Praktyczne zastosowanie imputacji wielokrotnej metodą FCS na potrzeby analizy biznesowej zbioru danych z brakami danych.

Autor: Adrian Żelazek

Cel projektu: Praktyczne wykorzystanie imputacji wielokrotnej na potrzeby analizy biznesowej zbioru danych z brakami danych. Imputacja właściwa została wykonana za pomocą procedury PROC MI przy użyciu opcji FCS. Projekt został wykonany przy wykorzystaniu języka SAS 4GL oraz narzędzia SAS Enterprise Guide.

Spis treści

1. Wprowadzenie

Występowanie braków danych w badaniach społecznych jest sytuacją powszechną. Powodów takiego stanu rzeczy może być wiele: odmowa udziału w badaniu, niechęć bądź brak możliwości udzielenia odpowiedzi na pytanie przez respondenta czy przypadkowy błąd ankietera. Niekompletność danych może prowadzić do zniekształcenia rozkładów analizowanych zmiennych czy wzrostu wariancji i obciążenia estymatorów. W efekcie ignorowanie problemu występowania braków danych może zniekształcać otrzymane wyniki oraz skutkować wyciągnięciem przez badacza błędnych wniosków.

W celu zbadania wpływu imputacji danych na realnych danych przeprowadzono niniejsze badanie. Pierwszym etapem pracy było wybranie na podstawie literatury fachowej zmiennych mogących mieć wpływ na wielkość sprzedaży oraz postawienie hipotezy badawczej. Kolejnym krokiem było dokonanie oceny jakości danych oraz przedstawienie charakterystyki zmiennych wybranych do modelowania by w ostatecznym kroku zbudować model odpowiedni do posiadanych danych, w tym wypadku uogólniony model regresji liniowej dla zmiennej o rozkładzie Gamma. W celu porównania wpływu braków danych model został wykonany dwukrotnie: przed i po imputacji.

2. Pytanie badawcze i przegląd rozważań na podjęty w pracy temat

Analiza sprzedaży jest niezwykle istotnym procesem polegającym na analizowaniu, interpretacji i ewaluacji nieustannie zbieranych informacji w zakresie sprzedaży produktów bądź usług danego przedsiębiorstwa oraz różnorakich czynników warunkujących wzrost lub spadek sprzedaży, a także odchyleń od planowanych wyników sprzedażowych. Jej celem jest identyfikacja determinantów, które ułatwiają bądź utrudniają uzyskanie planowanych wielkości sprzedaży. Wnioski płynące z analizy sprzedaży mogą być wykorzystane w celu dopracowania produktu lub usługi, dopasowania i skorygowania strategii oraz metod sprzedaży, planowania sprzedaży, oceny skuteczności i efektywności pracy pracowników, oceny satysfakcji klientów oraz kreowania odpowiedniej komunikacji sprzedażowej z potencjalnymi klientami.

Przed przystąpieniem do wyboru zmiennych objaśniających dokonano przeglądu literatury fachowej, aby jak najlepiej dopasować je do charakteru przeprowadzanej analizy. Z trzeciej edycji raportu *Salesforce: State of Sales* wynika, że wpływ na wielkość sprzedaży mają czynniki takie jak: lokalizacja biznesu, stosowanie nowoczesnych metod marketingowych oraz

wielkość przedsiębiorstwa. Na tej podstawie w niniejszej pracy postawiono hipotezę, że **większe przedsiębiorstwa pod względem liczby zatrudnionych pracowników, operujące w dużych miastach oraz korzystające z nowoczesnych kanałów marketingowych osiągają większą sprzedaż.**

Na bazie artykułów *R&D: Its relationship to company performance* oraz *R&D Spending And Profitability: What's The Link?*, zgodnie z ich autorami, sformułowano hipotezę głoszącą, że **wzrost wydatków na badania i rozwój nie skutkuje wzrostem sprzedaży.**

Rosnąca świadomość konsumentów i przywiązanie do marki sprawiają, że przedsiębiorstwa muszą przykładać dużą wagę do jakości wytwarzanych produktów i świadczonych usług. Uzyskanie powszechnie uznawanego certyfikatu jakości przekłada się na zauważalny wzrost jakości produkowanych wyrobów. **W pracy założono, że posiadanie przez przedsiębiorstwo międzynarodowego certyfikatu jakości skutkuje wzrostem sprzedaży.**

W pracy postanowiono także sprawdzić, **czy większe doświadczenie najwyższego managera w danym sektorze ma pozytywny wpływ na sprzedaż przedsiębiorstwa**, co wydaje się być intuicyjnym założeniem. Jako ostatnią do analizy włączoną zmienną wyrażającą procentowy udział kapitału pożyczonego od banków w całkowitym kapitale obrotowym przedsiębiorstwa, w celu sprawdzenia wpływu na wielkość sprzedaży.

