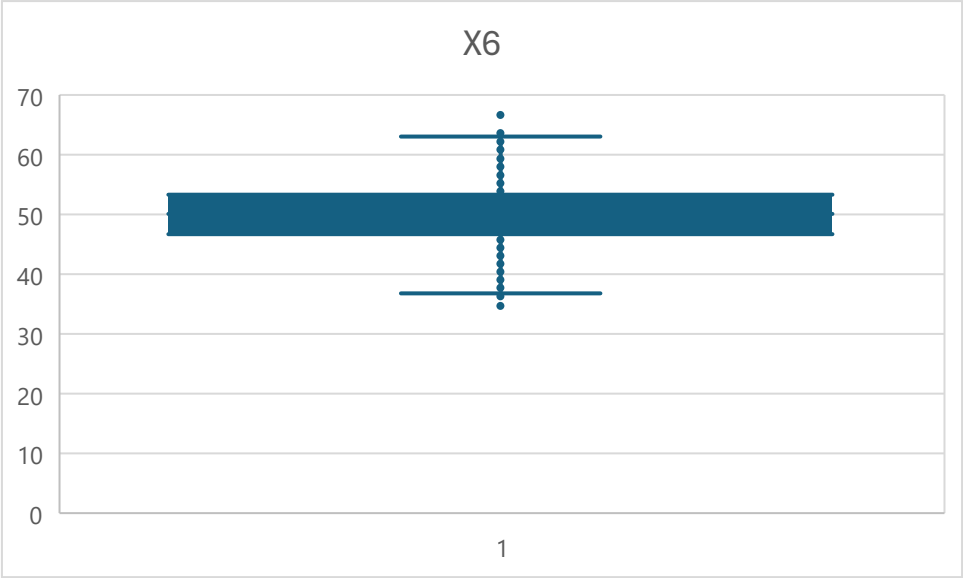


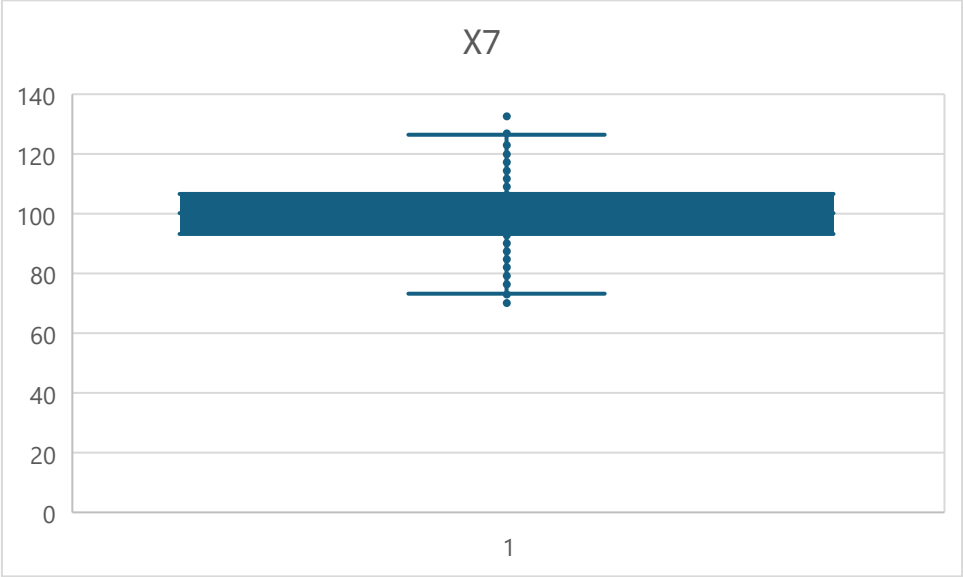
Usunięte zostaną wszystkie obserwacje spoza przedziału

$$(\mu - 3\sigma, \mu + 3\sigma)$$

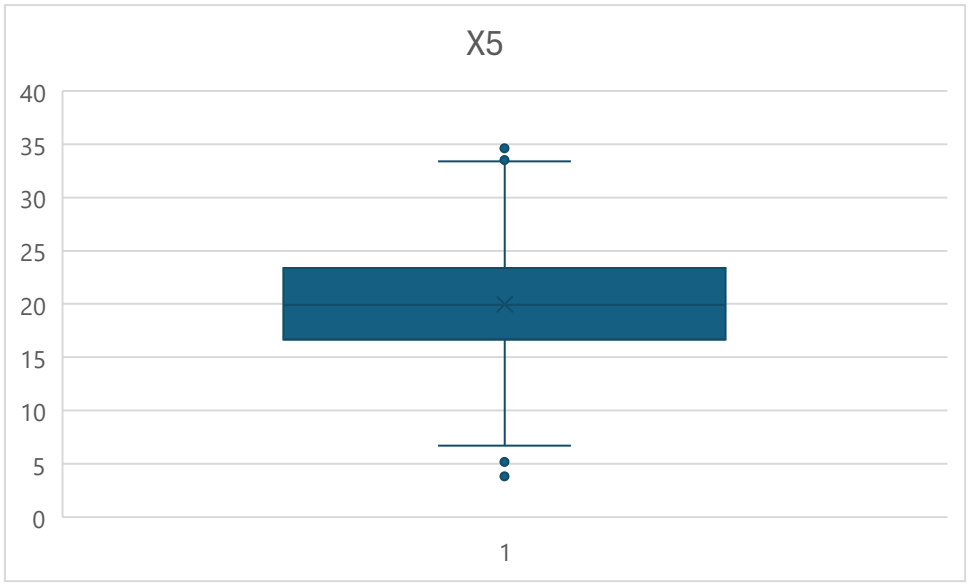
Jednak przy podejściu z 3σ wciąż występują obserwacje odstające w zmiennych X5, X6, X7 (plik odstajace3).

