

Eksploracja zbioru danych

Niniejszą analizę eksploracyjną zbioru danych wykonano w Systemie Zautomatyzowanego Tworzenia Opisu Statystycznego - SZTOS (Hryniewicz, Milewska, 2023) w module eksploracja.

W pierwszej części raportu została umieszczona informacja o liczbie obserwacji oraz liczbie zmiennych we wprowadzonej bazie danych. Aby usunąć z bazy danych a) obserwacje które nie zawierają żadnych danych, b) zmienne których liczba obserwacji jest mniejsza lub równa 1, zidentyfikowano takie przypadki. Jeśli takowe wystąpiły, zostały usunięte wraz z podaniem informacji o ich współrzędnych. Tej samej procedurze zostały poddane zmienne, które posiadały identyczne wartości.

Liczba obserwacji: 303

Liczba zmiennych: 8

- W bazie danych **nie występują** OBSERWACJE, które zawierają wyłącznie braki danych. Obserwacje, które zawierały wyłącznie braki danych to: brak.
- W bazie danych **nie wystąpiły** ZMIENNE, które zawierają wyłącznie braki danych lub 1 obserwację. Zmienne, które zawierały wyłącznie braki danych lub 1 obserwację to: brak.
- W bazie danych **nie wystąpiły** zmienne, które miały zerową zmienność. Zmienne, które posiadały identyczne wartości to: brak.

Stan braków danych w analizowanej bazie danych

W tak przygotowanej bazie danych zidentyfikowano zmienne, które posiadają odsetek braków danych większy niż ten wyznaczony przez użytkownika, a informację podano do raportu.

W bazie danych **nie wystąpiły** ZMIENNE, które zawierają ponad lub równo: 5% braków danych.

Zmienne, które zawierają równo lub ponad 5% braków danych, to: brak.

Dokładne wyniki przedstawia Tabela nr 2.

Tabela nr 1

Następnie zidentyfikowano obserwacje, które zawierają ponad połowę braków danych i podano ich współrzędne. Taki stan rzeczy może wynikać np. z porzucenia badania w trakcie jego trwania. Może to powodować ograniczoną przydatność takich obserwacji w analizach

Zmienna	% braków danych	n/N braków danych
Płeć	0.00%	0/303
Edukacja	0.00%	0/303
Grupa badana	0.00%	0/303
Obawy	0.00%	0/303
Cele	0.00%	0/303
Łatwość	0.00%	0/303
Użyteczność	0.00%	0/303
Intencja zakupu	0.00%	0/303

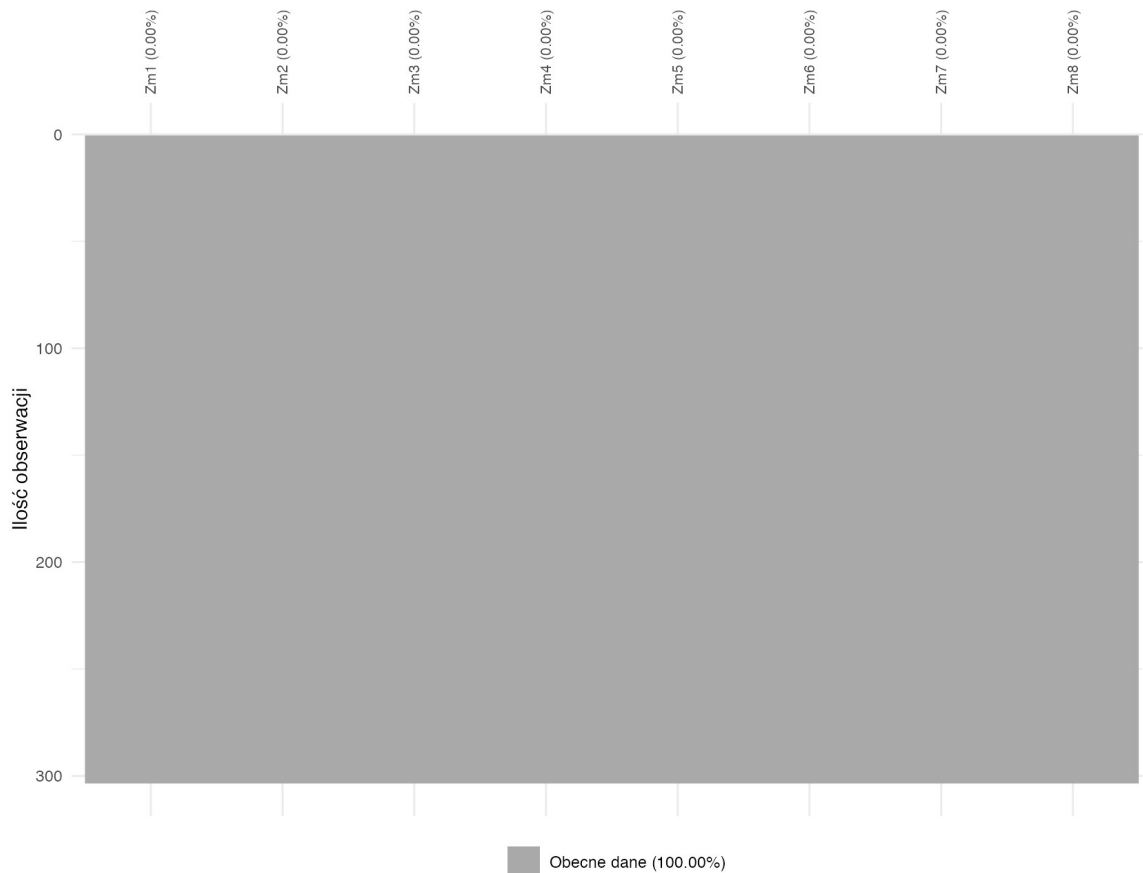
W bazie danych **nie wystąpiły** OBSERWACJE, które zawierają ponad połowę braków danych.

Numery wierszy, które zawierają ponad połowę braków danych: brak.

Wizualną prezentację braków danych stanowi Rysunek nr 1.

Wykres nr 1

Graficzna prezentacja braków danych występujących w zakresie zmiennych; Płeć, Edukacja, Grupa badana, Obawy, Cele, Łatwość, Użyteczność, Intencja zakupu .



Nota. Rysunek nr 1 Graficzna prezentacja braków danych występujących w zakresie zmiennych; Płeć, Edukacja, Grupa badana, Obawy, Cele, Łatwość, Użyteczność, Intencja zakupu . Nota : Wykres przedstawia następujące zmienne: Zm1 - Płeć, Zm2 - Edukacja, Zm3 - Grupa badana, Zm4 - Obawy, Zm5 - Cele, Zm6 - Łatwość, Zm7 - Użyteczność, Zm8 - Intencja zakupu.

Rozkład zmiennych jakościowych bez uwzględniania braków danych

Analiza dla zmiennej Płeć wykazała, że liczebność jej poszczególnych poziomów Kobieta, Mężczyzna wynosiła odpowiednio 173, 130 (ich odsetek wynosił odpowiednio 57.10, 42.90), co stanowiło sumę 303 wystąpień (100%).

