Sprawozdanie

Bednarz Weronika, Inżynieria Obliczeniowa, rok 2, gr. 1

Temat: Analiza zbioru + regresja – Statistica

Przebieg ćwiczenia 1:

1. Otworzono plik *insuranceSAD.xlsx.*

	insuranceSAD					
	age	bmi	children	smoker	charges	female
1	19	27,9	0	1	16884,924	1
2	18	33,77	1	0	1725,5523	0
3	28	33	3	0	4449,462	0
4	33	22,705	0	0	21984,4706	0
5	32	28,88	0	0	3866,8552	0
6	31	25,74	0	0	3756,6216	1
7	46	33,44	1	0	8240,5896	1
8	37	27,74	3	0	7281,5056	1
9	37	29,83	2	0	6406,4107	0
10	60	25,84	0	0	28923,1369	1
11	25	26,22	0	0	2721,3208	0
12	62	26,29	0	1	27808,7251	1
13	23	34,4	0	0	1826,843	0
14	56	39,82	0	0	11090,7178	1
15	27	42,13	0	1	39611,7577	0
16	19	24,6	1	0	1837,237	0
17	52	30,78	1	0	10797,3362	1
18	23	23,845	0	0	2395,17155	0
19	56	40,3	0	0	10602,385	0
20	30	35,3	0	1	36837,467	0
21	60	36,005	0	0	13228,847	1
22	30	32,4	1	0	4149,736	1
23	18	34,1	0	0	1137,011	0
24	34	31,92	1	1	37701,8768	1
25	37	28,025	2	0	6203,90175	0

2. Przygotowano dane do analizy:

a. ustalono, które zmienne to ilościowe, a które jakościowe.

zmienne ilościowe: *age, bmi, charges*
zmienne jakościowe: *smoker, children, female*

b. ustawiono etykiety dla zmiennych *smoker* i *female*

	insuranceSAD					
	age	bmi	children	smoker	charges	female
1	19	27,9	0	1	16884,924	K
2	18	33,77	1	0	1725,5523	M
3	28	33	3	0	4449,462	M
4	33	22,705	0	0	21984,4706	M
5	32	28,88	0	0	3866,8552	M
6	31	25,74	0	0	3756,6216	K
7	46	33,44	1	0	8240,5896	K
8	37	27,74	3	0	7281,5056	K
9	37	29,83	2	0	6406,4107	M
10	60	25,84	0	0	28923,1369	K
11	25	26,22	0	0	2721,3208	M
12	62	26,29	0	1	27808,7251	K
13	23	34,4	0	0	1826,843	M
14	56	39,82	0	0	11090,7178	K
15	27	42,13	0	1	39611,7577	M
16	19	24,6	1	0	1837,237	M
17	52	30,78	1	0	10797,3362	K
18	23	23,845	0	0	2395,17155	M
19	56	40,3	0	0	10602,385	M
20	30	35,3	0	1	36837,467	M

3. Obliczono statystyki opisowe dla zmiennych ilościowych.

	Descriptive Statistics (insuranceSAD in insuranceSAD)										
Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile	Range	Std.Dev.
age	1338	39,21	39,000	18,00000	69	18,000	64,00	27,000	51,00	46,00	14,05
bmi	1338	30,66	30,400	32,30000	13	15,960	53,13	26,290	34,70	37,17	6,10
charges	1338	13270,42	9382,033	1639,563	2	1121,874	63770,43	4738,268	16657,72	62648,55	12110,01

Średni wiek badanej grupy ludzi wynosi 39,21 lat, a odchylenie standardowe dla niego to 14,05 lat. Mediana, czyli wartość, która dzieli grupę na połowę, wynosi 39 lat. Można więc wnioskować, że rozkład wieku w tej grupie jest zbliżony do rozkładu normalnego. Nie ma jednak symetrii, ponieważ moda (najczęściej występujący wiek) jest inna (nie jest nawet zbliżona) niż mediana. Najwięcej osób miało 18 lat (69 osób). Dolny kwartyl, czyli wartość, poniżej której znajduje się 25% osób, to 27 lat, a górny kwartyl, czyli wartość, poniżej której znajduje się 75% osób, to 51 lat. Różnica pomiędzy najstarszą a najmłodszą osobą w grupie (rozstęp) wynosi 46 lat.

