PROJEKT Z ZAKRESU ANALIZY DANYCH

27 listopada 2024

Emil Szewczak
Kierunek: Inżynieria i Analiza Danych

Wstęp

Niniejszy projekt ma na celu przeprowadzenie kompleksowej analizy danych, obejmującej wyznaczenie podstawowych statystyk, budowę szeregu rozdzielczego, uzupełnianie braków danych oraz przygotowanie szczegółowego raportu. Analiza została przeprowadzona w oparciu o dostarczony zbiór danych i ma na celu zarówno pogłębienie umiejętności analizy danych, jak i praktyczne zastosowanie narzędzi wspierających proces eksploracji i przetwarzania danych.

Cele projektu:

1. Wyznaczenie i interpretacja podstawowych statystyk dla każdej zmiennej, takich jak średnia, mediana, odchylenie standardowe czy wartości skrajne, co pozwoli na lepsze zrozumienie struktury danych.
2. Budowa szeregu rozdzielczego przedziałowego dla zmiennej X1, składającego się z 10 przedziałów, wraz z wyznaczeniem odpowiednich miar statystycznych.
3. Uzupełnienie braków danych w zbiorze, co umożliwia dalsze wykorzystanie danych w analizie bez utraty wartości informacyjnej.
4. Opracowanie szczegółowego raportu dokumentującego wykonane kroki, zastosowane formuły oraz narzędzia, wraz ze zrzutami ekranowymi ilustrującymi proces analizy.

Wykorzystane narzędzia:

Do realizacji projektu wykorzystano dwa kluczowe narzędzia:

- Microsoft Excel – program do analizy danych, w którym przeprowadzono obliczenia statystyczne, budowę szeregu rozdzielczego oraz uzupełnianie braków danych w sposób manualny i automatyczny.
- KNIME Analytics Platform – platforma do zaawansowanej analizy danych, umożliwiająca automatyzację procesów analitycznych, takich jak wyznaczanie statystyk czy wizualizacja wyników, co pozwala na efektywną i intuicyjną analizę nawet dużych zbiorów danych.

Projekt łączy w sobie zarówno klasyczne, jak i nowoczesne podejście do analizy danych, wykorzystując możliwości obu narzędzi w celu uzyskania precyzyjnych wyników i optymalizacji procesu analitycznego.