3. Ocena jakości danych

3.1 Opis zbioru

Do niniejszego projektu użyto danych pochodzących z przeprowadzonego przez EBRD (skrót od European Bank for Reconstruction and Development - Europejski Bank Odbudowy i Rozwoju) badania panelowego BEEPS, poświęconego badaniu środowiska biznesowego i wydajności przedsiębiorstw. Dane wykorzystane w projekcie pochodzą z czwartej oraz piątej rundy badania. Zbiór wyjściowy składał się z 495 zmiennych oraz 9655 obserwacji.

3.2 Zmienne wybrane do modelu

Spośród wszystkich zmiennych badania, do projektu wybrano zmienne przedstawione w tabeli poniżej.

Tabela 1. Spis zmiennych wybranych do analizy

Nazwa zmiennej	Typ zmiennej	Opis zmiennej	Zakres przyjmowanych wartości/sposób kodowania zmiennych
d2	Ciągła, objaśniana	Łączna, roczna sprzedaż zakładu w ostatnim roku obrotowym	
a3	Porządkowa	Populacja miasta, w którym zakład prowadzi działalność	1 = Stolica 2 = Miasto z populacją większą niż 1 milion, ale inne niż stolica 3 = Populacja miasta ponad 250 000 do 1 miliona 4 = Populacja: 50000-250000 5 = Populacja mniejsza niż 50000
a6b	Porządkowa	Wielkość zakładu mierzona ilością zatrudnionych pracowników	0 = mikroprzedsiębiorstwo <5 1 = małe >=5 i <=19 2 = średnie >=20 i <=99 3 = duże >=100
b6b	Ilościowa	Rok, w którym zakład został formalnie zarejestrowany	-7 = zakład nigdy nie został zarejestrowany
b7	Ilościowa	Lata doświadczenia najwyższego managera w tym sektorze	1 = poniżej jednego roku
b8	Dychotomiczna	Czy zakład posiada certyfikat jakości?	1= tak 2 = nie
H1	Dychotomiczna	Czy w przeciągu ostatnich 3 lat został wprowadzony na rynek nowy produktu/usługi?	1 = tak 2 = nie
H5	Dychotomiczna	Czy w przeciągu ostatnich 3 lat został wprowadzone nowe metody marketingowe?	1 = tak 2 = nie

H6	Dychotomiczna	Czy w przeciągu ostatnich 3 lat zakład poniósł wydatki na badania i rozwój?	1 = tak 2 = nie
k3bc	Ilościowa	Procentowy udział kapitału pożyczonego od banków (prywatnych i państwowych) w całkowitym kapitale obrotowym przedsiębiorstwa)	0-100%
e2b	Ilościowa	Ilu konkurentów zetknęło się z głównym produktem/linią produkcyjną zakładu?	-4 = zbyt wielu, aby policzyć

Źródło: opracowanie własne na podstawie danych BEEPS

Analizowany zbiór składa się z 10 zmiennych objaśniających oraz jednej zmiennej celu. W ramach dalszej analizy zbiór został ograniczony do przeanalizowania badanego zagadnienia do Polski oraz strefy EURO, czyli do 7 państw: Czech, Słowacji, Słowenii, Chorwacji, Niemiec, Grecji oraz Polski. Z racji tego, iż w Polsce występuje inna waluta niż w państwach strefy EURO, to kwoty podane w złotówkach zostały przeliczone na euro w celu ujednolicenia waluty (przyjęty kurs PLN/EURO: 4.1925). Ponad to ze zbioru usunięto również rekordy, które odznaczały się brakiem danych, bądź których odpowiedzią na zadane pytanie, były takie odpowiedzi, jak: nie wiem, brak odpowiedzi. Po wszystkich operacjach na zmiennych finalny zbiór danych wykorzystanych do projektu liczył 11 zmiennych oraz 578 obserwacji.