4. Przygotowano tabele liczności dla zmiennych jakościowych.

Category	Frequency table: children (insuranceSAD in insuranceSAD)			
	Count	Cumulative Count	Percent	Cumulative Percent
0	574	574	42,89985	42,8999
1	324	898	24,21525	67,1151
2	240	1138	17,93722	85,0523
3	157	1295	11,73393	96,7862
4	25	1320	1,86846	98,6547
5	18	1338	1,34529	100,0000
Missing	0	1338	0,00000	100,0000

Tabela powyżej przedstawia informacje na temat liczby dzieci. Najmniejszą grupą jest ta, która składa się z 18 dzieci.

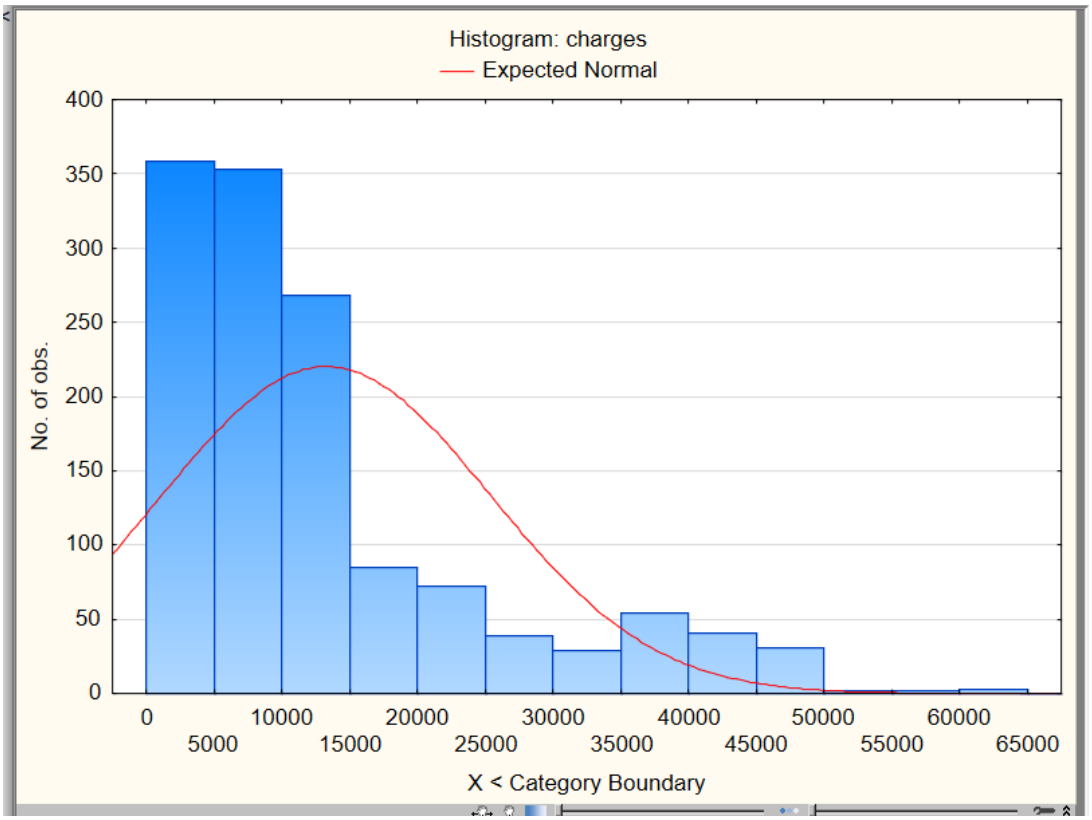
Category	Frequency table: female (insuranceSAD in insuranceSAD)			
	Count	Cumulative Count	Percent	Cumulative Percent
M	676	676	50,52317	50,5232
K	662	1338	49,47683	100,0000
Missing	0	1338	0,00000	100,0000

Tabela powyżej przedstawia informacje dotyczące płci. Liczba mężczyzn i kobiet jest prawie równa.