Spis treści

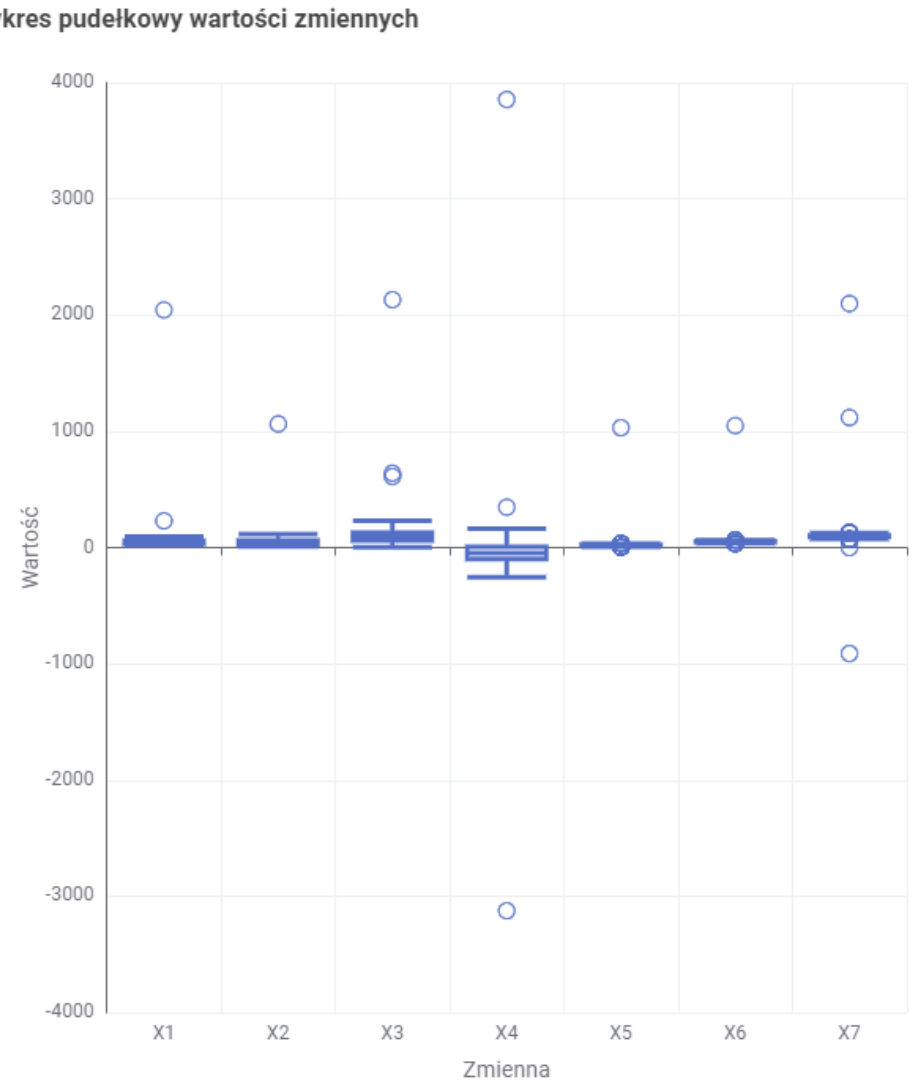
Przygotowanie danych

Na potrzeby analizy otrzymałem zbiór danych charakteryzujący się następującymi właściwościami:

- **Liczba zmiennych**: 7
- **Liczba obserwacji dla każdej zmiennej**: 2000

Wizualizacja danych

Aby przeanalizować rozkład wartości dla każdej zmiennej, przygotowałem wykres pudełkowy (*box plot*). Do jego stworzenia wykorzystałem węzeł `Box Plot` dostępny w programie **Knime**. Wykres przedstawia pełny rozkład wartości, w tym potencjalne wartości odstające.

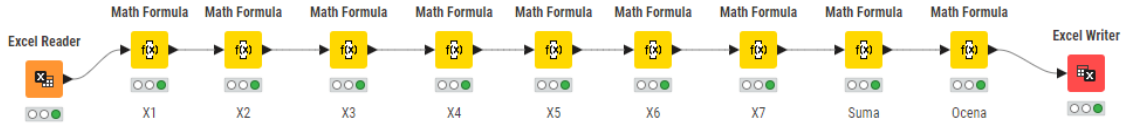


Rysunek 1: Wykres pudełkowy wartości zmiennych przed usunięciem wartości odstających

Jak widać na powyższym wykresie, w każdej zmiennej występują wartości odstające.

Identyfikacja wartości odstających

W celu zidentyfikowania wartości odstających skorzystałem z węzła `Math Formula` w programie **Knime**. Algorytm opiera się na formule omówionej podczas zajęć, a jego szczegóły zostały przedstawione na poniższym schemacie:



Rysunek 2: Schemat węzłów wykorzystanych do identyfikacji wartości odstających

W każdym węźle `Math Formula`, oznaczonym nazwami zmiennych, zastosowałem funkcję do identyfikacji wartości odstających. Jej działanie przedstawia poniższy obraz:

Expression

```
1 if(abs($X1$)<COL_MEAN($X1$)+COL_STDDEV($X1$),0,
2 if(abs($X1$)<COL_MEAN($X1$)+2*COL_STDDEV($X1$),0.1,
3 if(abs($X1$)<COL_MEAN($X1$)+3*COL_STDDEV($X1$),0.5,1)))
```

Rysunek 3: Węzeł `Math Formula` wykorzystany do identyfikacji wartości odstających

Formuła wykorzystuje zagnieżdżone instrukcje warunkowe (*if*) w celu przypisania odpowiedniej wartości w zależności od tego, jak daleko obserwacja $X1$ (lub innej zmiennej) znajduje się od średniej arytmetycznej, w jednostkach odchylenia standardowego. Kolejne warunki formuły sprawdzają:

- Jeśli wartość $X1$ mieści się w zakresie jednego odchylenia standardowego od średniej ($\mu \pm \sigma$), wynik to **0** (brak odstępstwa).
- Jeśli wartość $X1$ leży pomiędzy $\mu \pm \sigma$ a $\mu \pm 2\sigma$, wynik to **0.1** (niewielkie odstępstwo).
- Jeśli wartość $X1$ leży pomiędzy $\mu \pm 2\sigma$ a $\mu \pm 3\sigma$, wynik to **0.5** (umiarkowane odstępstwo).
- Jeśli wartość $X1$ jest większa niż $\mu \pm 3\sigma$, wynik to **1** (silne odstępstwo).

Następnie, wyniki z poszczególnych zmiennych zostały zsumowane w nowej kolumnie. W kolejnym węźle oceniłem, czy obserwacja jest wartością odstającą, wykorzystując następującą funkcję:

```
if($wynik_outlier$ < 0.5, 0, 1)
```