4. Analiza struktury w świetle analizy właściwej

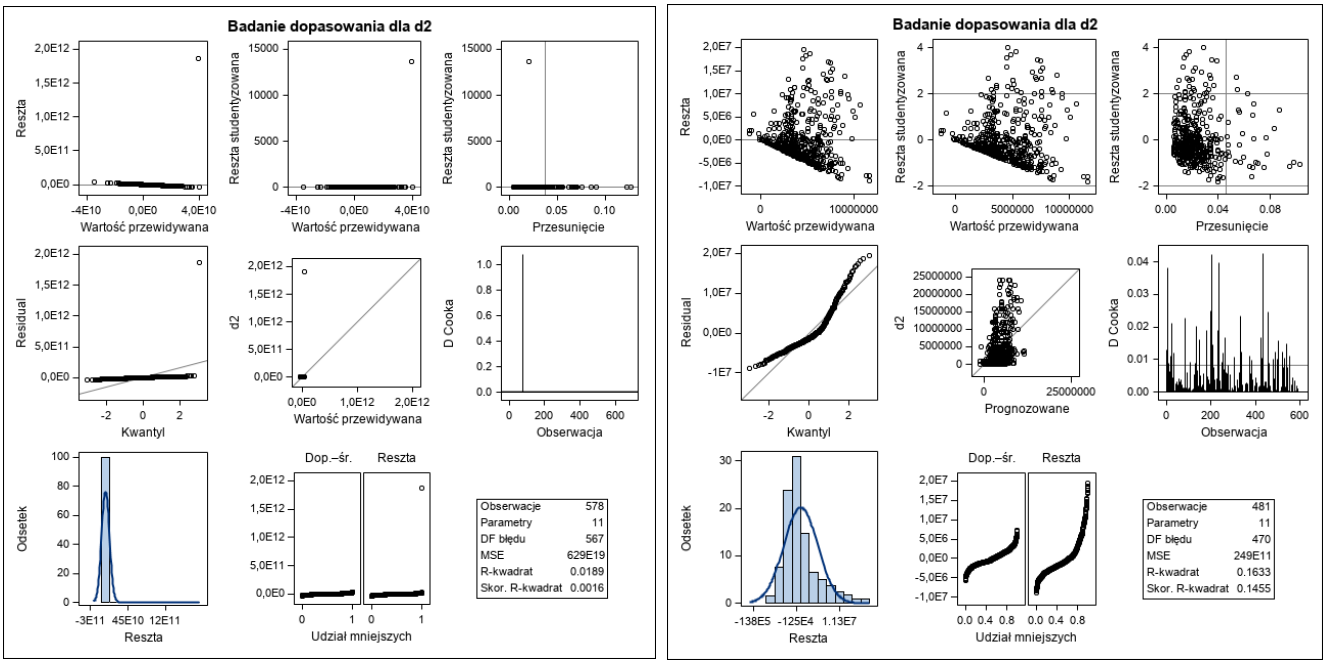
W niniejszym rozdziale została przedstawiona analiza struktury, która miała na celu przybliżenie opisu statystycznego próby, która była również wykorzystana w analizie właściwej

4.1 Badanie współliniowości i rozkład zmiennej celu

W celu zbadania współliniowości zmiennych została wyznaczona statysytka VIF – Współczynnik wariancji inflacji, który pozwala określić czy dany predyktor nie jest skorelowany z innymi predyktorami występującymi w modelu. Wartość statystyki VIF dla wszystkich zmiennych przyjmuje wartość 1, oznacza to brak współliniowości predyktorów.

Następnie zostało zweryfikowane dopasowanie zmiennej objaśnianej d2 najpierw dla całego zbioru, a następnie po usunięciu obserwacji odstających – finalnie 481 obserwacji. Wyniki badania rozkładu zmiennej d2 przed i po usunięciu obserwacji odstających, zostały zaprezentowane na poniższych wykresach.