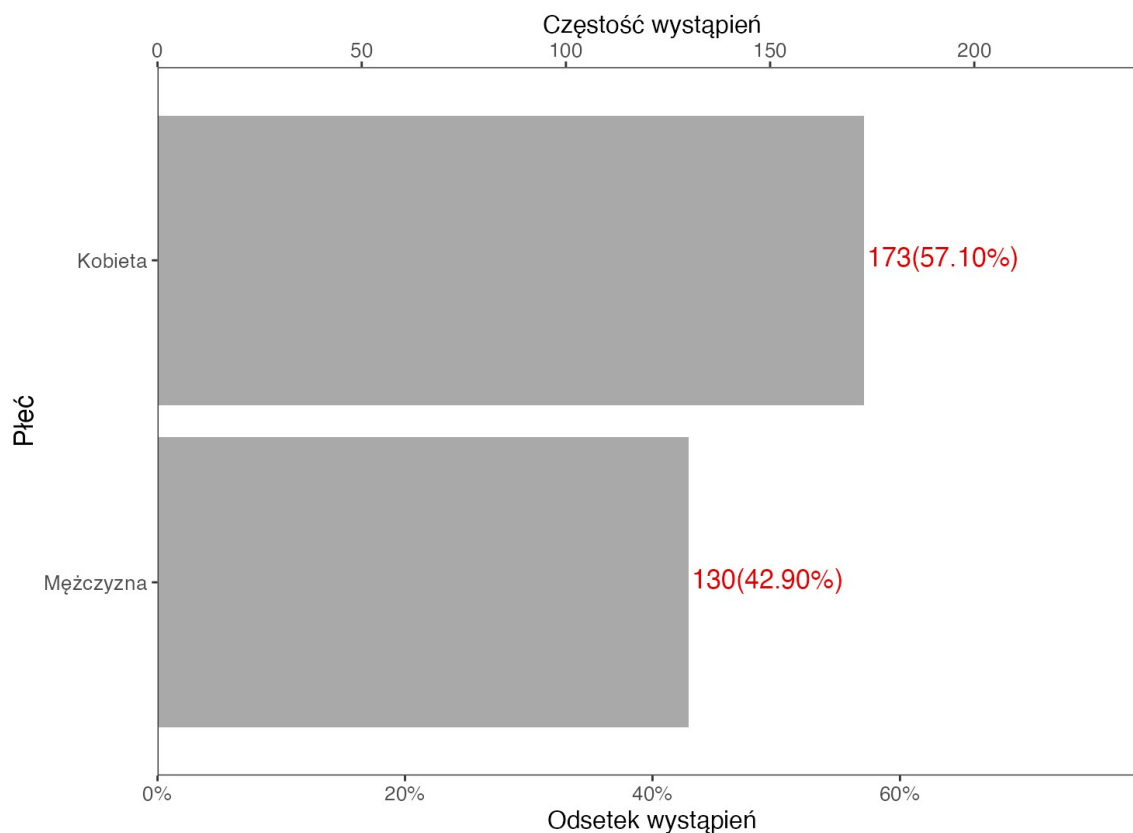
Analiza dla zmiennej Edukacja wykazała, że liczebność jej poszczególnych poziomów Wyższe, Średnie, Podstawowe wynosiła odpowiednio 171, 130, 2 (ich odsetek wynosił odpowiednio 56.44, 42.90, 0.66), co stanowiło sumę 303 wystąpień (100%).

Analiza dla zmiennej Grupa badana wykazała, że liczebność jej poszczególnych poziomów Bandziorki, Papugi, Policjanci wynosiła odpowiednio 101, 101, 101 (ich odsetek wynosił odpowiednio 33.33, 33.33, 33.33), co stanowiło sumę 303 wystąpień (100%).

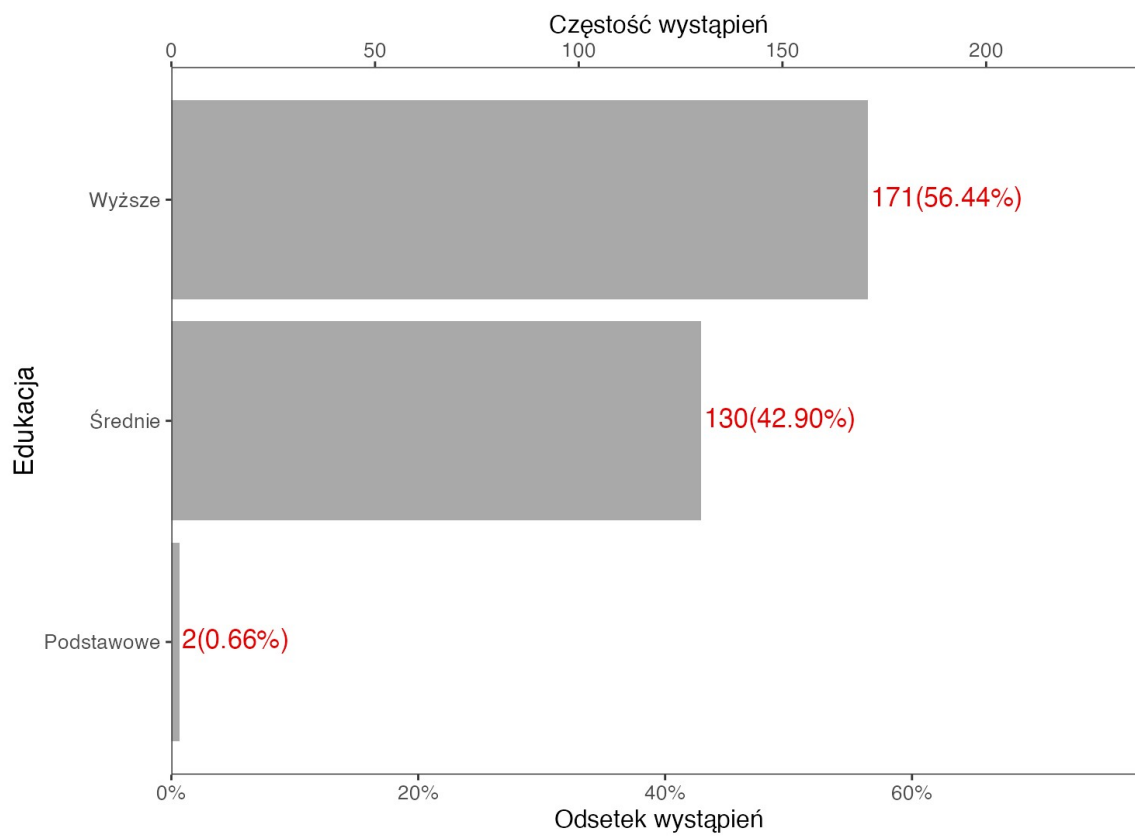
Otrzymane wyniki przedstawiają Rysunki od nr 2 do nr 11.