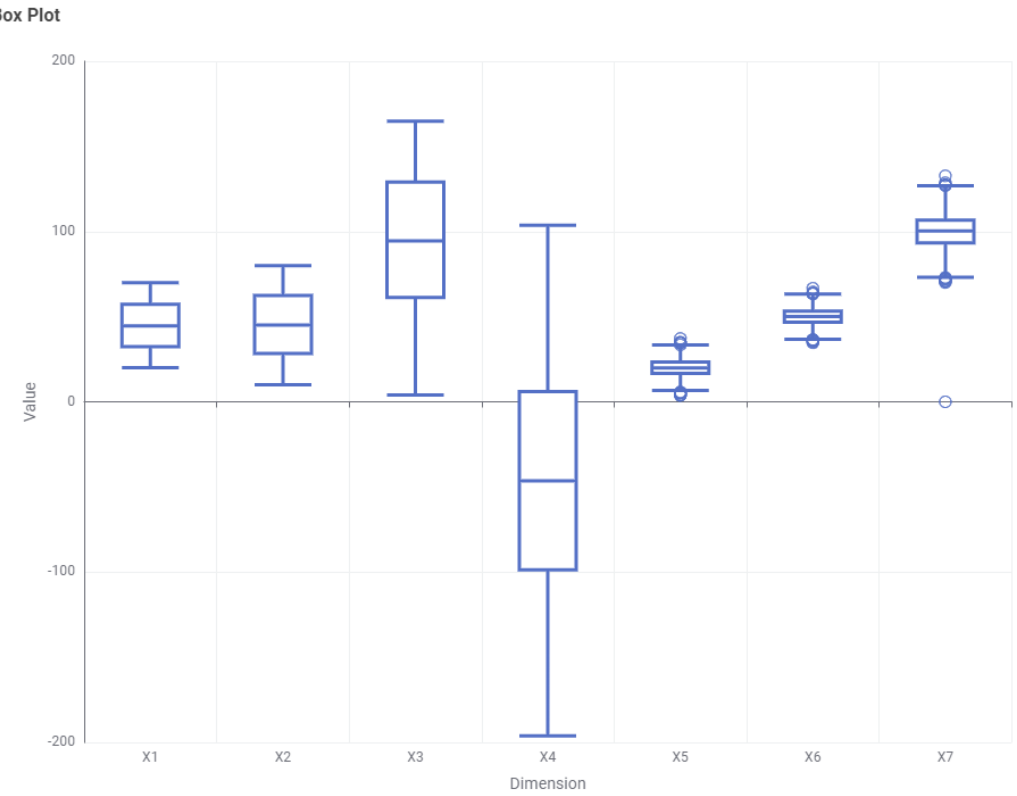
Ta funkcja zwraca:

- **1**, jeśli obserwacja została oznaczona jako odstająca (wartość skumulowana przekracza próg 0.5),
- **0**, jeśli obserwacja mieści się w akceptowalnym zakresie.

Wyniki analizy zapisano w **Arkuszu3** pliku *Projekt2* w programie Excel. Następnie z danych usunięto wszystkie obserwacje oznaczone jako odstające, co pozwoliło na dalsze przetwarzanie oczyszczonego zbioru danych.

Dane po usunięciu wartości odstających

Po wyeliminowaniu wartości odstających przygotowałem ponownie wykres pudełkowy, aby sprawdzić, jak zmienił się rozkład wartości.



Rysunek 4: Wykres pudełkowy wartości zmiennych po usunięciu wartości odstających

Jak wynika z powyższego wykresu, usunięcie wartości odstających pozwoliło na uzyskanie bardziej jednorodnego rozkładu danych, co znacząco ułatwi dalsze analizy.

1 Wyznaczenie i interpretacja podstawowych statystyk dla każdej zmiennej

1.1 Obliczanie podstawowych statystyk

Do obliczenia podstawowych statystyk wykorzystałem program **Excel**. Proces obliczeń został przedstawiony poniżej:

1. Podstawowe staystystki							
	X1	X2	X3	X4	X5	X6	X7
średnia	44,6670225	45,21860641	94,77127724	-46,17992955	19,99714361	50,03436359	100,0232791
Odchylenie sdandardowe	14,49737732	20,04230179	40,22590222	66,59139823	5,014493611	4,971653322	10,19794303
Mediana	44,58546085	45,13016708	94,55805087	-46,46908243	19,91410479	50,13322932	100,2809244
Q1	32,3555077	28,46998837	61,35270917	-98,69747493	16,6978424	46,66693324	93,28867542
Q3	57,2221597	62,38536992	129,0806077	5,974410988	23,38014646	53,33566288	106,7185177
Q0 (min)	20,03558053	10,02088298	4,061506522	-196,2189705	3,797429216	34,68974427	-0,00603431
Q4 (max)	69,95633723	79,98009888	164,7678241	103,6960101	37,18594868	66,67775199	132,698314
max - min	49,9207567	69,95921591	160,7063176	299,9149806	33,38851946	31,98800772	132,7043483
liczba pustych	1	3	2	1	2	1	3

Rysunek 5: Obliczenia statystyk opisowych w programie Excel

Wykorzystałem następujące funkcje programu Excel do obliczenia kluczowych miar statystycznych:

- **Średnia**: `=ŚREDNIA(A2:A1991)`
- **Odchylenie standardowe**: `=ODCH.STAND.POPUL(A2:A1991)`
- **Mediana**: `=MEDIANA(A2:A1991)`
- **Pierwszy kwartyl (Q1, rząd 0.25)**: `=KWARTYL(A2:A1991; 1)`
- **Trzeci kwartyl (Q3, rząd 0.75)**: `=KWARTYL(A2:A1991; 3)`
- **Wartość minimalna (Q0)**: `=MIN(A2:A1991)`
- **Wartość maksymalna (Q4)**: `=MAX(A2:A1991)`
- **Rozstęp danych (max-min)**: `=L9-L8`
- **Liczba pustych komórek**: `=LICZ.PUSTE(A2:A1991)`

1.2 Interpretacja statystyk opisowych dla każdej zmiennej

Poniżej przedstawiono interpretację wybranych statystyk opisowych dla danych:

- **Średnia**: To wartość przeciętna zmiennej, sugerująca, że większość danych oscyluje wokół tej wartości.
- **Odchylenie standardowe**: Wskazuje na rozrzut danych wokół średniej. Im większe odchylenie standardowe, tym większe zróżnicowanie danych. W tym przypadku dane charakteryzują się umiarkowanym zróżnicowaniem.
- **Mediana**: Jest to środkowa wartość zbioru danych, gdy dane są uporządkowane rosnąco. Zbliżenie mediany do średniej sugeruje, że rozkład danych jest symetryczny.

- **Pierwszy kwartyl (Q1)**: Wartość, poniżej której znajduje się 25% najmniejszych obserwacji. Oznacza dolną granicę danych, reprezentującą ”niższe” wartości w rozkładzie.
- **Trzeci kwartyl (Q3)**: Wartość, poniżej której znajduje się 75% obserwacji, a powyżej pozostaje 25%. Reprezentuje ”wyższe” wartości w rozkładzie danych.
- **Rozstęp (max-min)**: Różnica między wartością maksymalną i minimalną, wskazująca na pełen zakres zmienności danych.

1.3 Ogólne wnioski

Analiza statystyk opisowych wskazuje, że:

- Rozkład danych wydaje się symetryczny, ponieważ średnia i mediana są do siebie zbliżone.
- Dane charakteryzują się umiarkowanym zróżnicowaniem, co potwierdzają zarówno odchylenie standardowe, jak i rozstęp danych.
- Odchylenie standardowe i zakres międzykwartylowy sugerują, że dane są stabilne, bez dużych anomalii lub znacznych odstępstw.

2 Utworzenie dla zmiennej X_1 szeregu rozdzielczego oraz wyznaczenie podstawowych statystyk

2.1 Wyznaczanie szeregu rozdzielczego

Do wyznaczenia szeregu rozdzielczego dla zmiennej X_1 użyto programu Excel. Na początku obliczono szerokość przedziału według wzoru:

$$\text{szerokość przedziału} = \frac{\max - \min}{10}.$$

Następnie utworzono tabelę przedstawioną na poniższym rysunku:

2. Szereg rozdzielczy 10 elementowy Dla X1						
szerokość:	4,99207567					
	dolne	górne	ni	Xi	ni*xi	ni*(xi)^2
granice przedziałów:	20,03558053	25,0276562	209	22,53161837	4709,108239	106103,8297
	25,0276562	30,01973187	210	27,52369404	5779,975748	159086,284
	30,01973187	35,01180754	193	32,51576971	6275,543553	204054,129
	35,01180754	40,00388321	221	37,50784538	8289,233828	310911,3007
	40,00388321	44,99595888	175	42,49992105	7437,486183	316092,5756
	44,99595888	49,98803455	207	47,49199672	9830,84332	466886,3787
	49,98803455	54,98011022	178	52,48407239	9342,164885	490314,8581
	54,98011022	59,97218589	208	57,47614806	11955,0388	687129,5799
	59,97218589	64,96426156	187	62,46822373	11681,55784	729726,1684
	64,96426156	69,95633723	201	67,4602994	13559,52018	914729,291
suma:			1989		88860,47257	4385034,395

Rysunek 6: Wyznaczenie szeregu rozdzielczego 10-elementowego dla zmiennej X_1

Granice przedziałów wyznaczano, dodając szerokość przedziału do poprzedniej granicy, zaczynając od wartości minimalnej zmiennej.

Częstotliwości występowania elementów w każdym przedziale (n_i) obliczono za pomocą formuły:

$$\texttt{=LICZ.WARUNKI(}A1\texttt{:}A1991\texttt{;">="\&L17;}\ A1\texttt{:}A1991\texttt{;"<="\&M17)}.$$

Do tabeli dodano również kolumnę X_i, która reprezentuje środek każdego przedziału, oraz kolumny $n_i \cdot X_i$ i $n_i \cdot (X_i)^2$, wykorzystywane później do obliczeń podstawowych statystyk. Na końcu tabeli dodano wiersz podsumowujący, zawierający sumy poszczególnych kolumn.