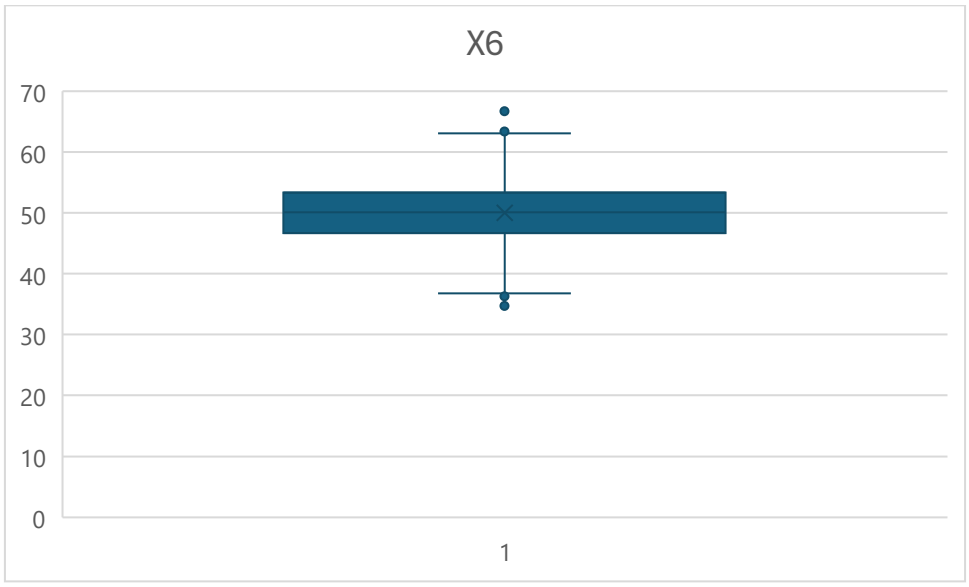


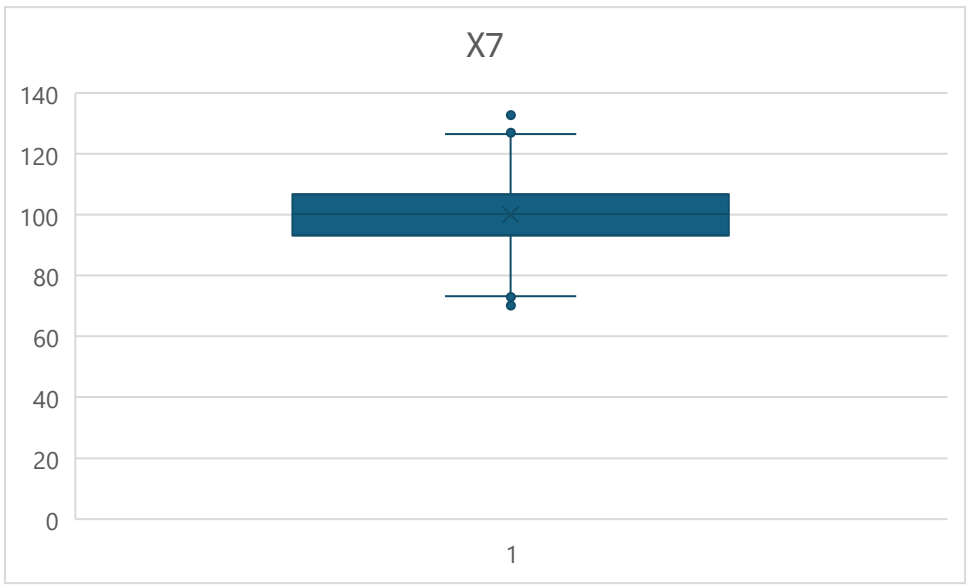




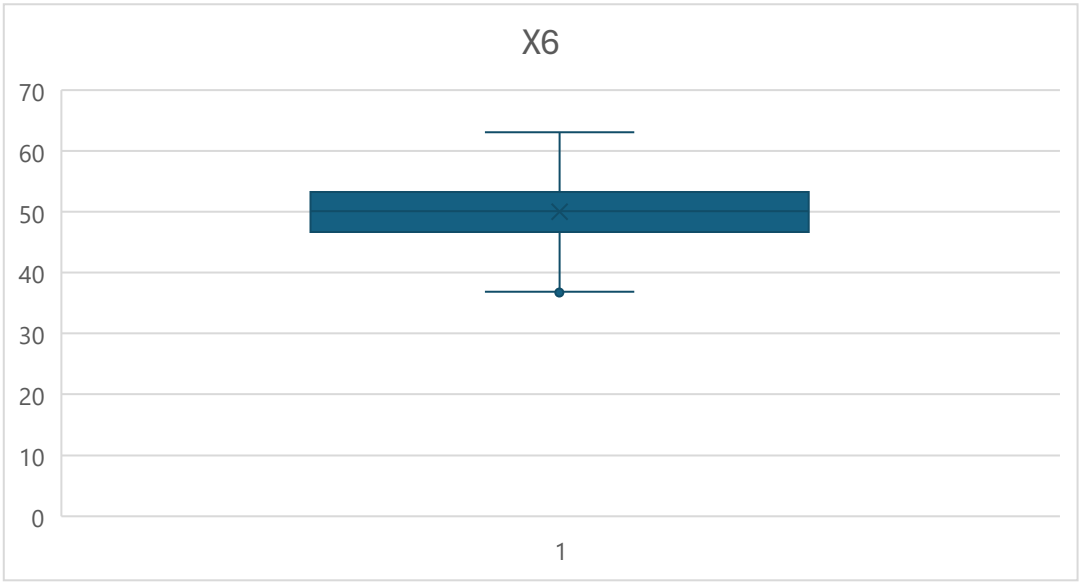
Próba usunięcia wartości odstających za pomocą kryterium 2σ wykazała, że wciąż występują obserwacje odstające, co zostało zapisane w pliku odstajace2.