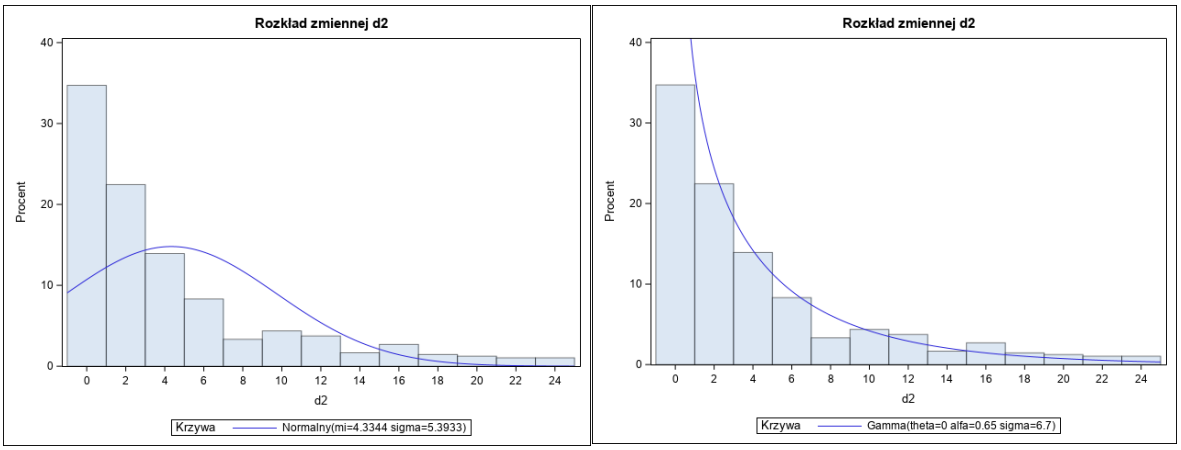
Rysunek 1. Rozkład zmiennej d2 przed i po usunięciu obserwacji odstających



Źródło: opracowanie własne

Zmienna objaśniana d2 została przeskalowana poprzez podzielenie wartości przez 1 milion jednostek w celu poprawienia czytelności danych przy analizie danych i interpretacji wyników. Następnie został zbadany rozkład zmiennej d2. Lepsze dopasowanie do danych uzyskał rozkład gamma co widać na rysunku 2., dlatego to on został użyty w dalszym modelowaniu.

Rysunek 2. Dopasowanie rozkładu zmiennej d2 do rozkładu normalnego i do rozkładu gamma