2.2 Wyznaczanie i interpretacja podstawowych statystyk szeregu rozdzielczego

Podstawowe statystyki, takie jak średnia, wariancja i odchylenie standardowe, obliczono na podstawie danych z tabeli.

- **Średnia** obliczona według wzoru:
$$\bar{X} = \frac{\sum(n_i \cdot X_i)}{n}.$$
- **Wariancja** obliczona jako:
$$s^2 = \frac{\sum(n_i \cdot (X_i)^2)}{n} - \bar{X}^2.$$

- **Odchylenie standardowe** wyznaczono jako pierwiastek z wariancji:

$$s = \sqrt{s^2}.$$

Wyniki zaprezentowano w poniższej tabeli:

średnia	44,67595
wariancja	208,7019
Odchylenie standardowe	14,44652

Rysunek 7: Wyliczanie podstawowych statystyk szeregu rozdzielczego 10-elementowego dla zmiennej X_1

2.3 Interpretacja wyników

Różnica w średnich pomiędzy szeregiem rozdzielczym a danymi oryginalnymi wynosi około $0,009$. Oznacza to, że szereg rozdzielczy bardzo dobrze odwzorowuje tendencję centralną danych, a agregacja nie wpływa znacząco na wartość średniej.

Wariancja szeregu rozdzielczego jest nieznacznie mniejsza od wariancji danych oryginalnych ($\approx 1{,}472$). Wynika to z efektu wygładzania, które pojawia się w wyniku grupowania danych w przedziały.

Odchylenie standardowe również jest minimalnie mniejsze dla szeregu. To potwierdza, że grupowanie danych delikatnie zmniejsza rozrzut danych, choć różnica jest niewielka.

2.4 Wnioski

Szereg rozdzielczy dobrze reprezentuje dane bez istotnej utraty informacji. Średnia, wariancja i odchylenie standardowe dla obu zestawów są zbliżone, co potwierdza, że agregacja danych w szeregach rozdzielczych nie prowadzi do znaczących zniekształceń wyników. Szeregi rozdzielcze są więc użyteczne przy analizie danych, szczególnie w sytuacjach wymagających uproszczenia i uogólnienia przy zachowaniu kluczowych charakterystyk statystycznych.

3 Uzupełnianie braków danych

Na początku zbudowałem macierz korelacji między zmiennymi, aby sprawdzić, czy są jakieś zmienne które z siebie wynikają. Wykorzystałem narzędzie Excela analiza danych do zbudowania tej macierzy.

	X1	X2	X3	X4	X5	X6	X7
X1	1,00000						
X2	0,01146	1,00000					
X3	0,01273	0,99998	1,00000				
X4	0,42356	-0,89930	-0,89983	1,00000			
X5	-0,01692	-0,01421	-0,01625	0,00785	1,00000		
X6	0,00207	0,01509	0,01260	-0,00982	0,00759	1,00000	
X7	0,00645	0,01004	0,00791	-0,00366	0,00677	0,99834	1,00000

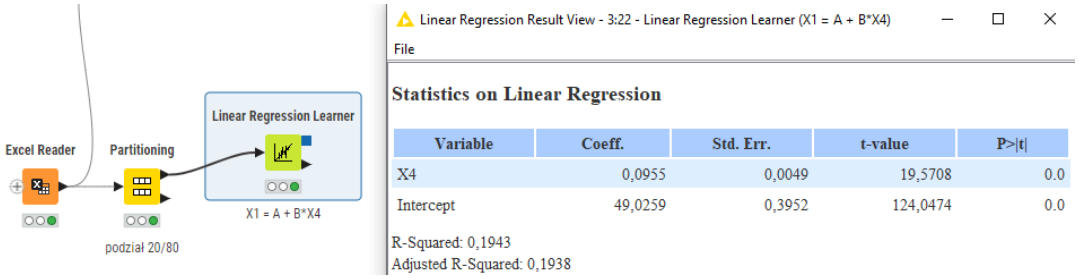
Rysunek 8: Macierz korelacji zmiennych

Na podstawie tej korelacji zbuduję modele regresji liniowej aby uzupełnić braki danych.

3.1 Propozycje modeli

3.1.1 Model dla zmiennej X_1

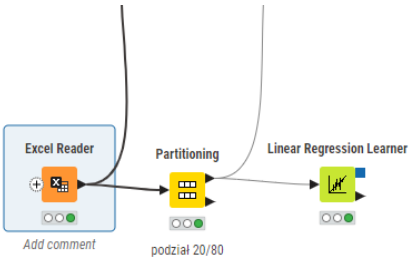
Analiza korelacji wykazała, że X_1 ma niskie korelacje z pozostałymi zmiennymi. Najwyższa korelacja, wynosząca 0,42356, została odnotowana w odniesieniu do X_4. W związku z tym w pierwszym kroku zbudowano model regresji liniowej z X_4 jako jedynym predyktorem w celu oceny jego znaczenia.