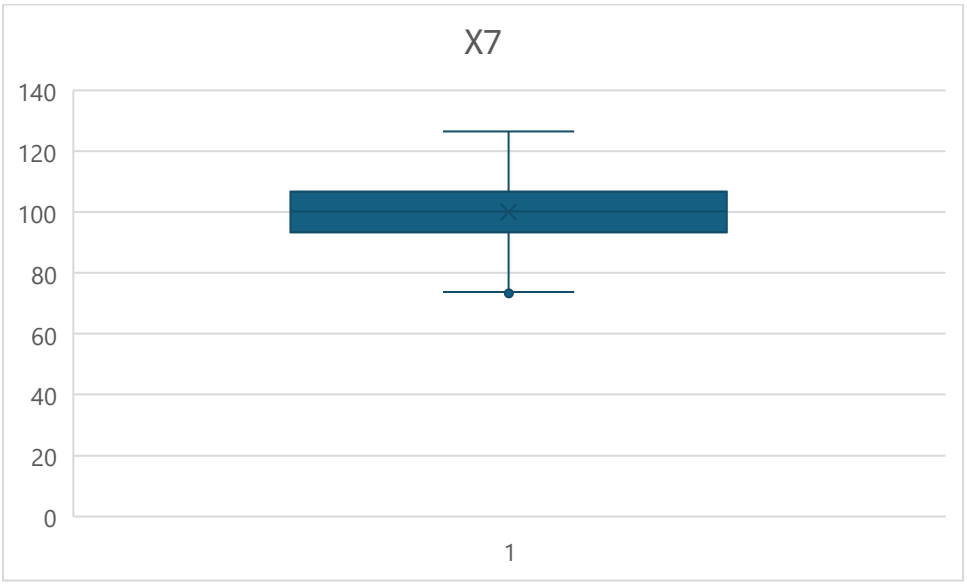

X5
40
35
30
25
20
15
10
5
0
1


X6
70
60
50
40
30
20
10
0
1

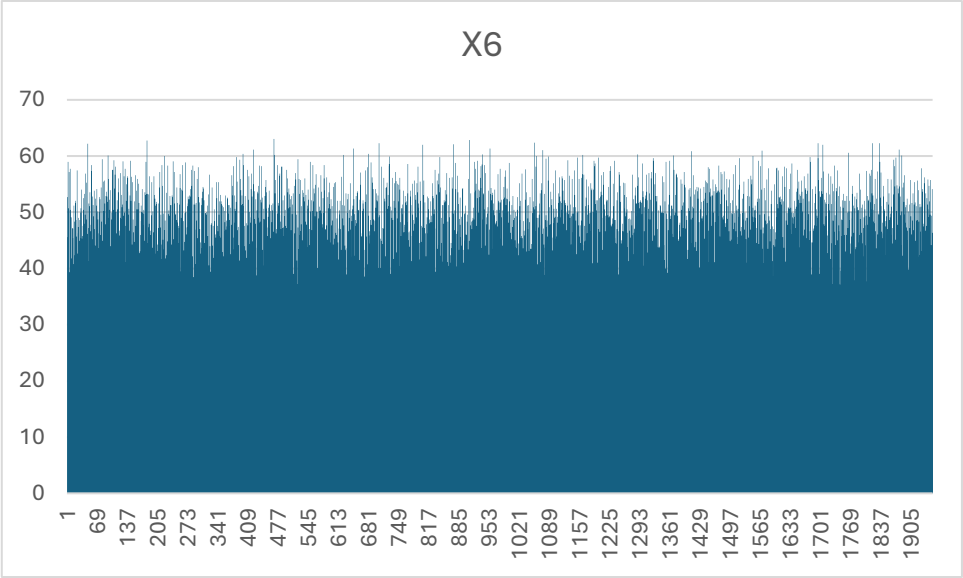

X7
140
120
100
80
60
40
20
0
1

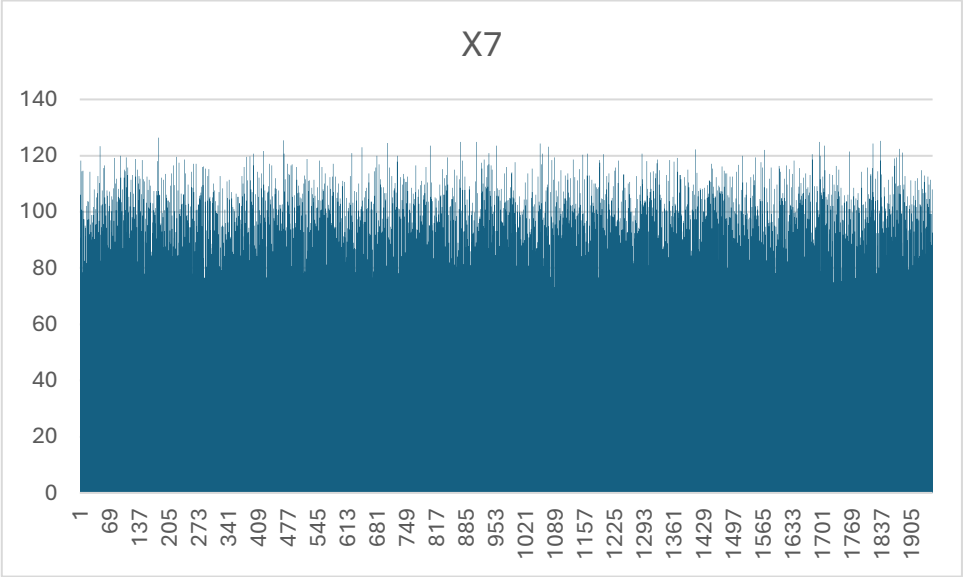
Podjęto zatem próbę usunięcia wartości odstających za pomocą kryterium 1,5σ (plik odstajace1.5). W wyniku analizy stwierdzono, że dla zmiennych X6 i X7 pojawiły się obserwacje na pograniczu (widoczne na boxplotach), które jednak nie zostały uznane za wartości odstające. Ostatecznie usunięto 45 obserwacji (2,25%), co stanowi stosunkowo niewielki odsetek. Usunięte obserwacje mogą zawierać błędne pomiary lub inne nieprawidłowe dane, które mogłyby zaburzać model, dlatego ich eliminacja może przyczynić się do poprawy jakości analizy.