Wykres nr 2

Częstość występowania wartości zmiennej Płeć (bez uwzględniania braków danych)

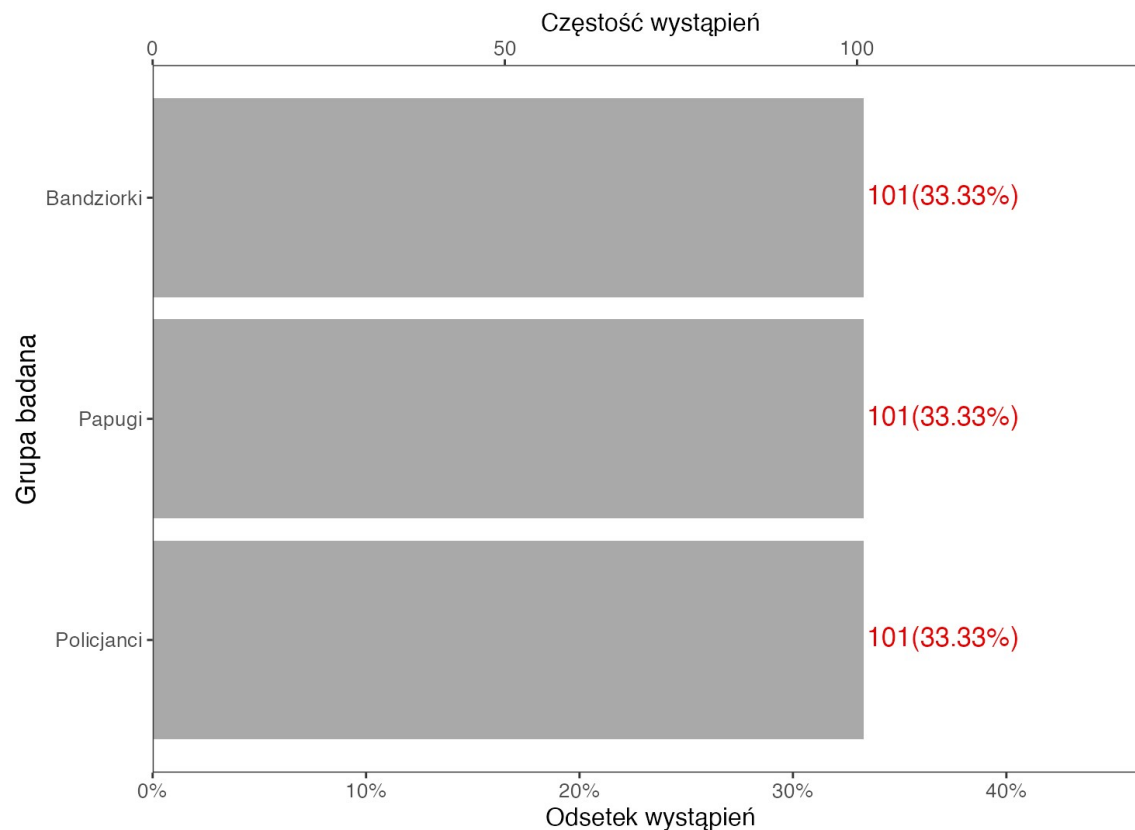


Wykres nr 3



Nota. n (%); n = liczebność w podzbiorze. Rysunek nr 2 Częstość występowania wartości zmiennej Edukacja (bez uwzględniania braków danych)

Wykres nr 4



Nota. n (%); n = liczebność w podzbiorze. Rysunek nr 3 Częstość występowania wartości zmiennej Grupa badana (bez uwzględniania braków danych)

Nota. n (%); n = liczebność w podzbiorze.

[1] “”

Statystyki opisowe i testy normalności analizowanych zmiennych

Analiza normalności rozkładu zmiennych: Obawy, Cele, Łatwość, Użyteczność, Intencja zakupu.

W celu weryfikacji założeń dotyczących normalności rozkładów wyników analizowanych zmiennych przeprowadzono dwa testy statystyczne, był to test Shapiro - Wilka (stosowany dla małoliczebnych próbek badawczych $N < 50$ lub $N < 100$) (Royston, 1982) oraz test Kolmogorova Smirnova z poprawką Lilieforce'a (stosowany dla dużych prób $N > 50$ lub $N > 100$) (Dallal i Wilkinson, 1986). Zdecydowano się na taką analizę ze względu na brak

jednoznacznych wytycznych dotyczących stosowania obu testów przy danej liczebności badanych obserwacji.

W celu weryfikacji normalności rozkładów zmiennych: Obawy, Cele, Łatwość, Użyteczność, Intencja zakupu, przeprowadzono serię analiz weryfikujących podobieństwo rozkładu z próby do teoretycznego rozkładu normalnego testem Shapiro-Wilka (Royston, 1982). Analiza wykazała, że:

Rozkład wyników zmiennej Obawy był istotnie różny od rozkładu normalnego $SW = 0.97; p < 0.001$

Rozkład wyników zmiennej Cele był istotnie różny od rozkładu normalnego $SW = 0.96; p < 0.001$

Rozkład wyników zmiennej Łatwość był istotnie różny od rozkładu normalnego $SW = 0.95; p < 0.001$

Rozkład wyników zmiennej Użyteczność był istotnie różny od rozkładu normalnego $SW = 0.98; p < 0.001$

Rozkład wyników zmiennej Intencja zakupu był istotnie różny od rozkładu normalnego $SW = 0.97; p < 0.001$

W celu dodatkowej weryfikacji normalności rozkładów zmiennych: Obawy, Cele, Łatwość, Użyteczność, Intencja zakupu, przeprowadzono serię analiz weryfikujących podobieństwo rozkładu z próby do teoretycznego rozkładu normalnego testem Kolmogorova-Smirnova z poprawką Liliefors'a (Dallal i Wilkinson, 1986). Analiza wykazała, że:

Rozkład wyników zmiennej Obawy był istotnie różny od rozkładu normalnego $KS = 0.12; p < 0.001$

Rozkład wyników zmiennej Cele był istotnie różny od rozkładu normalnego $KS = 0.13; p < 0.001$

Rozkład wyników zmiennej Łatwość był istotnie różny od rozkładu normalnego $KS = 0.17; p < 0.001$

Rozkład wyników zmiennej Użyteczność był istotnie różny od rozkładu normalnego $KS = 0.10; p < 0.001$

Rozkład wyników zmiennej Intencja zakupu był istotnie różny od rozkładu normalnego $KS = 0.08$; $p < 0.001$

Wyniki analiz przedstawia Tabela nr 2.

Tabela nr 2

Wyniki analizy rozkładu normalności i statystyk opisowych dla zmiennych: Obawy, Cele, Łatwość, Użyteczność, Intencja zakupu