Rysunek 9: Model regresji liniowej ze zmienną najbardziej skorelowaną

Dane zostały podzielone na zbiór uczący (80%) oraz testowy (20%). Model regresji liniowej, zbudowany z wykorzystaniem węzła *Linear Regression Learner* w programie KNIME, poprawnie opisał jedynie 19,4% przypadków, co potwierdza wcześniejsze przypuszczenia o niskiej jakości modelu.

Następnie zbudowano model z użyciem wszystkich zmiennych jako predyktorów w celu analizy ich istotności statystycznej.



Linear Regression Result View - 3:23 - Linear Regression Learner

File

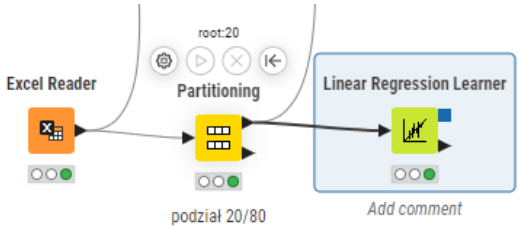
Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
X2	1,5	2,12E-14	7,09E13	0.0
X3	6,98E-15	1,06E-14	0,6597	0,5095
X4	0,5	1,02E-16	4,88E15	0.0
X5	4,70E-16	5,92E-16	0,794	0,4273
X6	1,37E-14	1,04E-14	1,3228	0,1861
X7	-6,36E-15	5,18E-15	-1,2274	0,2198
Intercept	-2,26E-13	5,82E-14	-3,8823	0,0001

R-Squared: 1
Adjusted R-Squared: 1

Rysunek 10: Model z wszystkimi zmiennymi

Model uwzględniający wszystkie zmienne poprawnie opisał 100% przypadków, jednakże okazał się zbyt skomplikowany. Z modelu usunięto zmienne o współczynnikach bliskich zeru oraz niskiej istotności statystycznej, uzyskując uproszczoną wersję.



Linear Regression Result View - 3:23 - Linear Regression Learner

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
X2	1,5	3,33E-16	4,50E15	0.0
X4	0,5	9,96E-17	5,02E15	0.0
Intercept	-7,46E-14	1,15E-14	-6,5136	9,83E-11

R-Squared: 1
Adjusted R-Squared: 1

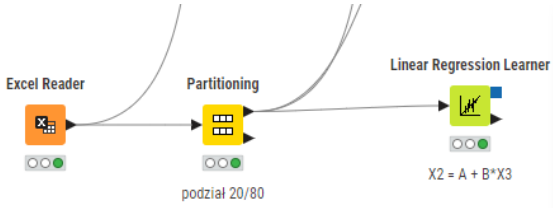
Rysunek 11: Uproszczony model dla zmiennej X_1

Ostateczny model poprawnie opisuje 100% przypadków i jest mniej złożony. Wartość wyrazu wolnego, bliska zeru, została pominięta. Model zastosowany do uzupełnienia braków danych w zmiennej X_1 jest następujący:

$$X_1 = 1.5 \cdot X_2 + 0.5 \cdot X_4$$

3.1.2 Model dla zmiennych X_2 i X_3

Zmienne X_2 i X_3 są niemal liniowo skorelowane (korelacja 0,99998). Aby uniknąć problemów z wielokolinearnością, wybrano tylko jedną z nich jako zmienną zależną. Ostateczny model jest przedstawiony poniżej.



Linear Regression Result View - 3:24 - Linear Regression Learner (X2 = A + B*X3)

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
X3	0,4999	8,85E-5	5 647,9224	0.0
Intercept	-2,2351	0,0091	-246,0826	0.0

R-Squared: 1
Adjusted R-Squared: 1

Rysunek 12: Model dla zmiennych X_2 i X_3

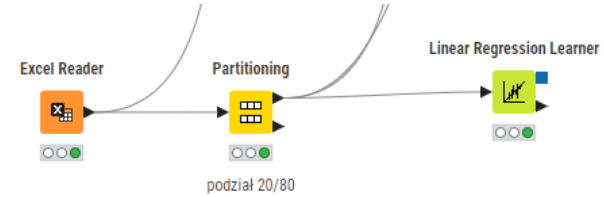
Model poprawnie opisuje 100% przypadków, a jego struktura jest mało obciążona. Modele wykorzystane do uzupełnienia braków danych są następujące:

$$X_2 = 0.5 \cdot X_3 - 2.24$$

$$X_3 = 2 \cdot X_2 + 4.48$$

3.1.3 Model dla zmiennej X_4

Zmienne X_1 i X_4 wykazują umiarkowaną dodatnią korelację (0,42356), natomiast X_4 jest silnie skorelowana w sposób negatywny z X_2 i X_3. Aby uniknąć wielokolinearności, w modelu uwzględniono X_1 i jedną ze zmiennych X_2 lub X_3.