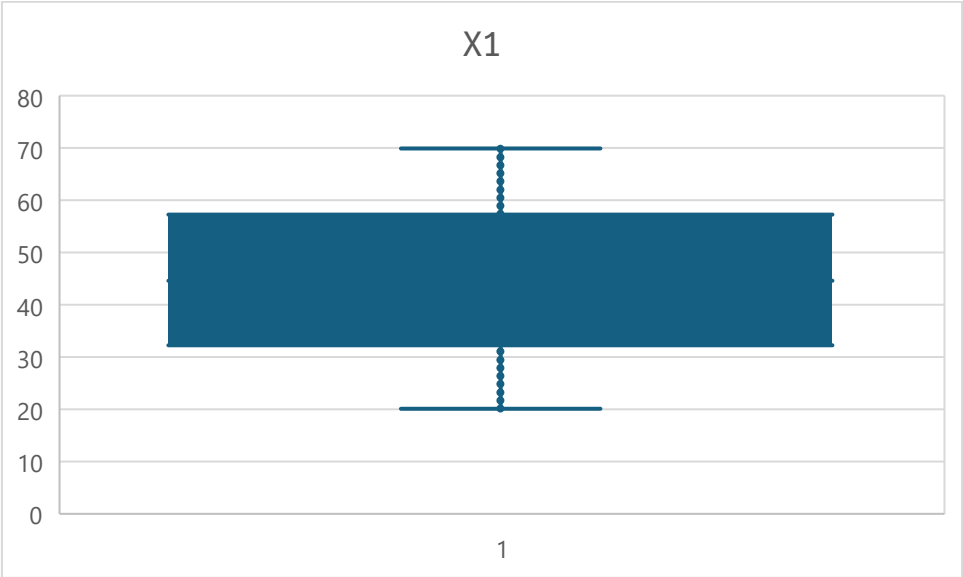


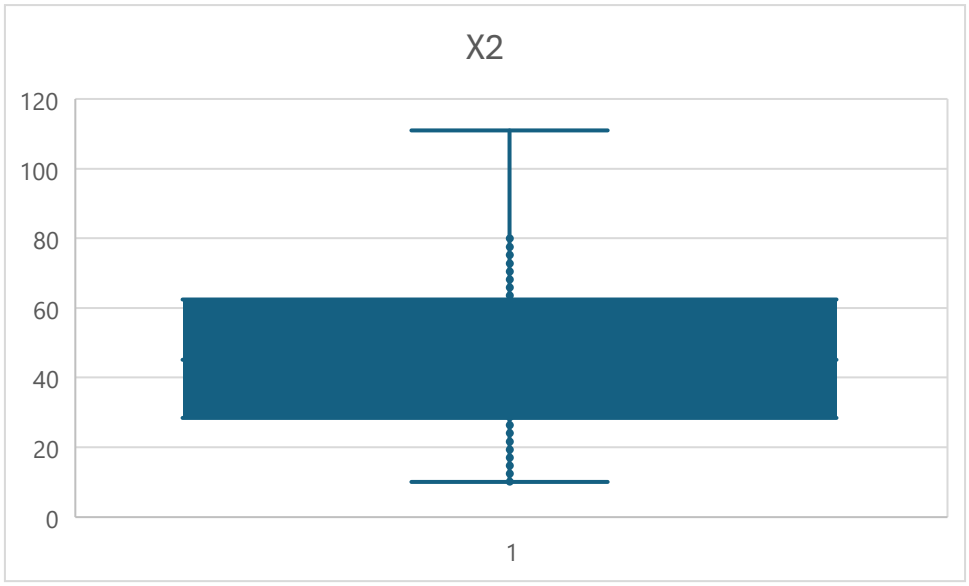
Z histogramów również nie stwierdzono znaczących wartości odstających.