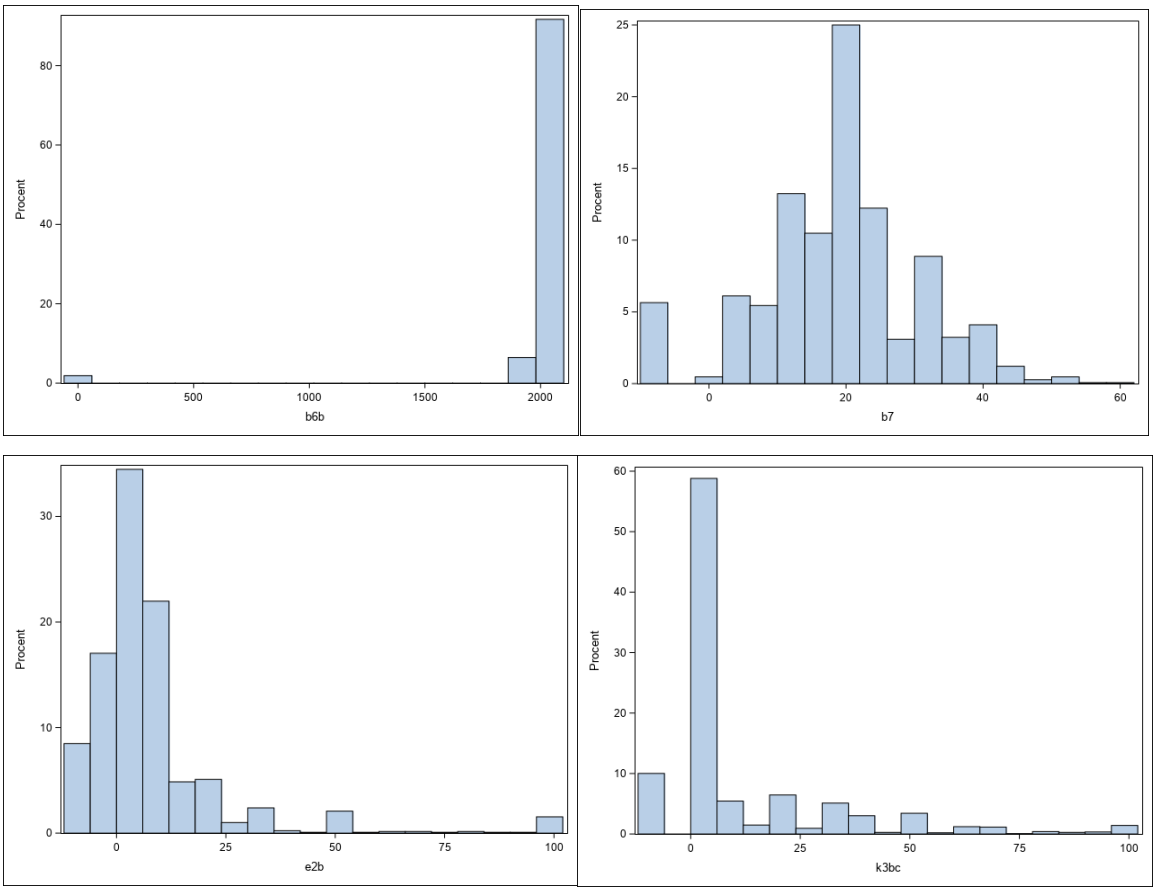
Źródło: opracowanie własne

4.2 Rozkład zmiennych objaśniających

W poniższej tabeli zostały przedstawione zmienne, ich histogramy, przedstawiające rozkłady zmiennych oraz informacje dodatkowe. W przypadku zmiennych binarnych - o ich liczebności, jak i podstawowe statystyki dla zmiennych ciągłych.

Na poniższym zestawie histogramów dla zmiennych, które podczas analizy właściwej (po imputacji) okazały się nieistotne statystycznie stwierdzono, że wiek menedżerów określony jest rozkładem normalnym, podobnie jak liczbę konkurentów badanych firm w zakresie głównego produktu, jednak w tym przypadku rozkład jest skośny prawostronnie. Zdecydowana większość firm nie finansowała działalności kredytem.

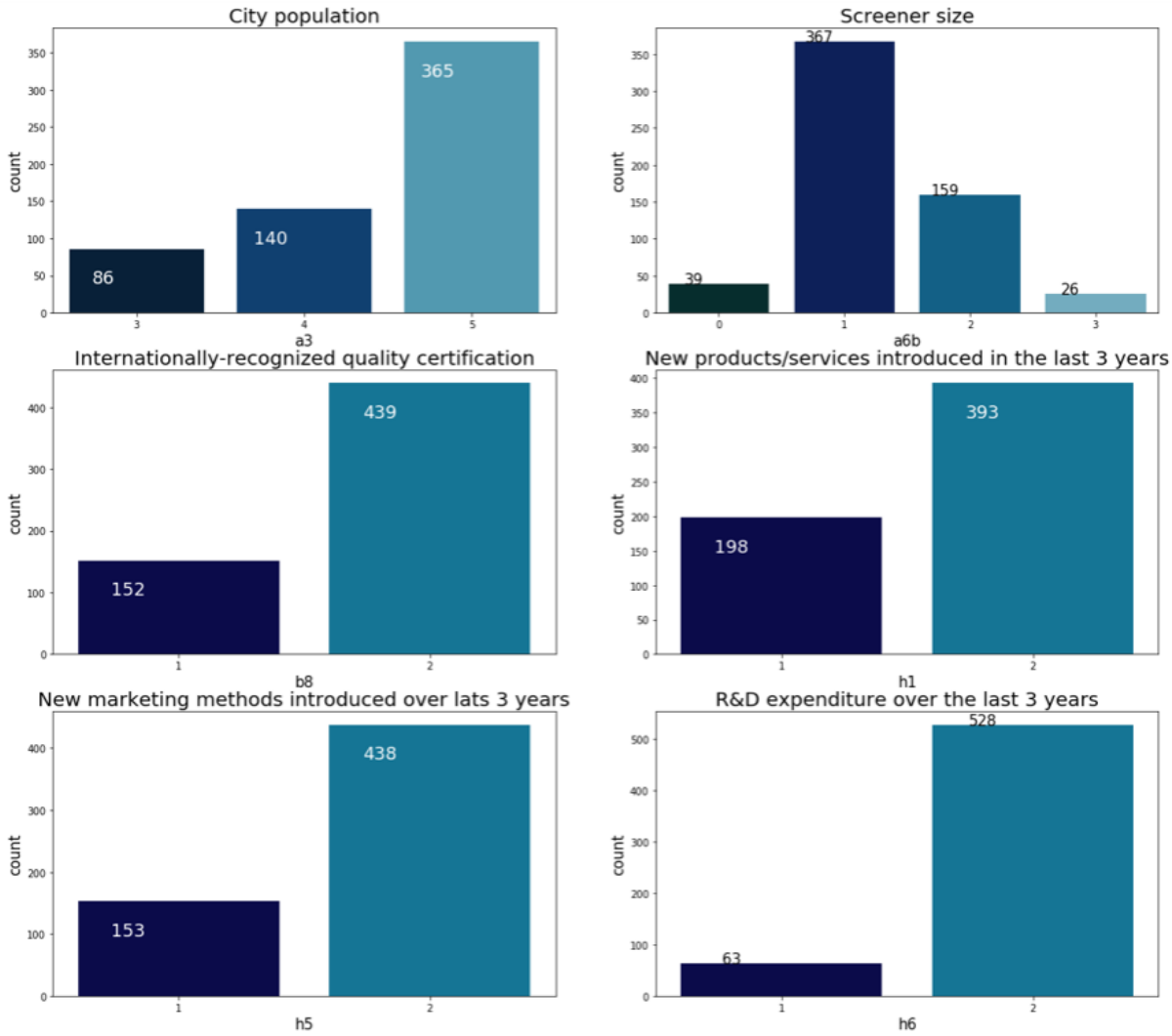
Rysunek 2. Histogramy zmiennych ilościowych