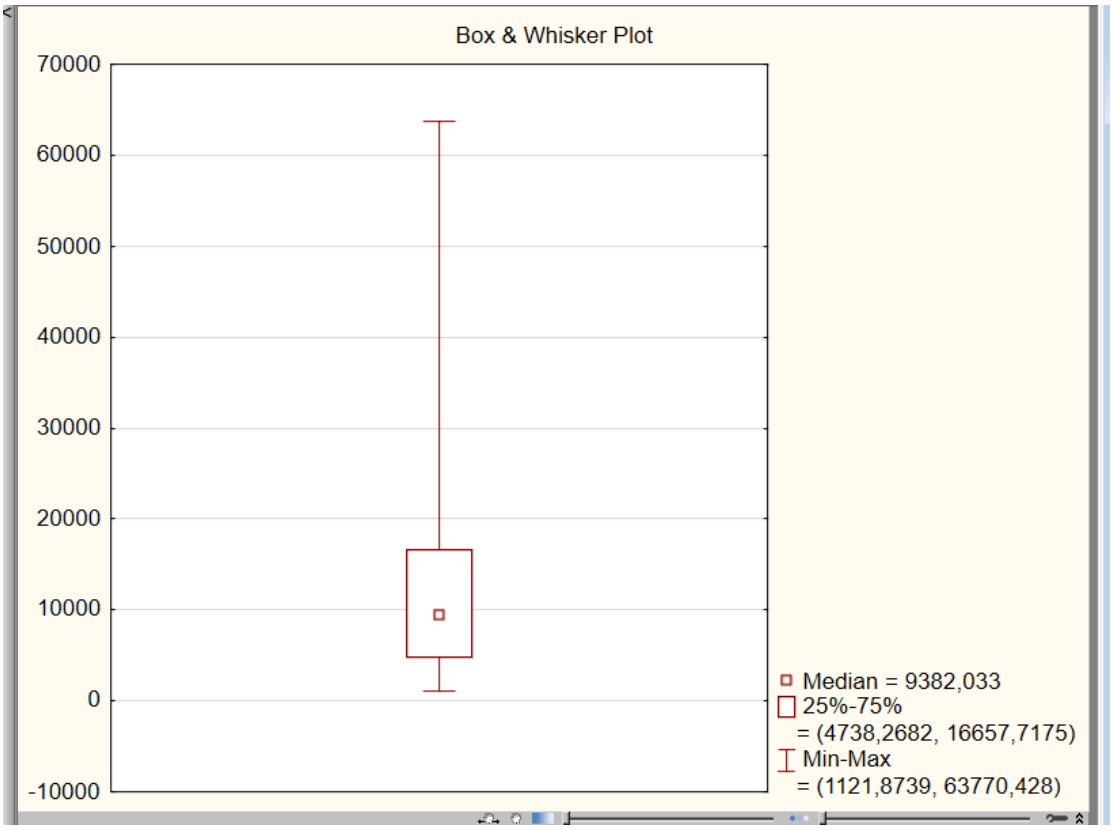
Category	Frequency table: smoker (insuranceSAD in insuranceSAD)			
	Count	Cumulative Count	Percent	Cumulative Percent
0	1064	1064	79,52167	79,5217
1	274	1338	20,47833	100,0000
Missing	0	1338	0,00000	100,0000

Ta tabela odnosi się do zwyczaju palenia. Jest dużo więcej ludzi, którzy nie palą, niż tych, którzy palą.