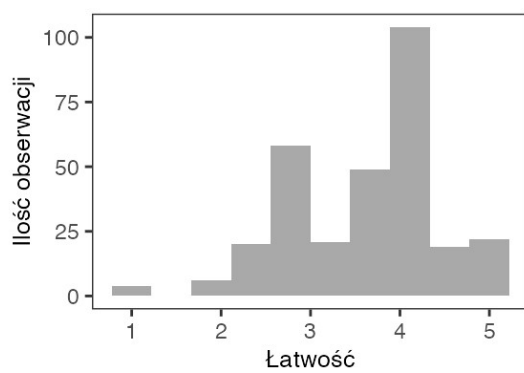
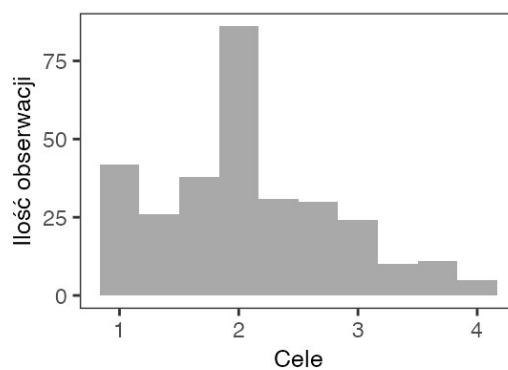
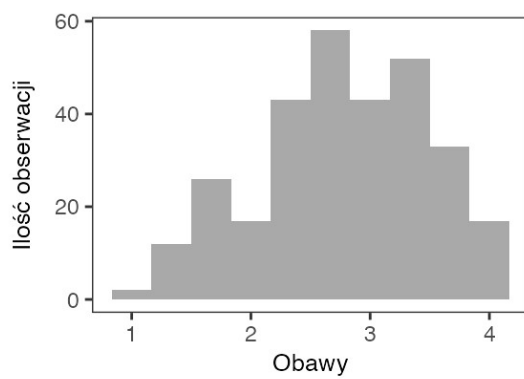
Zmienna	N	Min	Max	M	SD	SE	Me	Miary symetrii rozkładu		test KS		test SW	
								Skośność	Kurtoza	KS	p	SW	p
Obawy	303	1.00	4.00	2.79	0.71	0.04	2.80	-0.31	-0.57	0.12	< 0.001	0.97	< 0.001
Cele	303	1.00	4.00	2.15	0.74	0.04	2.00	0.45	-0.21	0.13	< 0.001	0.96	< 0.001
Łatwość	303	1.00	5.00	3.61	0.79	0.05	3.75	-0.52	0.51	0.17	< 0.001	0.95	< 0.001
Użyteczność	303	1.00	5.00	3.24	0.89	0.05	3.25	-0.32	-0.05	0.10	< 0.001	0.98	< 0.001
Intencja zakupu	303	1.00	5.00	2.76	1.03	0.06	2.75	0.03	-0.68	0.08	< 0.001	0.97	< 0.001

Nota. N = Liczebność; Min = Wartość minimalna; Max = Wartość maksymalna; M = Średnia arytmetyczna; SD = Odchylenie standardowe; SE = Błąd standardowy średniej; Me = Mediana; p = Istotność statystyczna. Wyniki testu KS (Kolmogorov - Smirnov) i SW (Shapiro - Wilk), których $p < 0.05$ dla danej zmiennej, wskazują na istotną różnicę między rozkładem wyników w próbie a teoretycznym rozkładem normalnym.

Histogramy (Wykresy skumulowane)

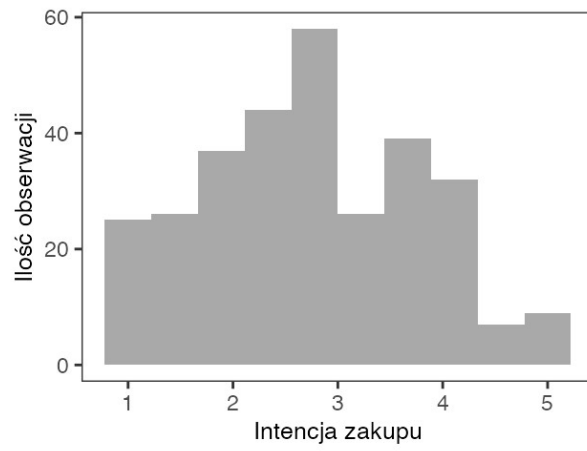
Wykres nr 5

Rozkład wartości zmiennych: Obawy, Cele, Łatwość, Użyteczność



Wykres nr 6

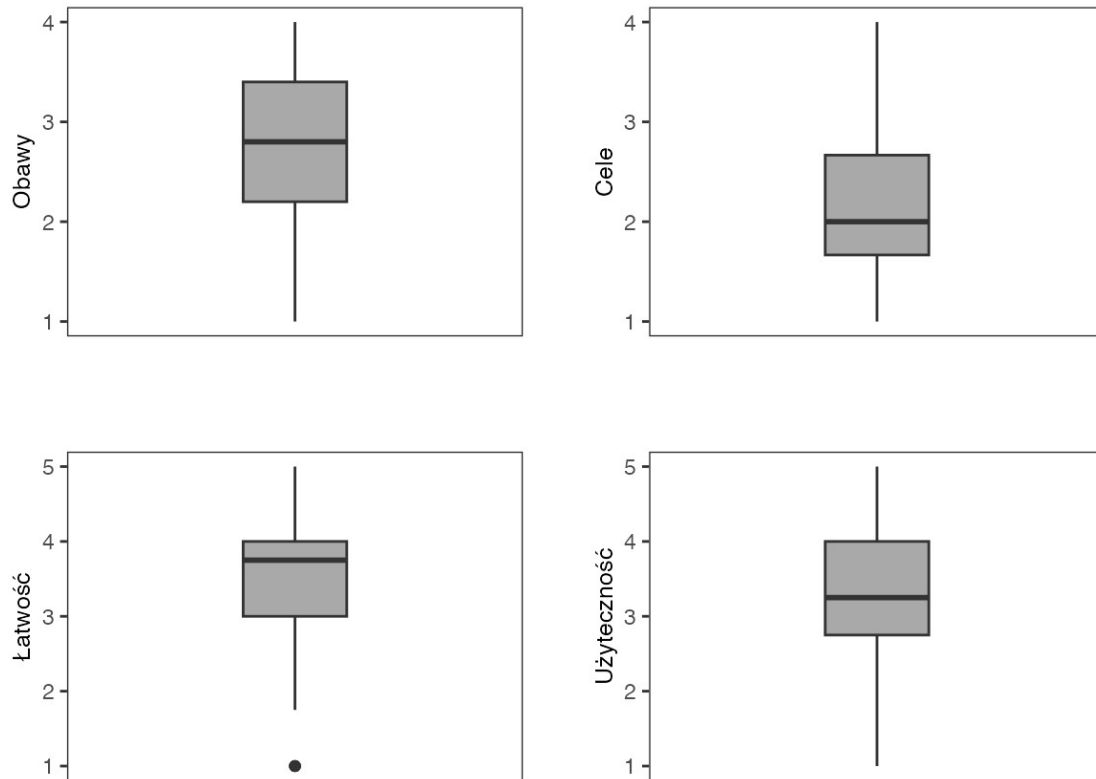
Rozkład wartości zmiennych: Intencja zakupu



Box-ploty (Wykresy skumulowane)

Wykres nr 7

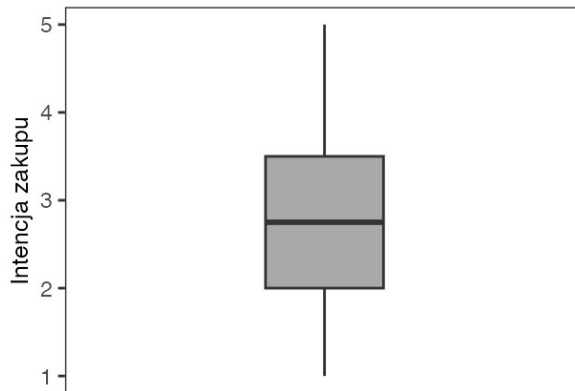
Rozkład wartości zmiennych: Obawy, Cele, Łatwość, Użyteczność



Nota. Pozioma linia przedstawia Medianę. Pudełko prezentuje Rozstęp międzykwartyłowy IQR. Wąsy sięgają do wartości powyżej/poniżej $1.5 \cdot \text{IQR}$. Kropki poza wąsami oznaczają wartości powyżej/poniżej $1.5 \cdot \text{IQR}$.