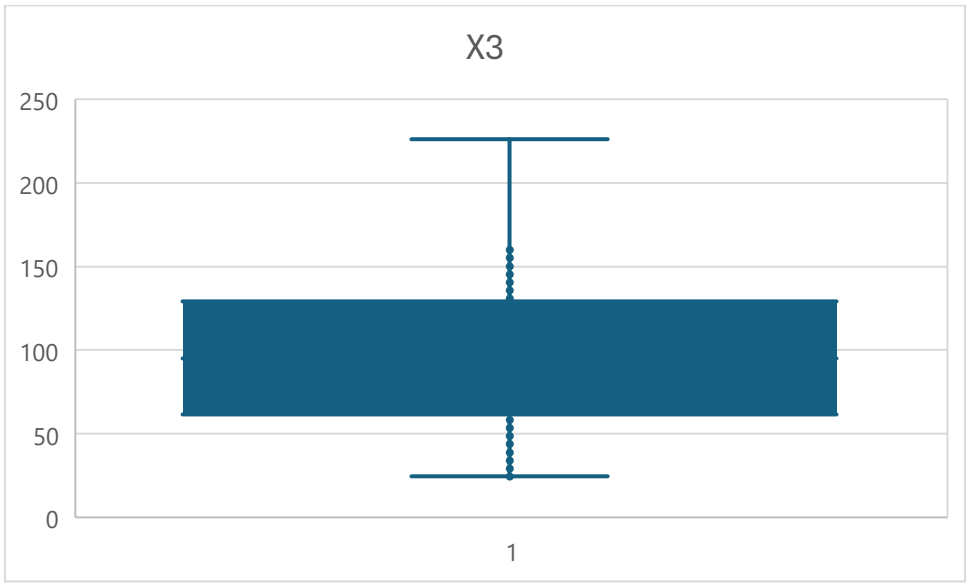


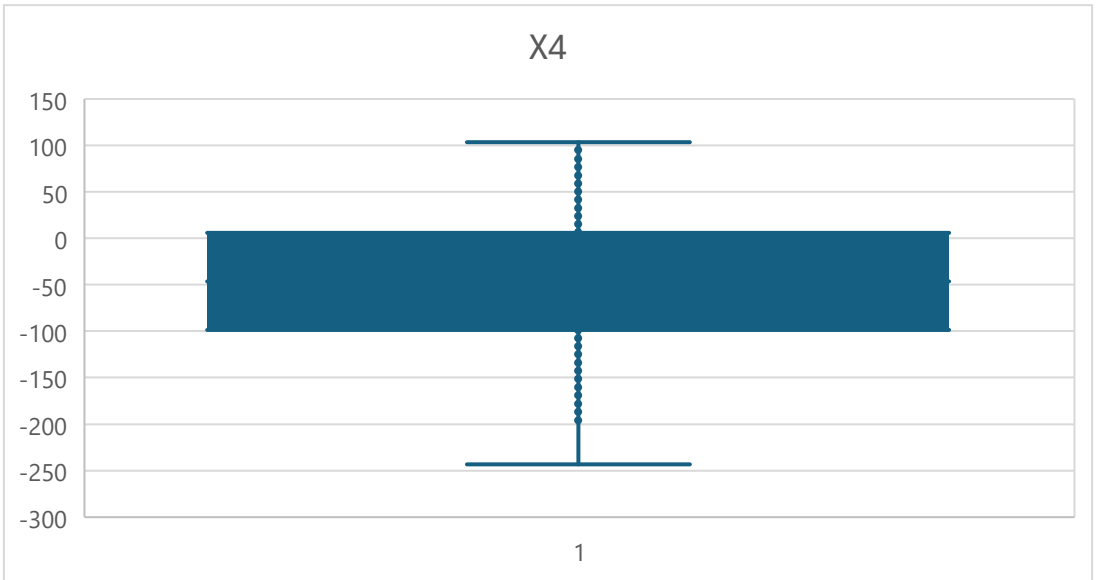


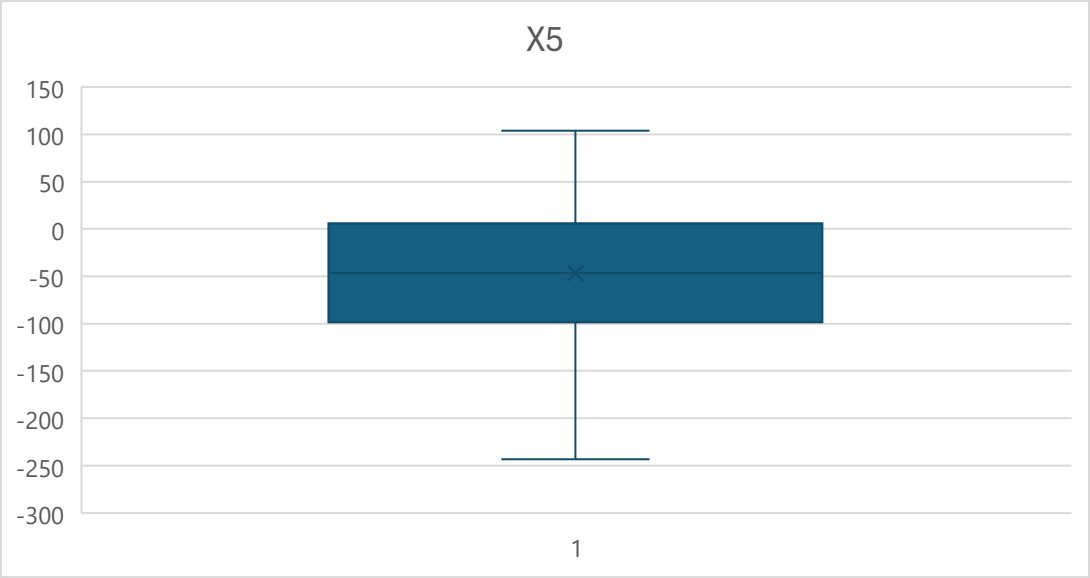
Dla pozostałych zmiennych również nie stwierdzono wartości odstających:




X2
120
100
80
60
40
20
0
1


X3
250
200
150
100
50
0
1

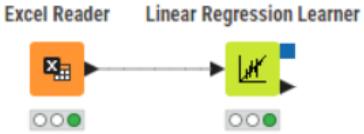

X4
150
100
50
0
-50
-100
-150
-200
-250
-300
1



Po usunięciu wartości odstających macierz korelacji prezentuje się lepiej w porównaniu do sytuacji, w której wartości odstające nie zostały usunięte. Zauważalne są wyższe korelacje między zmiennymi oraz większa liczba zmiennych, które można wyznaczyć za pomocą regresji liniowej.

korelacje	X1	X2	X3	X4	X5	X6	X7
X1	1	0,012141	0,013574	0,421653	-0,0208	-0,00031	0,002647
X2	0,012141	1	0,999975	-0,89993	-0,01524	0,023539	0,021706
X3	0,013574	0,999975	1	-0,90037	-0,01726	0,01954	0,017861
X4	0,421653	-0,89993	-0,90037	1	0,006163	-0,01799	-0,01526
X5	-0,0208	-0,01524	-0,01726	0,006163	1	0,008842	0,006371
X6	-0,00031	0,023539	0,01954	-0,01799	0,008842	1	0,998259
X7	0,002647	0,021706	0,017861	-0,01526	0,006371	0,998259	1

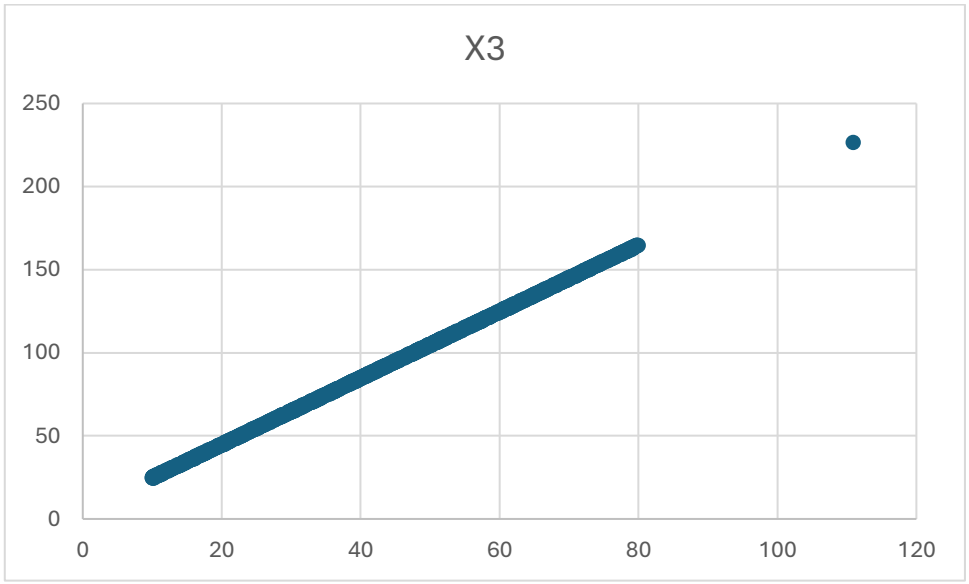
Zmienna X1 wykazuje dość słabe korelacje z pozostałymi zmiennymi, z wyjątkiem umiarkowanej korelacji między X1 a X4. Mimo tego, w programie KNIME udało się wyznaczyć tę zmienną za pomocą równania regresji liniowej, wykorzystując wszystkie pozostałe zmienne jako zmienne niezależne. Proces ten został zrealizowany przy użyciu node'ów Excel Reader oraz Linear Regression Learner (folder do wglądu – projekt_uzupelniania).

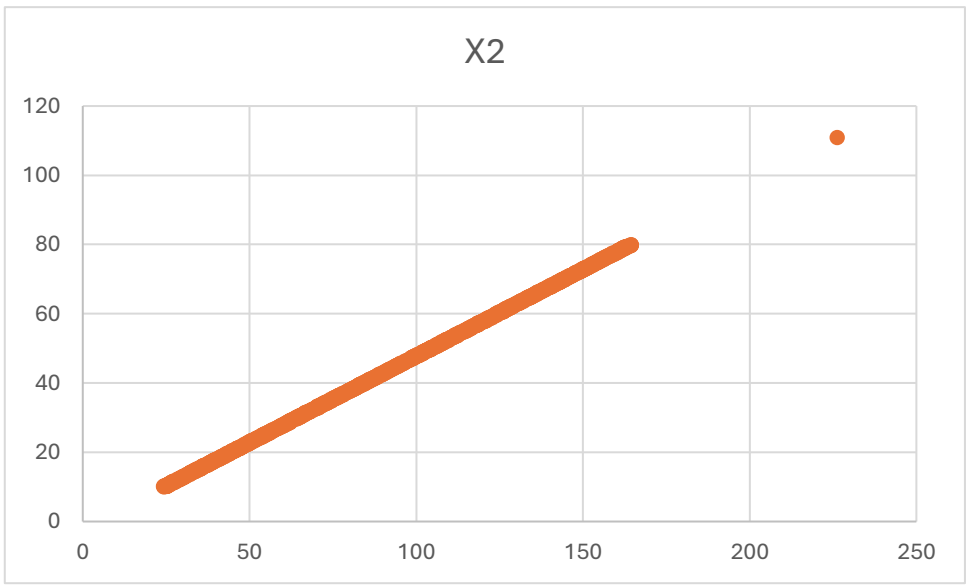


Na poniższym zrzucie ekranu przedstawiono uzyskane wartości współczynników.