5. Dla zmiennej *charges* wykonano histogram dla 5000 kroków, zaczynając od 0:

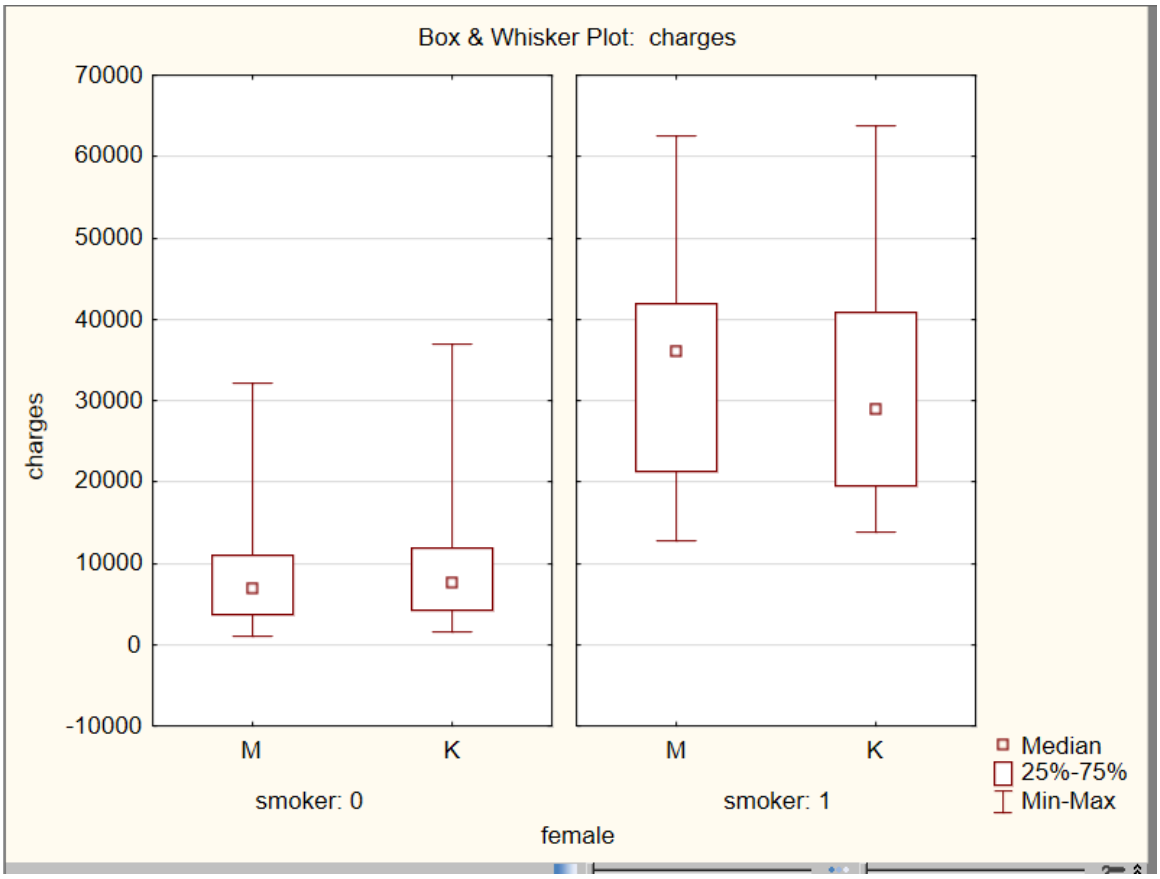


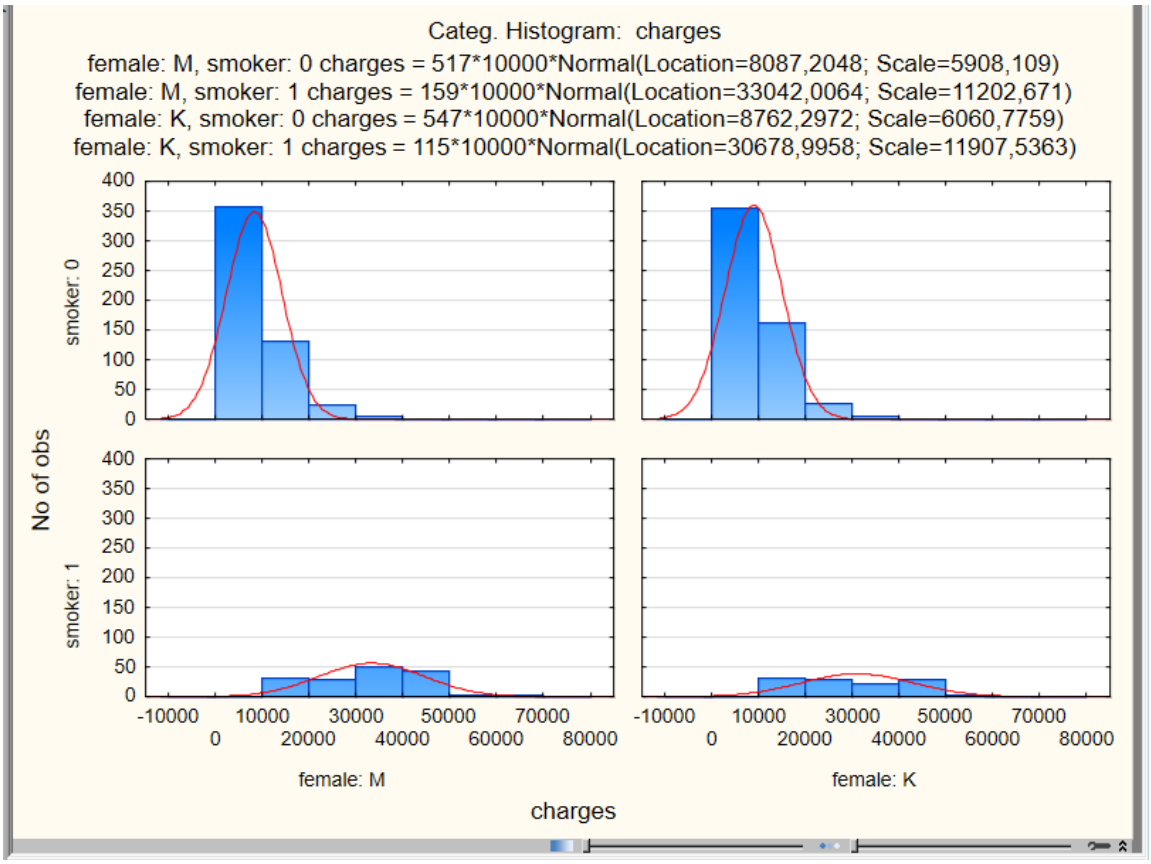
Wykonano wykres ramka-wąsy (mediana/kwartyle).



6.. Dla zmiennej *charges* wykonano skategoryzowane histogramy i wykresy ramka-wąsy.

Kategoryzacji dokonano ze względu na płeć i palenie..





Przebieg ćwiczenia 2:

1. Wczytano zbiór danych insurance_train.xlsx.

	insurance_X_train					
	age	bmi	children	smoker	female	charges
1	46	19,95	2	0	1	9193,8385
2	47	24,32	0	0	1	8534,6718
3	52	24,86	0	0	1	27117,9938
4	39	34,32	5	0	1	8596,8278
5	54	21,47	3	0	1	12475,3513
6	63	41,47	0	0	0	13405,3903
7	22	24,3	0	0	1	2150,469
8	18	21,565	0	1	0	13747,8724
9	40	41,23	1	0	0	6610,1097
10	37	34,2	1	1	0	39047,285
11	34	22,42	2	0	0	27375,9048
12	50	37,07	1	0	0	9048,0273
13	49	29,925	0	0	1	8988,15875
14	64	39,33	0	0	1	14901,5167
15	46	25,8	5	0	0	10096,97
16	50	32,205	0	0	0	8835,26495
17	36	33,4	2	1	0	38415,474
18	25	26,22	0	0	0	2721,3208
19	51	37,73	1	0	1	9877,6077
20	52	41,8	2	1	0	47269,854

2. Sporządzono macierz korelacji dla wszystkich zmiennych.

	Correlations (insurance_X_train in insurance_train) Marked correlations are significant at p < ,05000 N=1070 (Casewise deletion of missing data)							
Variable	**Means**	Std.Dev.	age	bmi	children	smoker	female	charges
age	39,36	14,07	1,000000	0,118274	0,060999	-0,052035	0,008459	0,281721
bmi	30,56	6,04	0,118274	1,000000	-0,005040	-0,003450	-0,015293	0,197316
children	1,11	1,22	0,060999	-0,005040	1,000000	0,013994	-0,017080	0,071885
smoker	0,21	0,40	-0,052035	-0,003450	0,013994	1,000000	-0,070908	0,780063
female	0,49	0,50	0,008459	-0,015293	-0,017080	-0,070908	1,000000	-0,056802
charges	13346,09	12019,51	0,281721	0,197316	0,071885	0,780063	-0,056802	1,000000

Macierz jest symetryczna, ma jedynki na przekątnej.