Źródło: Opracowanie własne

Dane, które okazały się istotne dla modelu zostały przedstawione w rys. nr 3 wygenerowanym w Python 3. Odpowiedzi zmiennych binarnych były kodowane wg reguły 1= tak, 2 nie (patrz tabela x). Większość badanych przedsiębiorstw nie posiadało certyfikatu, nie wprowadzało w ciągu ostatnich 3 lat nowego produktu, nie korzystało z nowych kanałów dystrybucji marketingowej, zdecydowana większość nie ponosiło kosztów na badania i rozwój, co oddaje stan próby jako składających się głównie z naśladowców (89%). Większość badanych spółek prowadziła działalność na terenie miejscowości poniżej 50.000 mieszkańców. Zdecydowana większość badanych firm to firmy małe.

Rysunek 3. Histogramy zmiennych jakościowych



Źródło: Opracowanie własne

Podstawowe statystyki zmiennych użytych w modelu zostały zawarte w tabeli nr 2.

Tabela 2. Statystyki opisowe zmiennych objaśniających

Procedura MEANS

Zmienna	N	Średnia	Mediana	Odch. std.	Skośność	Kurtoza	Minimum	Maksimum
a1	689	72.2699565	77.0000000	8.2657345	-0.8902830	-1.0400140	59.0000000	79.0000000
a3	689	4.4833091	5.0000000	0.7254290	-1.0300041	-0.3688301	3.0000000	5.0000000
a6b	689	1.4121916	1.0000000	0.7342707	0.5709439	-0.0355473	0	3.0000000
b6b	689	1993.86	1995.00	12.7854241	-3.2249298	16.4132932	1894.00	2012.00
b7	689	20.6705370	20.0000000	9.7951557	0.4917166	0.3636593	1.0000000	60.0000000
b8	689	1.7155298	2.0000000	0.4514895	-0.9575292	-1.0862995	1.0000000	2.0000000
e2b	689	12.1552975	6.0000000	16.6769225	3.4135136	13.4724089	0	100.0000000
h1	689	1.6487663	2.0000000	0.4777026	-0.6246540	-1.6145024	1.0000000	2.0000000
h5	689	1.7140784	2.0000000	0.4521805	-0.9496279	-1.1014125	1.0000000	2.0000000
h6	689	1.8708273	2.0000000	0.3356347	-2.2161409	2.9197473	1.0000000	2.0000000
k3bc	689	12.4092888	0	21.7704922	2.1317996	4.4316093	0	100.0000000
d2	578	3328320283	3163961.73	79369183998	24.0415104	577.9961382	7708.00	1.9081884E12

5. Analiza właściwa

W celu dokonania analizy sprzedaży wykorzystano uogólniony model liniowy, z rozkładem Gamma zmiennej celu. W celu zbadania wpływu imputacji zbudowane zostały dwa Uogólnione modele regresji liniowej z zmienną celu d2, które w dalszej części zostaną porównane pod kątem zmian jakie zaszły po wprowadzeniu imputacji wielokrotnej.