#	RowID	Variable String	Coeff. Number (double)	Std. Err. Number (double)	t-value Number (double)	P>|t| Number (double)
1	Row1	X2	1.5	0	80,021,998,377,284.14	0
2	Row2	X3	0	0	1.861	0.063
3	Row3	X4	0.5	0	5,499,113,227,014,182	0
4	Row4	X5	0	0	0.163	0.871
5	Row5	X6	0	0	1.452	0.147
6	Row6	X7	-0	0	-1.335	0.182
7	Row7	Intercept	-0	0	-3.523	0

Zmienne X2 i X3 zostały oszacowane za pomocą równania liniowego, ponieważ korelacja między nimi wynosi 0,99, co wskazuje na niemal idealną zależność liniową. Analiza wykresów zależności potwierdza, że taki sposób modelowania jest uzasadniony (szczegóły w pliku „odstające1.5”, arkusz „wyznaczanie X2, X3, X6, X7”).





Równania dla tych zmiennych wyglądają następująco:

$$X2 = 0{,}499826516 * X3 - 2{,}228939252$$

oraz

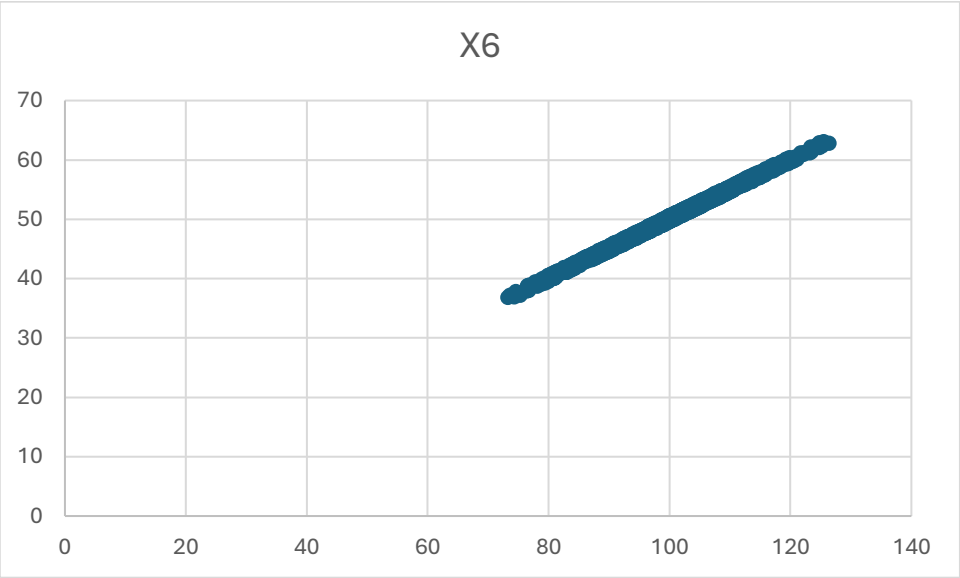
$$X3 = 2{,}000595896 * X2 + 4{,}463870912$$

Zmienną X4 wyznaczono w analogiczny sposób jak X1.

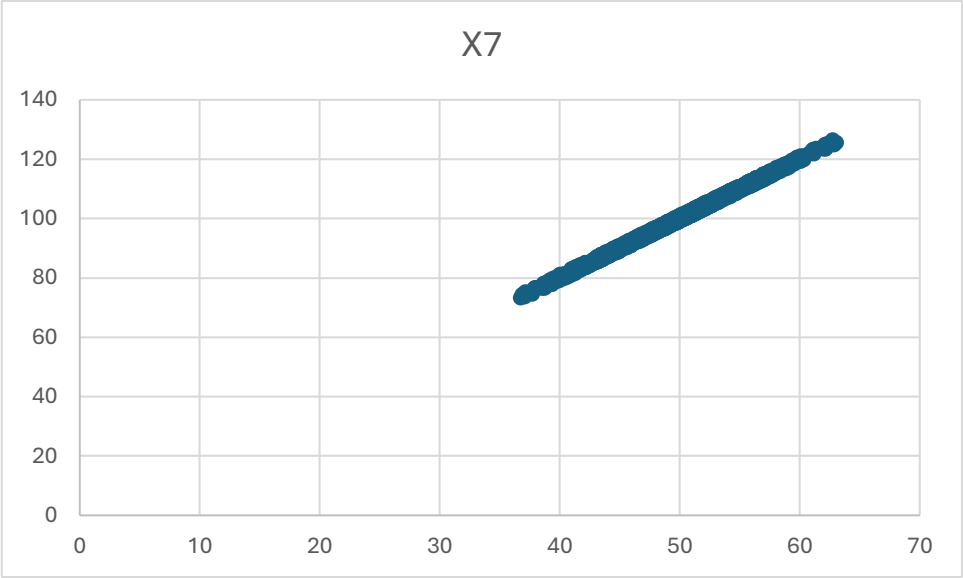
#	RowID	Variable *String*	Coeff. *Number (double)*	Std. Err. *Number (double)*	t-value *Number (double)*	P>|t| *Number (double)*
1	Row1	X1	2	0	4,985,385,411,269,382	0
2	Row2	X2	-3	0	-72,507,974,140,203.17	0
3	Row3	X3	-0	0	-0.06	0.952
4	Row4	X5	-0	0	-0.04	0.968
5	Row5	X6	-0	0	-0.416	0.677
6	Row6	X7	0	0	0.297	0.766
7	Row7	Intercept	-0	0	-1.859	0.063

Z kolei zmienne X6 i X7 w analogiczny sposób jak X2 i X3.

$$X6 = 0{,}49871841 * X7 + 0{,}123676986$$



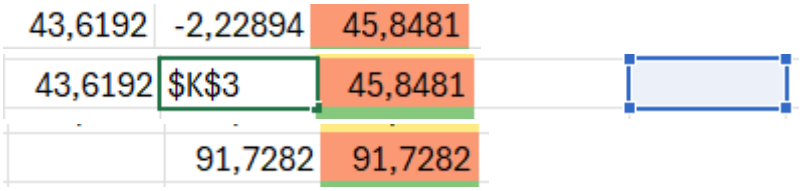
$$X7 = 1{,}99814061 * X6 + 0{,}102206924$$



Zmienna X5 wykazuje bardzo niskie korelacje z pozostałymi zmiennymi, a te korelacje są praktycznie nieistotne. W związku z tym, dla zmiennej X5 zostanie zastosowana prostsza metoda uzupełniania braków danych, mianowicie zastąpienie brakujących wartości medianą wszystkich dostępnych obserwacji.