Zmienne, które są skorelowane ze zmienną *charges: age, smoker*

3. Zbudowano model regresji wielorakiej dla zmiennej *charges*

	Regression Summary for Dependent Variable: charges (insurance_X_train in insurance_train) R= ,86086518 R2= ,74108886 Adjusted R2= ,73987217 F(5,1064)=609,10 p<0,0000 Std.Error of estimate: 6130,3					
N=1070	b*	Std.Err. of b*	b	Std.Err. of b	t(1064)	p-value
Intercept			-12121,4	1088,421	-11,1367	0,000000
age	0,301009	0,015762	257,1	13,462	19,0966	0,000000
bmi	0,164683	0,015713	327,5	31,251	10,4807	0,000000
children	0,043224	0,015634	427,3	154,537	2,7647	0,005795
smoker	0,795712	0,015662	23653,9	465,565	50,8069	0,000000
female	0,000331	0,015643	8,0	375,979	0,0211	0,983130

Wzór na model regresji:

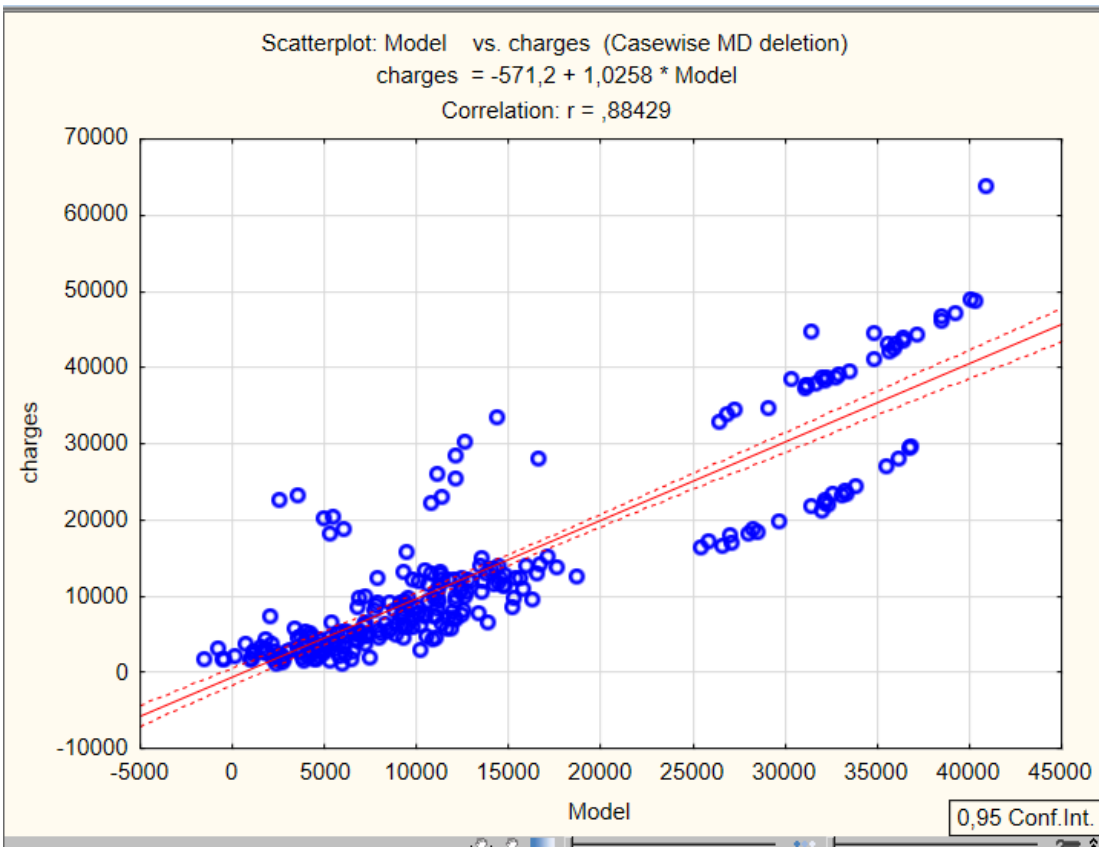
charges = 257,1*age + 327,5*bmi + 427,3*children + 23653,9*smoker + 8*female – 12121,4