Do budowy pierwszego modelu został wykorzystany zbiór z brakami danych w zmiennej D2. Procedura proc genmod, wykorzystana do budowy modelu automatycznie wyeliminowała z analizy obserwacje posiadające braki. Wyniki dopasowania dla poszczególnych zmiennych przedstawia Tabela 3. Analiza parametrów modelu UML przed imputacją2. Na podstawie statystyki Chi-kwadrat w modelu istotne okazały się zmienne a6b, b7 i b8 (przyjęte p-value=0.05).

Tabela 3. Analiza parametrów modelu UML przed imputacją

Testy efektów stałych typu 3			
Efekt	**DF**	**Chi-kwadrat**	**Pr. > chi-kw.**
a3	2	4,84	0,089
a6b	3	68,38	<.0001
b6b	1	0	0,9874
b7	1	4,60	0,0320
b8	1	4,88	0,0272
h1	1	0,08	0,7710
h5	1	0,14	0,7119
h6	1	2,24	0,1344
k3bc	1	0,20	0,6586
e2b	1	0,92	0,3367

Źródło: Opracowanie własne

Imputacja właściwa została wykonana za pomocą procedury PROC MI przy użyciu opcji FCS. Z powodu braku rozkładu normalnego zmiennej celu użycie metody MCMC – Markov chain Monte Carlo nie jest zalecane. Analiza tej procedury pozwoliła na sprawdzenie struktury braków danych w zbiorze co przedstawia Tabela 4. Struktura braków danych dla zmiennych3. Braki danych występują w 18.61% zbioru.

Tabela 4. Struktura braków danych dla zmiennych

Grupa	d2	a3	a6b	b6b	b7	b8	h1	h5	h6	k3bc	e2b	Liczebn.	Procent
1	X	X	X	X	X	X	X	X	X	X	X	481	81,39
2	.	X	X	X	X	X	X	X	X	X	X	110	18,61

Źródło: Opracowanie własne

Tabela 5. Średnie dla zmiennych ciągłych użytych przy imputacji4 przedstawia porównanie średnich dla każdej zmiennej ciągłej z brakami i bez. Tabela 6. Ocena parametru zmiennej celu po imputacji.6 zawiera wyniki oceny parametrów zmiennej celu po imputacji, pokazuje błąd standardowy i 95% przedziały ufności dla oszacowanej nowej średniej.

Tabela 5. Średnie dla zmiennych ciągłych użytych przy imputacji

	Średnie grup									
d2	**a3**	**a6a**	**b6b**	**b7**	**b8**	**h1**	**h5**	**h6**	**k3bc**	**e2b**
4,33	4,5	1,27	1995,35	20,90	1,74	1,66	1,73	1,89	11,63	12,17
.	4,36	1,4	1995,01	19,37	1,76	1,69	1,78	1,93	10,24	14,73

Źródło: Opracowanie własne

Tabela 6. Ocena parametru zmiennej celu po imputacji.

Oceny parametrów (20 imputacji)									
Zmienna	**Średnia**	**Błąd std.**	**Przedział ufności 95%**		**DF**	**Minimum**	**Maksimum**	**Mi0**	**t dla H0: średnia=mi0**	**Pr. > |t|**
d2	4,353981	0,23968	3,882644	4,825317	368,98	4,248572	4,499012	0	18,17	<.0001

Źródło: Opracowanie własne

Drugi model zbudowano przy zastosowaniu zaimputowanych braków danych. Zbudowano model z tymi samymi parametrami. Wyniki przedstawia Tabela 7. Analiza parametrów modelu UML po imputacji6 i zauważyć można, że wyeliminowana została zmienna b7, która na poziomie 0.05 została odrzucona. W modelu pozostały istotne dwie zmienne: a6b - Wielkość zakładu mierzona ilością zatrudnionych pracowników oraz b8 - Czy zakład posiada certyfikat jakości.