Wykres nr 8

Rozkład wartości zmiennych: Intencja zakupu



Nota. Pozioma linia przedstawia Medianę. Pudełko prezentuje Rozstęp międzykwartyłowy IQR. Wąsy sięgają do wartości powyżej/poniżej $1.5 \cdot \text{IQR}$. Kropki poza wąsami oznaczają wartości powyżej/poniżej $1.5 \cdot \text{IQR}$.

Opis przeprowadzonej normalizacji

W celu unormowania rozkładów w analizowanych danych, wykorzystano bibliotekę języka R o nazwie „bestNormalize” (Peterson, 2021; Peterson i inni, 2020) , która pozwala na testowanie wielu przekształceń unormalniających rozkład. Logika tej funkcji polega na testowaniu 15 możliwych przekształceń, wraz z jednoczesnym ustaleniem optymalnych parametrów przekształceniowych np. hiperparametru Lambda w przekształceniu Boxa Coxa (Box & Cox, 1964). Istnieje wiele przypadków, w których badacze mogą chcieć znormalizować zmienną. Oto jeden z najczęstszych:

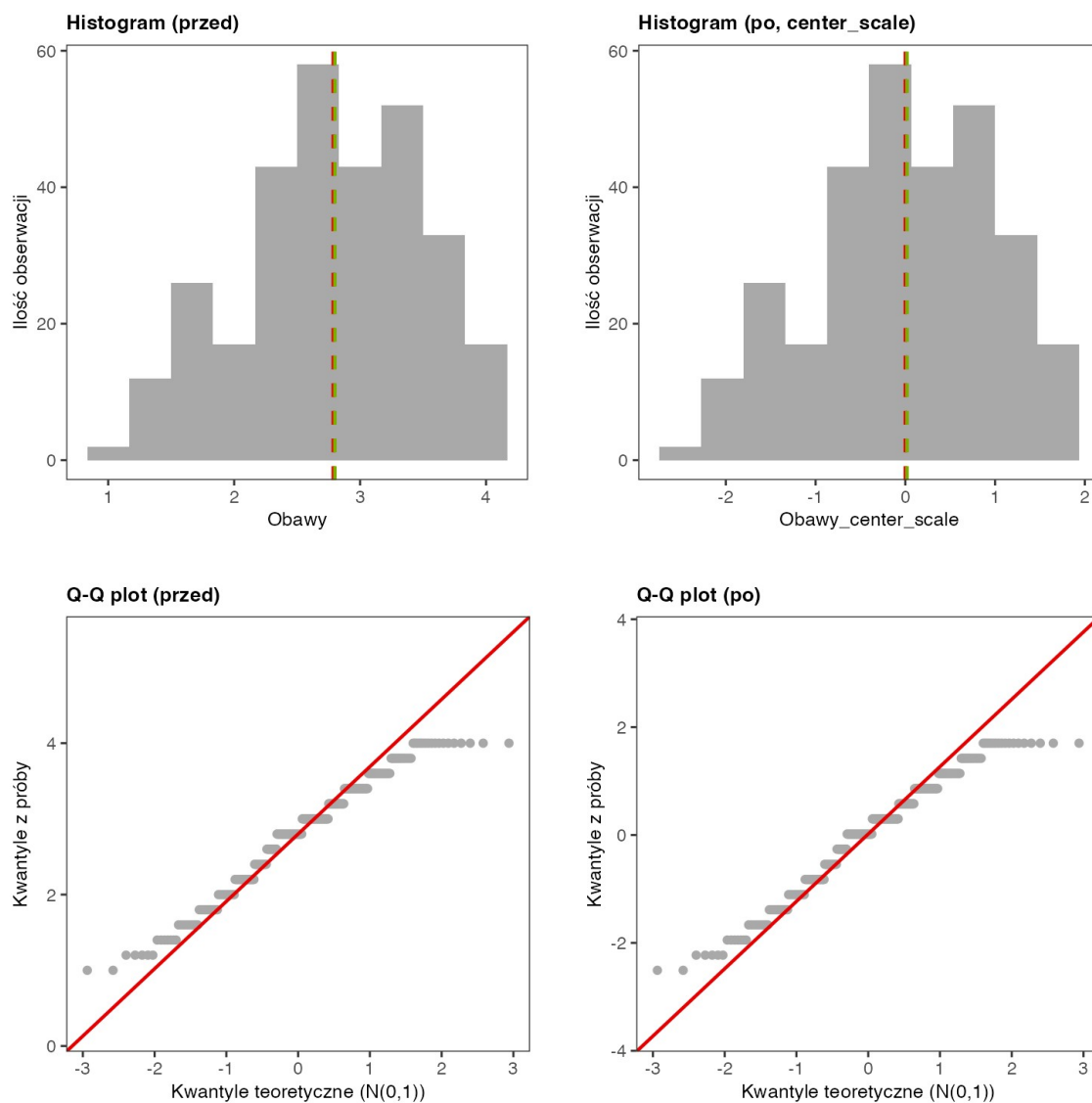
Często problematyczne założenie o normalności zmiennej zależnej (uwarunkowanego współzmiennymi w modelu) w klasycznym problemie regresji liniowej. Przez lata stosowano wiele metod, aby złagodzić to założenie: uogólnione modele liniowe, regresja kwantylowa, modele przetrwania itp. Jedną z technik, która wciąż jest dość popularna w tym kontekście, jest transformacja danych, aby wyglądały normalnie. Może to być coś tak prostego jak transformacja logarytmiczna tzw. $\log()$ lub coś tak złożonego jak transformacja Yeo-Johnsona.

W celu wyboru metody normalizującej zmienną Obawy wykonano serię przekształceń. Analiza wykazała, że najlepsze właściwości normalizacyjne zwracała technika `center_scale`. W porównaniu do innych popularnych przekształceń, metoda ta zwróciła najniższą wartość statystyki dopasowania, $P/df = 1.61$. Po transformacji rozkład zmiennej stał się mniej symetryczny (skośność: -0.31 vs -0.31 przed). W celu wyboru metody normalizującej zmienną Cele wykonano serię przekształceń. Analiza wykazała, że najlepsze właściwości normalizacyjne zwracała technika `orderNorm`. W porównaniu do innych popularnych przekształceń, metoda ta zwróciła najniższą wartość statystyki dopasowania, $P/df = 1.64$. Po transformacji rozkład zmiennej stał się bardziej symetryczny (skośność: 0.13 vs 0.45 przed). W celu wyboru metody normalizującej zmienną Łatwość wykonano serię przekształceń. Analiza wykazała, że najlepsze właściwości normalizacyjne zwracała technika `boxcox`. W porównaniu do innych popularnych przekształceń, metoda ta zwróciła najniższą wartość statystyki dopasowania, $P/df = 3.21$. Po transformacji rozkład zmiennej stał się bardziej symetryczny (skośność: -0.04 vs -0.52 przed). W celu wyboru metody normalizującej zmienną Użyteczność wykonano serię przekształceń. Analiza wykazała, że najlepsze właściwości normalizacyjne zwracała technika `center_scale`. W porównaniu do innych popularnych przekształceń, metoda ta zwróciła najniższą wartość statystyki dopasowania, $P/df = 1.53$. Po transformacji rozkład zmiennej stał się bardziej symetryczny (skośność: -0.32 vs -0.32 przed). W celu wyboru metody normalizującej zmienną Intencja zakupu wykonano serię przekształceń. Analiza wykazała, że najlepsze właściwości normalizacyjne zwracała technika `orderNorm`. W porównaniu do innych popularnych przekształceń, metoda ta zwróciła najniższą wartość statystyki dopasowania, $P/df = 1.39$. Po transformacji rozkład zmiennej stał się mniej symetryczny (skośność: 0.08 vs 0.03 przed).