4. Otworzono plik insurance_test.xlsx i obliczono wartości przewidywane za pomocą modelu.

	insurance_X_test					
	age	bmi	children	smoker	female	charges
1	45	25,175	2	0	1	9095,06825
2	36	30,02	0	0	1	5272,1758
3	64	26,885	0	1	1	29330,9832
4	46	25,745	3	0	0	9301,89355
5	19	31,92	0	1	0	33750,2918
6	34	42,9	1	0	0	4536,259
7	19	22,515	0	0	1	2117,33885
8	64	37,905	0	0	0	14210,536
9	28	17,29	0	0	1	3732,6251
10	49	28,69	3	0	0	10264,4421
11	30	24,4	3	1	0	18259,216
12	41	30,59	2	0	0	7256,7231
13	29	29,59	1	0	1	3947,4131
14	46	42,35	3	1	0	46151,1245
15	60	40,92	0	1	0	48673,5588
16	47	38,94	2	1	0	44202,6536
17	49	42,68	2	0	1	9800,8882
18	47	36,63	1	1	1	42969,8527

	insurance_X_test							
	age	bmi	children	smoker	female	charges	Model	Reszta
1	45	25,175	2	0	1	9095,06825	8555,5125	539,55575
2	36	30,02	0	0	1	5272,1758	6973,75	1701,5742
3	64	26,885	0	1	1	29330,9832	36799,7375	7468,75435
4	46	25,745	3	0	0	9301,89355	9418,5875	116,69395
5	19	31,92	0	1	0	33750,2918	26871,2	6879,0918
6	34	42,9	1	0	0	4536,259	11097,05	6560,791
7	19	22,515	0	0	1	2117,33885	145,1625	1972,17635
8	64	37,905	0	0	0	14210,536	16746,8875	2536,35155
9	28	17,29	0	0	1	3732,6251	747,875	2984,7501
10	49	28,69	3	0	0	10264,4421	11154,375	889,9329
11	30	24,4	3	1	0	18259,216	28518,4	10259,184
12	41	30,59	2	0	0	7256,7231	9292,525	2035,8019
13	29	29,59	1	0	1	3947,4131	5460,525	1513,1119
14	46	42,35	3	1	0	46151,1245	38510,625	7640,4995
15	60	40,92	0	1	0	48673,5588	40359,8	8313,7588
16	47	38,94	2	1	0	44202,6536	37223,65	6979,0036
17	49	42,68	2	0	1	9800,8882	15316,8	5515,9118
18	47	36,63	1	1	1	42969,8527	36047,825	6922,0277

Wykonano wykres rozrzutu wartości przewidywanych względem obserwowanych.



Współczynnik determinacji wyniósł:

- dla zbioru testowego: r^2 = 0,782

- dla zbioru treningowego: 0,741

Lepsze wyniki otrzymano dla zbioru testowego.

Wnioski:

Badanie danych pozwoliło na pozyskanie informacji o zmiennych, które są mierzone w sposób ilościowy i jakościowy. Dodatkowo, uzyskano opisowe statystyki, co pozwoliło na lepsze zrozumienie charakterystyk i rozkładu danych w każdej zmiennej. Aby przedstawić wyniki w bardziej zrozumiały sposób, wykorzystano histogramy i wykresy typu ramka-wąsy. Analiza korelacji między zmiennymi wykazała, że największy wpływ na wartość zmiennej *charges* miały zmienne *age* i *smoker*. Wartości korelacji między innymi zmiennymi były niskie lub nieistotne.

Przy użyciu modelu regresji wielorakiej, było możliwe przewidywanie wartości zmiennej *charges* na podstawie innych zmiennych niezależnych. Porównanie wyników współczynnika determinacji dla zbioru treningowego i testowego pokazało, że model jest dobrze dopasowany i skutecznie przewiduje wartości dla nowych danych.