Linear Regression Result View - 3:24 - Linear Regression Learner

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
X1	2	4,20E-16	4,77E15	0.0
X2	-3	3,05E-16	-9,84E15	0.0
Intercept	-1,14E-13	2,41E-14	-4,7119	2,67E-6

R-Squared: 1
Adjusted R-Squared: 1

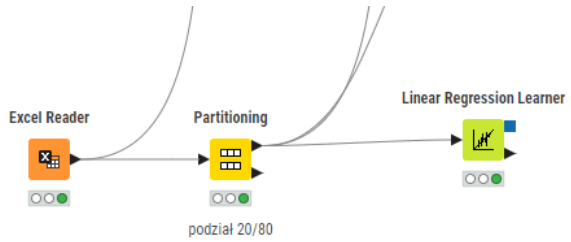
Rysunek 13: Model dla zmiennej X_4

Model poprawnie opisuje 100% przypadków i charakteryzuje się niskim obciążeniem. Wyraz wolny, z wartością bliską zeru, został pominięty. Ostateczny model wykorzystany do uzupełnienia braków danych w zmiennej X_4 to:

$$X_4 = 2 \cdot X_1 - 3 \cdot X_2$$

3.1.4 Model dla zmiennej X_5

Zmienne X_5 wykazują bardzo niskie korelacje z innymi zmiennymi, co utrudnia budowę sensownego modelu.



Linear Regression Result View - 3:24 - Linear Regression Learner

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
X1	5,35E11	1,04E12	0,5137	0,6075
X2	-8,02E11	1,56E12	-0,5137	0,6075
X3	-0,2006	0,4502	-0,4456	0,6559
X4	-2,67E11	5,21E11	-0,5137	0,6075
X6	-0,041	0,4416	-0,0928	0,9261
X7	0,031	0,2207	0,1405	0,8883
Intercept	20,1612	2,4254	8,3124	2,22E-16

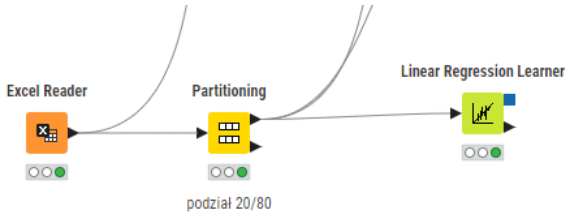
R-Squared: 0,001
Adjusted R-Squared: -0,0028

Rysunek 14: Model dla zmiennej X_5

Model opisuje niewielką liczbę przypadków, a współczynniki są istotne statystycznie na niskim poziomie. W tym przypadku braków danych nie uzupełniano, lecz zostały one usunięte.

3.1.5 Model dla zmiennych X_6 i X_7

Zmienne X_6 i X_7 są silnie skorelowane (korelacja 0,99834). Aby uniknąć wielokolinearności, jedna z nich została wybrana jako zmienna zależna.



Linear Regression Result View - 3:24 - Linear Regression Learner

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
X7	0,499	0,0007	685,9533	0.0
Intercept	0,1027	0,073	1,4063	0,1598

R-Squared: 0,9966
Adjusted R-Squared: 0,9966

Rysunek 15: Model dla zmiennych X_6 i X_7

Model poprawnie opisuje prawie 100% przypadków i jest mało obciążony. Wyraz wolny, ze względu na niską istotność statystyczną, został pominięty. Modele zastosowane do uzupełnienia braków danych to:

$$X_6 = 0.5 \cdot X_7$$

$$X_7 = 2 \cdot X_6$$

3.2 Podsumowanie uzupełnienia braków danych

Przedstawione modele wykorzystam do uzupełnienia braków danych w programie Excel. Policzę także ponownie podstawowe statystyki w celu porównania skuteczności uzupełnienia braków danych.