Porównanie rozkładów przed i po normalizacji

Wykres nr 9

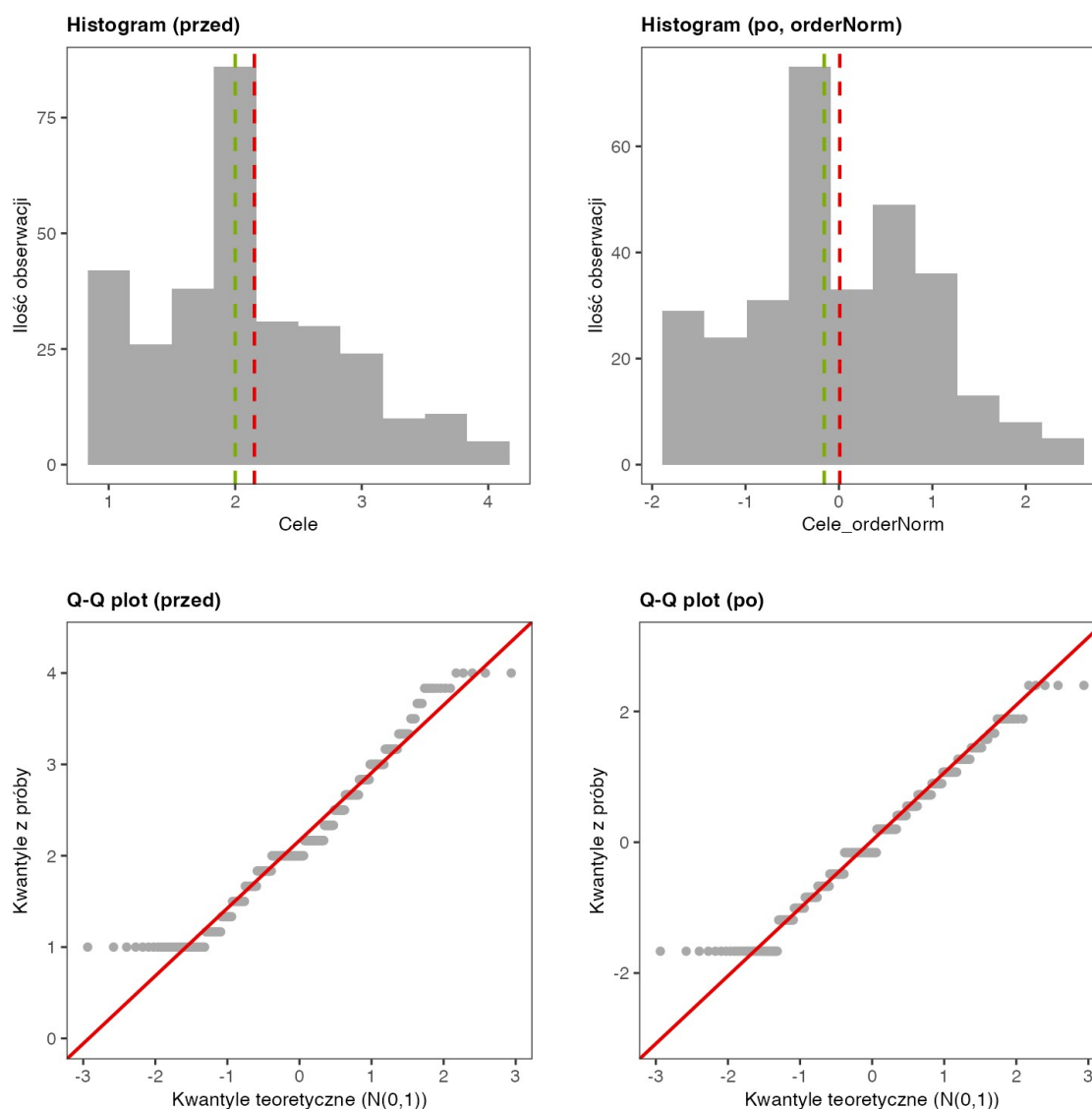
Aby umożliwić wizualną ocenę efektu zastosowanych przekształceń, dla każdej zmiennej numerycznej zestawiono histogram rozkładu oryginalnego (przed normalizacją) z histogramem rozkładu po normalizacji oraz wykresy kwantyl-kwantyl (Q-Q) względem teoretycznego rozkładu normalnego. Wizualizacje wyników przedstawia seria rysunków od nr 7 do nr 13. Rysunek nr 10 Porównanie rozkładu wartości zmiennej Obawy przed i po normalizacji (metoda: center_scale).



Nota. Aby umożliwić wizualną ocenę efektu zastosowanych przekształceń, dla każdej zmiennej numerycznej zestawiono histogram rozkładu oryginalnego (przed normalizacją) z histogramem rozkładu po normalizacji oraz wykresy kwantyl-kwantyl (Q-Q) względem teoretycznego rozkładu normalnego. Wizualizacje wyników przedstawia seria rysunków od nr 1 do nr 5. Rysunek nr 1 Porównanie rozkładu wartości zmiennej Obawy przed i po normalizacji (metoda: center_scale). *Nota :* Na histogramach kolor czerwony przedstawia Średnią ; kolor zielony przedstawia Medianę . Na Q-Q plot czerwona linia odpowiada idealnemu dopasowaniu do rozkładu normalnego — im bliżej punktów do tej linii, tym rozkład bardziej zbliżony do normalnego.