Tabela 7. Analiza parametrów modelu UML po imputacji

Testy efektów stałych typu 3			
Efekt	**DF**	**Chi-kwadrat**	**Pr. > chi-kw.**
a3	2	4,62	0,0990
a6b	3	73,93	<.0001
b6b	1	0,04	0,8343
b7	1	3,27	0,0704
b8	1	5,16	0,0232
h1	1	0,14	0,7119
h5	1	0,12	0,7263
h6	1	2,23	0,1352
k3bc	1	0,06	0,8065
e2b	1	1,13	0,2869

Źródło: Opracowanie własne

Następnie procedurą PROC MINANALYZE została wykonana analiza parametrów istotnych w modelu regresji (Tabela 8. Tabela wariancji i imputacji7), która pokazuje wariancję między imputacjami, względny wzrost wariancji spowodowany brakującymi wartościami i względną wydajnością dla każdej imputowanej zmiennej. Wydajność dla wszystkich parametrów jest bardzo wysoka, osiąga wartość bliską jedności.

Tabela 8. Tabela wariancji i imputacji

Informacje o wariancji (20 imputacji)							
Parametr	**Wariancja**			**DF**	**Wzrost względny wariancji**	**Informacje o udziale braków danych**	**Wydajność względna**
	Pomiędzy	**Wewnątrz**	**Suma**				
a3	0,540	20,540	21,080	18000000,000	0,026	0,026	0,9999980
a6b	0,426	2,093	2,519	412678,000	0,204	0,169	0,9999860
b7	96,634	521,711	618,352	483864,000	0,185	0,156	0,9999870
b8	0,191	3,228	3,420	3790000,000	0,059	0,056	0,9999950
H6	0,095	3,680	3,775	18600000,000	0,026	0,025	0,9999980
e2b	306,758	466,748	773,532	75140,000	0,657	0,397	0,9999660

Źródło: Opracowanie własne

Tabela 9. Wyniki estymacji istotnych parametrów regresji liniowej9 pokazuje błąd szacunkowy i standardowy dla każdego istotnego parametru regresji. Tabela wyświetla tez 95% przedział ufności i

test t powiązany wartością p dla hipotezy, że parametr jest równy wartości określonej za pomocą opcji THETA0. Wszystkie parametry wskazują na brak podstaw do odrzucenia hipotezy zerowej.

Tabela 9. Wyniki estymacji istotnych parametrów regresji liniowej

Oceny parametrów (20 imputacji)										
Parametr	**Ocena**	**Błąd std.**	**Przedział ufności 95%**		**DF**	**Minimum**	**Maksimum**	**Theta0**	**t dla H0: parametr= theta0**	**Pr. > |t|**
a3	4,472	4,591	-4,527	13,471	18000000,000	3,000	5,000	0,000	0,970	0,330
a6b	1,291	1,587	-1,820	4,402	412678,000	0,000	3,000	0,000	0,810	0,416
b7	20,618	24,867	-28,120	69,356	483864,000	1,000	60,000	0,000	0,830	0,407
b8	1,743	1,849	-1,882	5,367	3790000,000	1,000	2,000	0,000	0,940	0,346
H6	1,893	1,943	-1,915	5,702	18600000,000	1,000	2,000	0,000	0,970	0,330
e2b	12,650	27,812	-41,863	67,162	75140,000	0,000	100,000	0,000	0,450	0,649

Źródło: Opracowanie własne

6. Podsumowanie

Celem projektu było zbadanie działania procedury imputacji w praktyce. Ponad to, kolejnym zagadnieniem poddanym badaniu było sprawdzenie, jakie czynniki zasadniczo wpływają na łączną, roczną sprzedaż danego przedsiębiorstwa w ostatnim badanym roku obrotowym. Za pomocą uogólnionego modelu liniowego dla zmiennej z rozkładu gamma oraz przy pomocy imputacji danych zbadano zmienne objaśniające i potwierdzono następujące hipotezy: Posiadanie przez przedsiębiorstwo międzynarodowego certyfikatu jakości skutkuje wzrostem sprzedaży.

Większe przedsiębiorstwa pod względem liczby zatrudnionych pracowników osiągają większą sprzedaż.

Nie potwierdzono zależności sprzedaży od zmiennych odnoszących się do doświadczenia najwyższego managera w danym sektorze oraz wielkości wydatków na badania i rozwój

7. Spis rysunków i tabel