1. Podstawowe staystystki po uzupełnieniu braków danych							
	X1	X2	X3	X4	X5	X6	X7
średnia	44,65	45,15	94,80	-46,17	20,00	50,00	100,02
Odchylenie sdandardowe	14,53	20,09	40,20	66,54	5,01	5,09	10,19
Mediana	44,58	45,06	94,56	-46,35	19,91	50,13	100,28
Q1	32,32	28,42	61,41	-98,67	16,70	46,66	93,27
Q3	57,22	62,34	129,08	5,86	23,38	53,33	106,73
Q0 (min)	0,00	-0,21	4,06	-196,22	3,80	0,00	-0,01
Q4 (max)	69,96	79,98	164,77	103,70	37,19	66,68	132,70
max - min	69,96	80,19	160,71	299,91	33,39	66,68	132,70
liczba pustych	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1. Podstawowe staystystki przed uzupełnieniem braków danych							
	X1	X2	X3	X4	X5	X6	X7
średnia	44,67	45,22	94,77	-46,18	20,00	50,03	100,02
Odchylenie sdandardowe	14,50	20,04	40,23	66,59	5,01	4,97	10,20
Mediana	44,59	45,13	94,56	-46,47	19,91	50,13	100,28
Q1	32,36	28,47	61,35	-98,70	16,70	46,67	93,29
Q3	57,22	62,39	129,08	5,97	23,38	53,34	106,72
Q0 (min)	20,04	10,02	4,06	-196,22	3,80	34,69	-0,01
Q4 (max)	69,96	79,98	164,77	103,70	37,19	66,68	132,70
max - min	49,92	69,96	160,71	299,91	33,39	31,99	132,70
liczba pustych	1,00	3,00	2,00	1,00	2,00	1,00	3,00

Rysunek 16: Porównanie statystyk przed i po uzupełnieniu braków danych

Na podstawie przedstawionych statystyk można zauważyć różnice między danymi przed i po uzupełnieniu braków. Oceniając, czy uzupełnienie braków danych zostało przeprowadzone poprawnie, warto wziąć pod uwagę następujące aspekty:

Średnie i odchylenie standardowe: Po uzupełnieniu średnie dla większości zmiennych są bardzo zbliżone do tych sprzed uzupełnienia, co sugeruje, że metoda uzupełnienia zachowała ogólną strukturę danych. Odchylenie standardowe również wykazuje minimalne różnice, co wskazuje na niewielkie zmiany w rozkładzie.

Mediana i kwartyle (Q1, Q3): Wartości mediany oraz kwartyle przed i po uzupełnieniu danych różnią się bardzo nieznacznie. Oznacza to, że uzupełnienie braków danych nie zaburzyło znacząco rozkładu.

Zakres wartości (max - min): Po uzupełnieniu dane zachowują te same wartości maksymalne i minimalne

w większości przypadków, co sugeruje, że brakujące dane zostały uzupełnione w sposób, który nie wpływa na istniejący zakres.

Wniosek: Uzupełnienie braków danych wydaje się być przeprowadzone poprawnie, ponieważ zmiany w podstawowych statystykach są minimalne, a struktura danych została zachowana.

Podsumowanie

W ramach niniejszego projektu przeprowadzono kompleksową analizę zbioru danych, obejmującą kilka istotnych etapów, takich jak obliczenie podstawowych statystyk opisowych, budowę szeregu rozdzielczego, uzupełnianie braków danych oraz dokumentację procesu analitycznego. Projekt wykorzystał zarówno narzędzia Microsoft Excel, jak i KNIME Analytics Platform, co pozwoliło na zastosowanie klasycznych oraz nowoczesnych metod analizy danych.

Wykonane czynności:

- **Podstawowe statystyki opisowe:** Wyznaczono kluczowe miary statystyczne dla każdej zmiennej w zbiorze danych, takie jak średnia, mediana, odchylenie standardowe, kwartyle oraz wartości minimalne i maksymalne. Pozwoliło to na uzyskanie ogólnego obrazu rozkładu danych, wykrycie ewentualnych anomalii oraz przygotowanie danych do dalszej analizy.
- **Budowa szeregu rozdzielczego:** Zbudowano szereg rozdzielczy dla zmiennej X1, składający się z 10 przedziałów. Na podstawie tego szeregu wyznaczone zostały wartości miar statystycznych, takich jak liczność i frekwencja, co pozwoliło na szczegółową analizę rozkładu tej zmiennej.
- **Uzupełnianie braków danych:** Zastosowano różne techniki uzupełniania braków danych. W zależności od charakterystyki zmiennej, brakujące dane zostały uzupełnione na podstawie modeli regresji, przy użyciu innych zmiennych w zbiorze. W przypadkach, gdzie nie udało się uzyskać trafnych modeli, brakujące dane zostały usunięte, aby nie wprowadzać błędów w dalszą analizę.
- **Raport:** Na każdym etapie analizy wykonano szczegółowe dokumentowanie kroków, obliczeń oraz narzędzi wykorzystywanych w procesie. Dodatkowo, przedstawiono wizualizacje wyników, które pozwalają na łatwiejsze zrozumienie przeprowadzonych analiz i wyników.

Wnioski: Dzięki przeprowadzonej analizie możliwe było uzyskanie pełniejszych i bardziej dokładnych wyników, co ma kluczowe znaczenie dla dalszej pracy z danymi. Przeanalizowane dane zostały wstępnie oczyszczone, uzupełnione, a wyniki statystyczne pozwalają na dalsze badanie zjawisk i zależności w zbiorze. Projekt pokazał również zalety i ograniczenia stosowania różnych narzędzi analitycznych, które mogą wspierać procesy eksploracji danych w praktyce.