Wykres nr 10

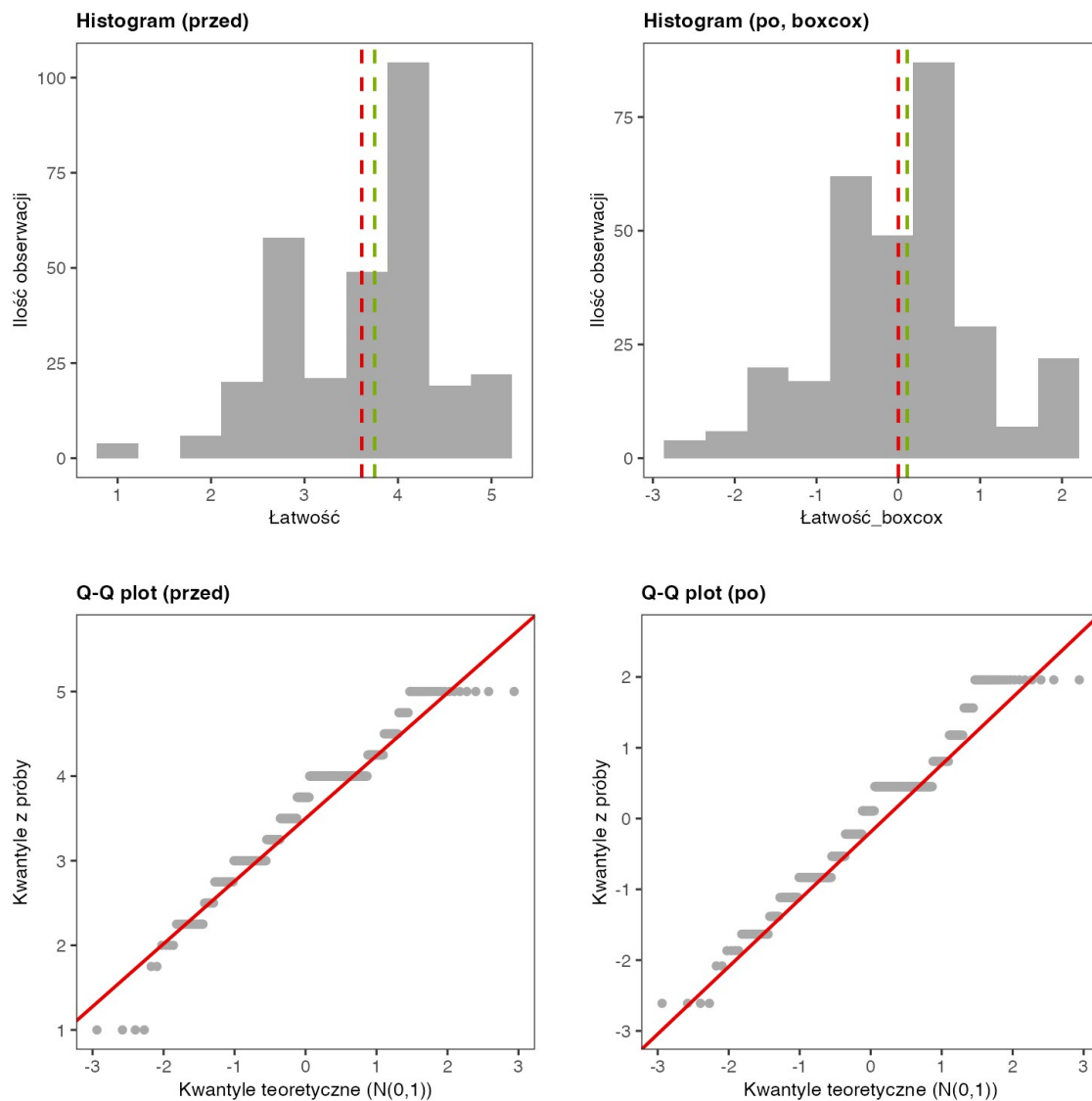
Porównanie rozkładu wartości zmiennej Cele przed i po normalizacji (metoda: orderNorm).



Nota. Rysunek nr 2 Porównanie rozkładu wartości zmiennej Cele przed i po normalizacji (metoda: orderNorm). *Nota :* Na histogramach kolor czerwony przedstawia Średnią ; kolor zielony przedstawia Medianę . Na Q-Q plot czerwona linia odpowiada idealnemu dopasowaniu do rozkładu normalnego — im bliżej punktów do tej linii, tym rozkład bardziej zbliżony do normalnego.

Wykres nr 11

Porównanie rozkładu wartości zmiennej Łatwość przed i po normalizacji (metoda: boxcox).

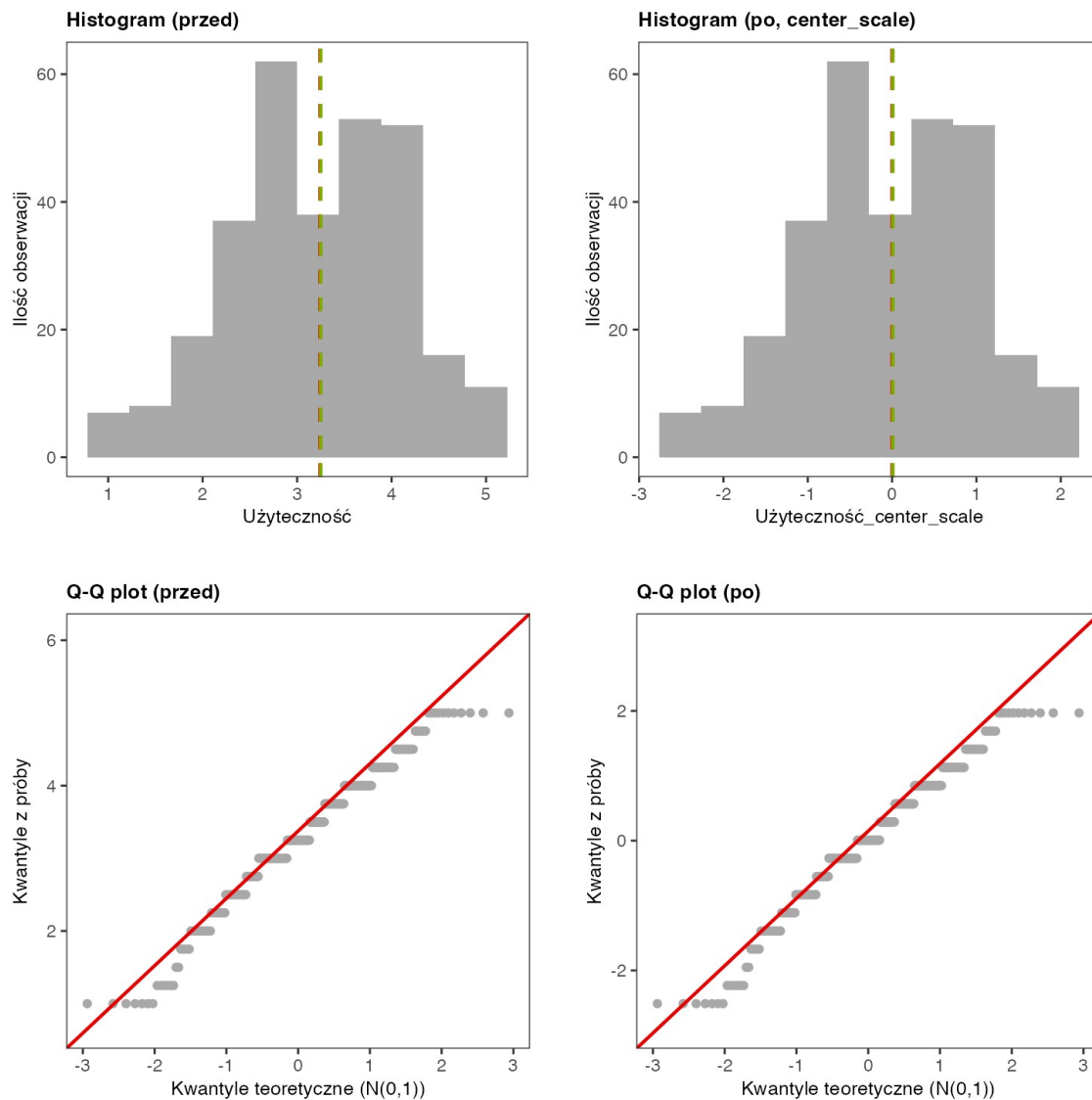


Nota. Rysunek nr 3 Porównanie rozkładu wartości zmiennej Łatwość przed i po normalizacji (metoda: boxcox).

Nota : Na histogramach kolor czerwony przedstawia Średnią ; kolor zielony przedstawia Medianę . Na Q-Q plot czerwona linia odpowiada idealnemu dopasowaniu do rozkładu normalnego — im bliżej punktów do tej linii, tym rozkład bardziej zbliżony do normalnego.