Po uzupełnieniu braków, błędy w poszczególnych modelach wyglądają następująco (plik do wglądu – odstajace1.5 arkusz modele):

- X1 i X4 – bledy wszedzie wychodza 0 (oprocz oczywiście miejsc, gdzie te zmienne miały braki danych)
- X2 i X3 oraz X6 i X7– wciąż jest ekstremalnie wysokich wartości, jednak one wychodzą dlatego, że zostały policzone na podstawie komórek gdzie były puste wartości, bądź jest akurat liczony brak danych, np.:



- X5 – tutaj nie wyznaczalismy bledu gdyz wszystkie puste komorki zostały po prostu zastąpione mediana

Po uzupełnieniu braków danych, wyznaczone zostały statystyki opisowe, które zostały zapisane w pliku odstajace1.5, arkusz: uzupelnione dane.

statystyki	X1	X2	X3	X4	X5	X6	X7
średnia	44,68718	45,10317	94,94466	-46,4516	20,00225	49,57867	99,45815
odchylenie	14,62824	20,28219	40,38327	66,85082	4,875201	6,839131	12,59919
mediana	44,62754	45,0145	94,77154	-46,585	19,90447	50,10593	100,2346
Q1	32,35959	28,33277	61,34199	-98,827	16,74655	46,58509	93,31972
Q3	57,22245	62,3287	129,4389	5,864433	23,34535	53,30566	106,6211
Q0	0	-2,22894	4,463871	-243,471	6,696283	0,123677	0,102207
Q4	117,2287	110,8627	226,3909	103,696	33,3693	63,17246	126,0818
dominanta	0	-2,22894	4,463871	#N/D	19,90447	0,123677	0,102207

Dla poszczególnych zmiennych mamy następujace wnioski:

- X1: rozkład danych jest symetryczny, ponieważ średnia jest zbliżona do mediany, co sugeruje brak dużych wartości odstających. Większość wartości mieści się w zakresie od 32,36 do 57,22, z pojedynczymi wartościami znajdującymi się poza tym zakresem. Dominanta wynosi 0, co wskazuje, że wartość 0 pojawia się najczęściej w analizowanych danych.
- X2: rozkład danych jest niemal symetryczny, ponieważ średnia jest zbliżona do mediany. Wartości są bardziej rozproszone niż w przypadku zmiennej X1, co sugeruje większe odchylenie standardowe. Dominanta wynosi -2,23, co może wskazywać na specyficzne zjawisko w danych, takie jak częste występowanie ujemnych wartości.
- X3: rozkład danych jest symetryczny, ponieważ średnia jest zbliżona do mediany. Dane są rozproszone, a odchylenie standardowe wynosi 40,38, co stanowi ponad 40% wartości średniej. Często występują wartości małe, a dominanta wynosi 4,46, co może wskazywać na asymetrię rozkładu i sugerować, że rozkład może być przesunięty w stronę niższych wartości.
- X4:rozkład danych jest silnie rozproszony, co wynika z wysokiego odchylenia standardowego. Istnieje prawdopodobieństwo, że rozkład danych jest niesymetryczny, z wyraźnym przesunięciem w stronę wyższych wartości. O tym świadczy różnica między średnią a medianą, a także fakt, że trzeci kwartyl (Q3) przyjmuje wartości dodatnie, co sugeruje przewagę wyższych obserwacji w danych. Brak dominującej wartości (oznaczone jako #N/D) wskazuje na równomierne rozłożenie danych w tej kolumnie.
- X5: rozkład danych jest symetryczny, ponieważ średnia jest zbliżona do mediany. Dane są skupione blisko średniej, co wskazuje na niskie odchylenie standardowe. Często występuje wartość 19,90, która jest dominantą, co oznacza, że ta wartość pojawia się najczęściej i jest równa medianie.
- X6: rozkład danych jest symetryczny, ponieważ średnia jest zbliżona do mediany. Dane są stabilne i mało rozproszone, co wskazuje na niskie odchylenie standardowe (wynoszące około 14% średniej). Często występuje wartość 0,12,

która jest dominantą, co może sugerować specyficzne cechy danych, takie jak częste występowanie zerowych lub bardzo małych wartości.

- X7: rozkład danych jest symetryczny, ponieważ średnia jest zbliżona do mediany. Dane są bardziej rozproszone niż w przypadku zmiennej X6, ale nadal stosunkowo stabilne. Często występuje wartość 0,10, która jest dominantą, co, podobnie jak w przypadku zmiennej X6, może wskazywać na specyficzne cechy danych, takie jak częste występowanie tej wartości w zbiorze danych.

Podsumowując: kolumny X1, X2, X3, X5, X6 i X7 mają średnią zbliżoną do mediany, co sugeruje symetryczność rozkładu. Wyjątkiem jest X4, która może mieć asymetryczny rozkład. Zmienne X4 i X3 mają największe rozproszenie (wysokie odchylenie standardowe), co wskazuje na dużą zmienność w wartościach. Zmienne X5 i X6 są najbardziej stabilne, z najmniejszym odchyleniem standardowym. Kolumny X3 i X4 mają najszersze zakresy wartości (Q0 do Q4), co potwierdza ich dużą zmienność.

Następnie skonstruowano szereg rozdzielczy przedziałowy (plik odstajace 1.5 arkusz szereg).

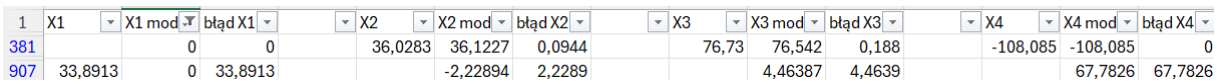
liczba obserwacji n=	2000
x_min=	0
x_max=	117,2287
rozpietosc R=	117,2287
dlugośc jednego przedziału h=	11,72287

szereg						
nr przedzialu	lewy koniec	prawy koniec	x_i	n_i	x_i*n_i	x_i^2*n_i
1	0	11,7228711	5,861436	2	11,72287	68,71285
2	11,72287114	23,4457423	17,58431	149	2620,062	46071,97
3	23,44574228	35,1686134	29,30718	476	13950,22	408841,5
4	35,16861342	46,8914846	41,03005	471	19325,15	792912
5	46,89148456	58,6143557	52,75292	456	24055,33	1268989
6	58,6143557	70,3372268	64,47579	445	28691,73	1849922
7	70,33722684	82,060098	76,19866	0	0	0
8	82,06009798	93,7829691	87,92153	0	0	0
9	93,78296912	105,50584	99,6444	0	0	0
10	105,5058403	117,228711	111,3673	1	111,3673	12402,67
				0		
				2000	88765,58	4379208
	srednia=	44,3827901	z szeregu			
	srednia=	58,6143557	dokladna			
	wariancja=	219,771742				
	odchylenie st=	14,8247004				