Wykres nr 12

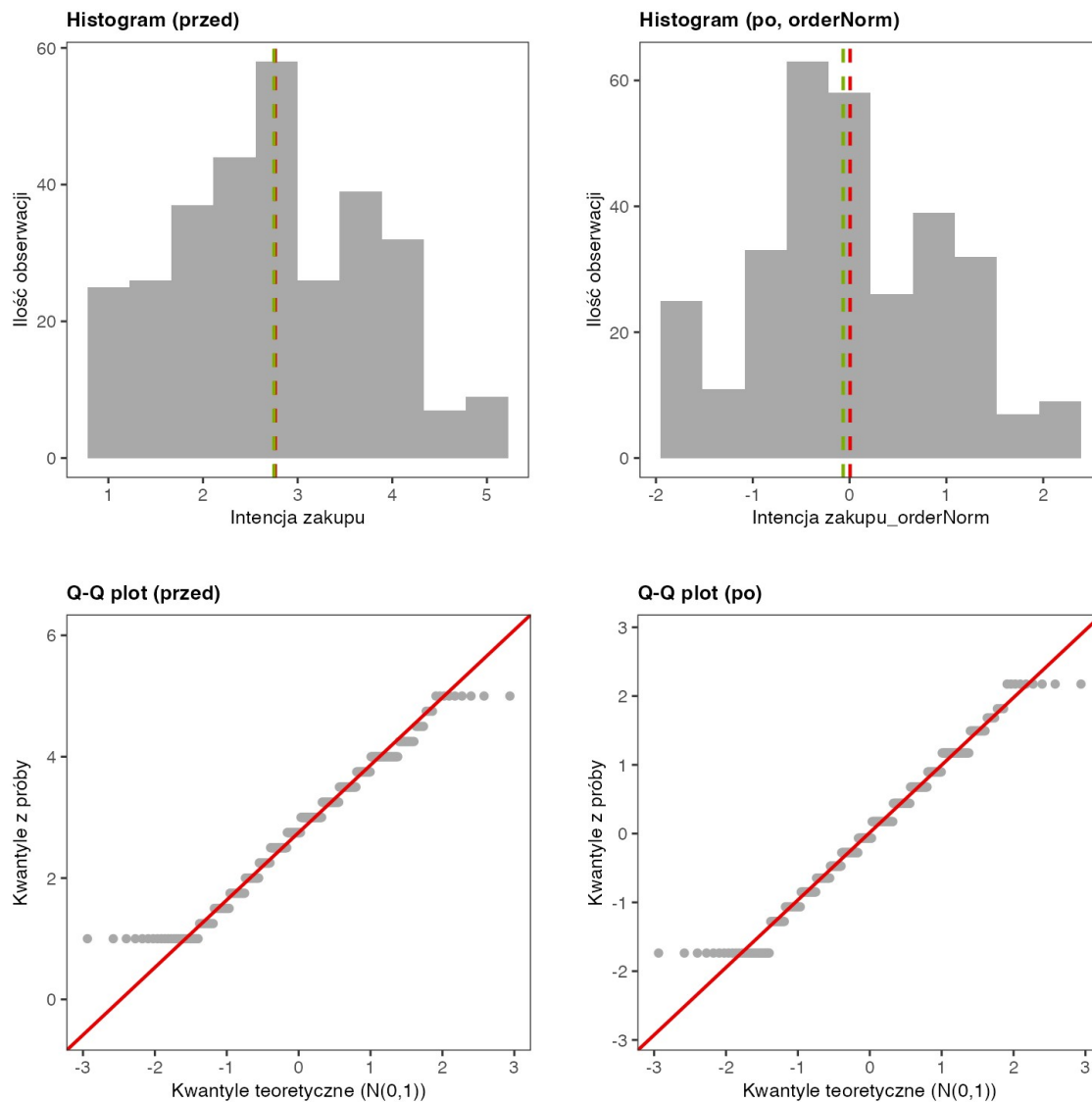
Porównanie rozkładu wartości zmiennej *Użyteczność* przed i po normalizacji (metoda: *center_scale*).



Nota. Rysunek nr 4 Porównanie rozkładu wartości zmiennej *Użyteczność* przed i po normalizacji (metoda: *center_scale*). *Nota :* Na histogramach kolor czerwony przedstawia Średnią ; kolor zielony przedstawia Medianę . Na Q-Q plot czerwona linia odpowiada idealnemu dopasowaniu do rozkładu normalnego — im bliżej punktów do tej linii, tym rozkład bardziej zbliżony do normalnego.

Wykres nr 13

Porównanie rozkładu wartości zmiennej *Intencja zakupu* przed i po normalizacji (metoda: *orderNorm*).



Nota. Rysunek nr 5 Porównanie rozkładu wartości zmiennej *Intencja zakupu* przed i po normalizacji (metoda: *orderNorm*). *Nota :* Na histogramach kolor czerwony przedstawia Średnią ; kolor zielony przedstawia Medianę . Na Q-Q plot czerwona linia odpowiada idealnemu dopasowaniu do rozkładu normalnego — im bliżej punktów do tej linii, tym rozkład bardziej zbliżony do normalnego.

Tabela nr 3

Statystyki opisowe oraz wyniki testów normalności (Shapiro-Wilka i Kolmogorova-Smirnova z poprawką Lilieforce'a) przed i po normalizacji

Zmienna	Technika	N	M	Przed normalizacją					Po normalizacji					
				SD	Skoś.	Kurt.	p SW	p KS	M	SD	Skoś.	Kurt.	p SW	p KS
Obawy	center_scale	303	2.79	0.71	-0.31	-0.57	< 0.001	< 0.001	0.00	1.00	-0.31	-0.57	< 0.001	< 0.001
Cele	orderNorm	303	2.15	0.74	0.45	-0.21	< 0.001	< 0.001	0.01	0.97	0.13	-0.44	< 0.001	< 0.001
Łatwość	boxcox	303	3.61	0.79	-0.52	0.51	< 0.001	< 0.001	0.00	1.00	-0.04	-0.20	< 0.001	< 0.001
Użyteczność	center_scale	303	3.24	0.89	-0.32	-0.05	< 0.001	< 0.001	0.00	1.00	-0.32	-0.05	< 0.001	< 0.001
Intencja zakupu	orderNorm	303	2.76	1.03	0.03	-0.68	< 0.001	< 0.001	0.01	0.97	0.08	-0.48	< 0.001	0.021

Nota. Skoś. = skośność (typ 2); Kurt. = kurtoza (typ 2); p SW = istotność testu Shapiro-Wilka; p KS = istotność testu Kolmogorova-Smirnova z poprawką Lilieforce'a.

Bibliografia

- Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26 (2), 211–243.
- Dallal, G.E. and Wilkinson, L. (1986) An analytic approximation to the distribution of Lilliefors' test for normality. *The American Statistician* , 40, 294–296.
- Hryniewicz, K., Milewska, A. (2023). SZTOS: System Zautomatyzowanego Tworzenia Opisu Statystycznego (Wersja SZTOS) [Oprogramowanie]. <https://sztos-it.com/>
- Joanes, D. N., & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47 (1), 183–189.
- Peterson, R. A. (2021). Finding Optimal Normalizing Transformations via bestNormalize. *The R Journal*, 13:1, 310-329, DOI:10.32614/RJ-2021-041 .
- Royston, J. P. (1982). An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* , 31(2), 115–124. <https://doi.org/10.2307/2347973>
- Ryan A. Peterson & Joseph E. Cavanaugh (2020). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*, 47:13-15, 2312-2327, DOI: 10.1080/02664763.2019.1630372