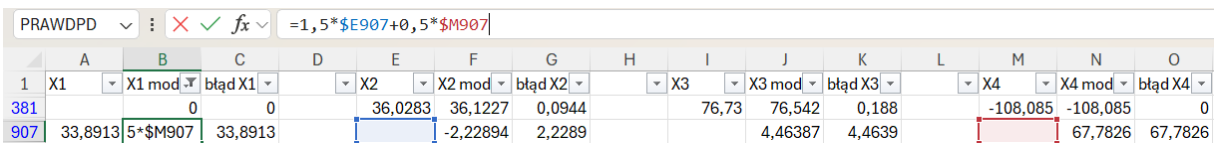
Po uzupełnieniu braków danych zauważono poprawę w rozłożeniu danych pomiędzy poszczególne przedziały oraz w zgodności średniej z szeregu z dokładną średnią. Mimo to, obserwacje nadal nie są równomiernie rozłożone, a średnia z szeregu i średnia dokładna wciąż się różnią.

Dodatkowo, zauważono, że prawdopodobnie wciąż istnieje wartość odstająca, ponieważ w przedziałach nr 7, 8 i 9 nie ma żadnych obserwacji, a w przedziale 10 znajduje się tylko jedna obserwacja. W związku z tym zindentyfikowano tą wartość – jest to po prostu wartość największa,a następnie zastąpiono tę pojedynczą obserwację medianą.

Ponadto, jedna obserwacja została usunięta, ponieważ zmienna X1 = 0 została obliczona na podstawie braków danych (dotyczy to dokładnie wiersza 907, który został uznany za podejrzany w arkuszu modele).



1	X1	X1 mod	błąd X1		X2	X2 mod	błąd X2		X3	X3 mod	błąd X3		X4	X4 mod	błąd X4
381		0	0		36,0283	36,1227	0,0944		76,73	76,542	0,188		-108,085	-108,085	0
907	33,8913	0	33,8913			-2,22894	2,2289			4,46387	4,4639			67,7826	67,7826



PRAWDPD =1,5*$E907+0,5*$M907

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	X1	X1 mod	błąd X1		X2	X2 mod	błąd X2		X3	X3 mod	błąd X3		X4	X4 mod	błąd X4
381		0	0		36,0283	36,1227	0,0944		76,73	76,542	0,188		-108,085	-108,085	0
907	33,8913	5*$M907	33,8913			-2,22894	2,2289			4,46387	4,4639			67,7826	67,7826

Po tych zmianach szereg wygląda następująco:

			szereg						
liczba obserwacji n=	2000		nr przedzialu	lewy koniec	prawy koniec	x_i	n_i	x_i*n_i	x_i^2*n_i
x_min=	0		1	0	6,99563372	3,497817	1	3,497817	12,23472
x_max=	69,95634		2	6,995633723	13,9912674	10,49345	0	0	0
rozpietosc R=	69,95634		3	13,99126745	20,9869012	17,48908	42	734,5415	12846,46
dlugośc jednego przedziału h=	6,995634		4	20,98690117	27,9825349	24,48472	279	6831,236	167260,9
			5	27,98253489	34,9781686	31,48035	291	9160,782	288384,7
mediana=	44,62754		6	34,97816862	41,9738023	38,47599	300	11542,8	444120,4
			7	41,97380234	48,9694361	45,47162	264	12004,51	545864,4
			8	48,96943606	55,9650698	52,46725	263	13798,89	723989,7
			9	55,96506979	62,9607035	59,46289	277	16471,22	979426,3
			10	62,96070351	69,9563372	66,45852	282	18741,3	1245519
							0		
							1999	89288,77	4407424
				srednia=	44,6667189	z szeregu			
				srednia=	34,9781686	dokladna			
				wariancja=	209,698785				
				odchylenie st=	14,4809801				

Wartość X1=0 prawdopodobnie również stanowi obserwację odstającą. Świadczy o tym fakt, że w pierwszym przedziale znajduje się tylko jedna obserwacja, podczas gdy kolejna większa liczba obserwacji występuje dopiero w przedziale trzecim. W związku z tym wartość ta została zastąpiona medianą.

Po wprowadzeniu tych zmian szereg przedstawia się w bardziej uporządkowanej formie (arkusz szereg(2)).

			szereg						
liczba obserwacji n=	2000		nr przedzialu	lewy koniec	prawy koniec	x_i	n_i	x_i*n_i	x_i^2*n_i
x_min=	20,03558		1	20,03558053	25,0276562	22,53162	209	4709,108	106103,8
x_max=	69,95634		2	25,0276562	30,0197319	27,52369	210	5779,976	159086,3
rozpietosc R=	49,92076		3	30,01973187	35,0118075	32,51577	193	6275,544	204054,1
dlugośc jednego przedziału h=	4,992076		4	35,01180754	40,0038832	37,50785	220	8251,726	309504,5
			5	40,00388321	44,9959589	42,49992	180	7649,986	325123,8
mediana=	44,62754		6	44,99595888	49,9880346	47,492	209	9925,827	471397,4
			7	49,98803455	54,9801102	52,48407	179	9394,649	493069,4
			8	54,98011022	59,9721859	57,47615	210	12069,99	693736,6
			9	59,97218589	64,9642616	62,46822	188	11744,03	733628,4
			10	64,96426156	69,9563372	67,4603	200	13492,06	910178,4
							1		
							1999	89292,89	4405883
				srednia=	44,6687807	z szeregu			
				srednia=	44,9959589	dokladna			
				wariancja=	208,743419				
				odchylenie st=	14,4479555				

Dane są równomiernie rozłożone a średnie z szeregu i dokładna są do siebie zbliżone.

Wnioski z projektu

Analiza danych pokazała, jak istotne znaczenie mają wartości odstające, ponieważ mogą one znacząco wpływać na wyniki badań. Możliwość przeprowadzenia bardziej precyzyjnej analizy byłaby większa, gdyby znany był charakter badanych danych. Niestety, w tym przypadku losowy charakter danych ograniczył takie możliwości. Warto jednak zauważyć, że bardziej zaawansowane metody, takie jak interpolacja, często okazują się skuteczniejsze, nawet jeśli dane nie spełniają wszystkich formalnych wymagań do ich zastosowania.