Higher School of Economics

Faculty of Computer Science

Financial Technologies and Data Analysis

Can you accurately predict insurance cost?

Выполнили:

Лобачев Никита

Озерова Дарья

Павлеева Мария

Moscow, 2023

Medical Cost Personal Datasets

Content

- **age:** age of primary beneficiary
- **sex:** insurance contractor gender, female, male
- **bmi:** Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- **children:** Number of children covered by health insurance / Number of dependents
- **smoker:** Smoking
- **region:** The beneficiary's residential area in the US, northeast, southwest, northwest.
- **charges:** Individual medical costs billed by health insurance

License

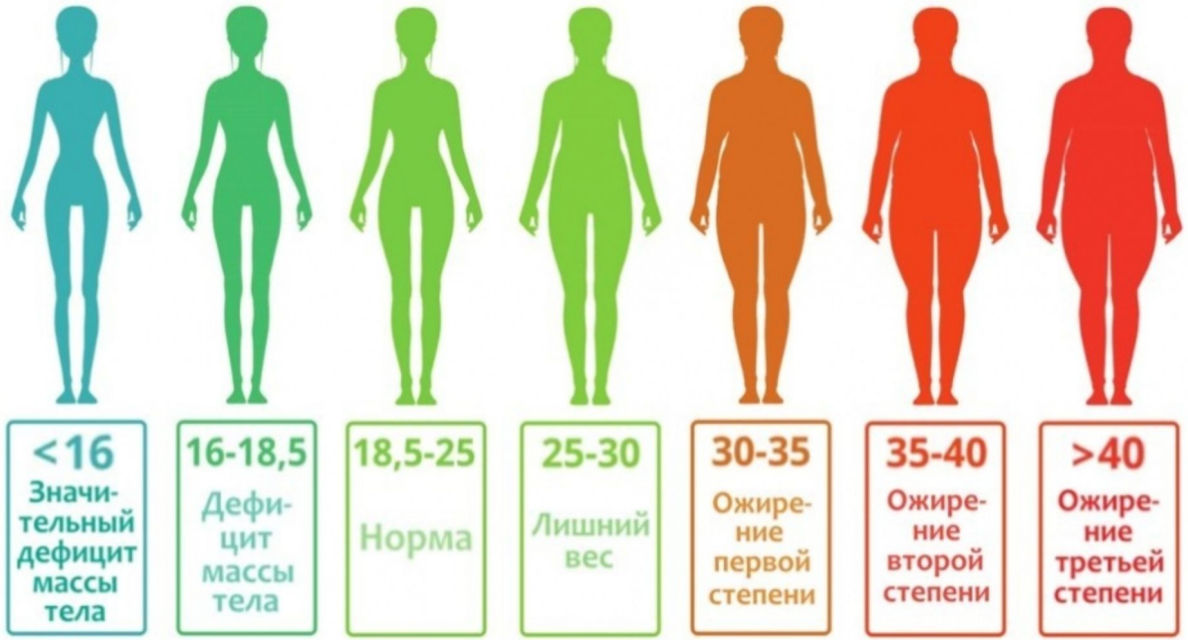
Database: Open Database,
Contents: Database Contents

Гипотезы

Индекс массы тела влияет на цену страховки

Индекс массы тела дает представление о теле: насколько его вес высок или низок по отношению к росту.

Объективный индекс массы тела (кг/м^2) в идеале от 18,5 до 24,9.

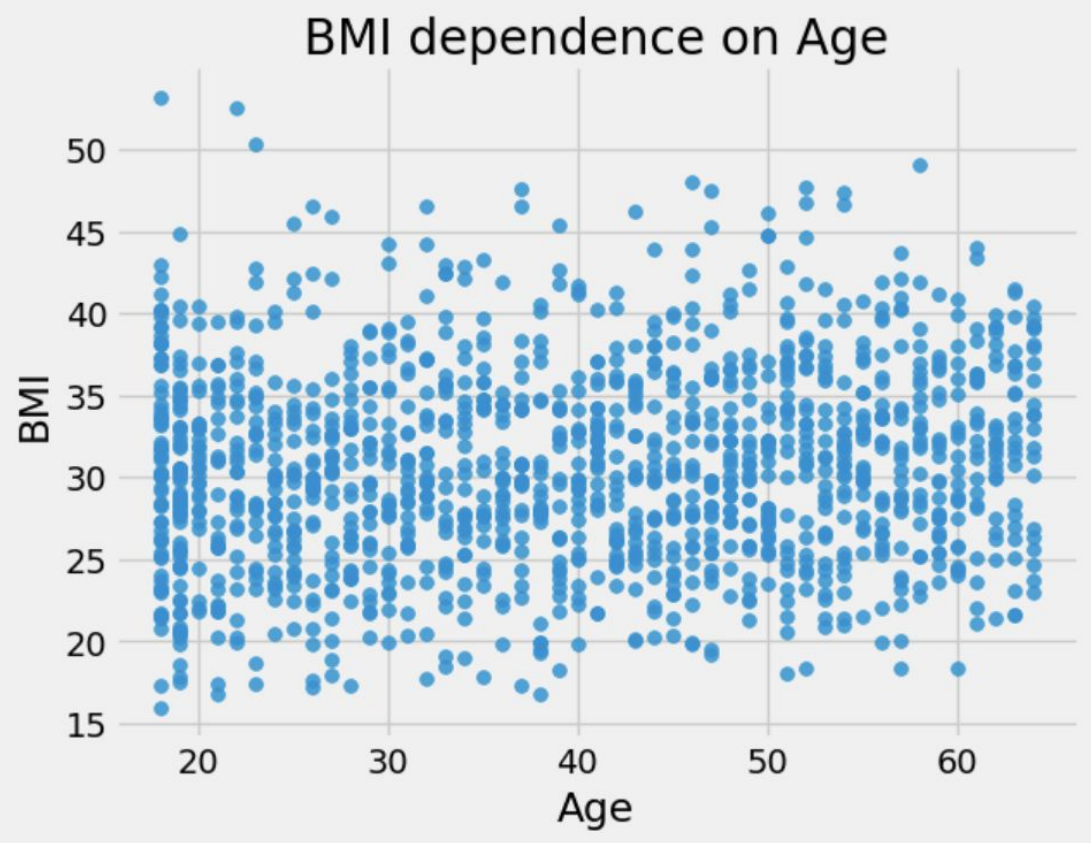


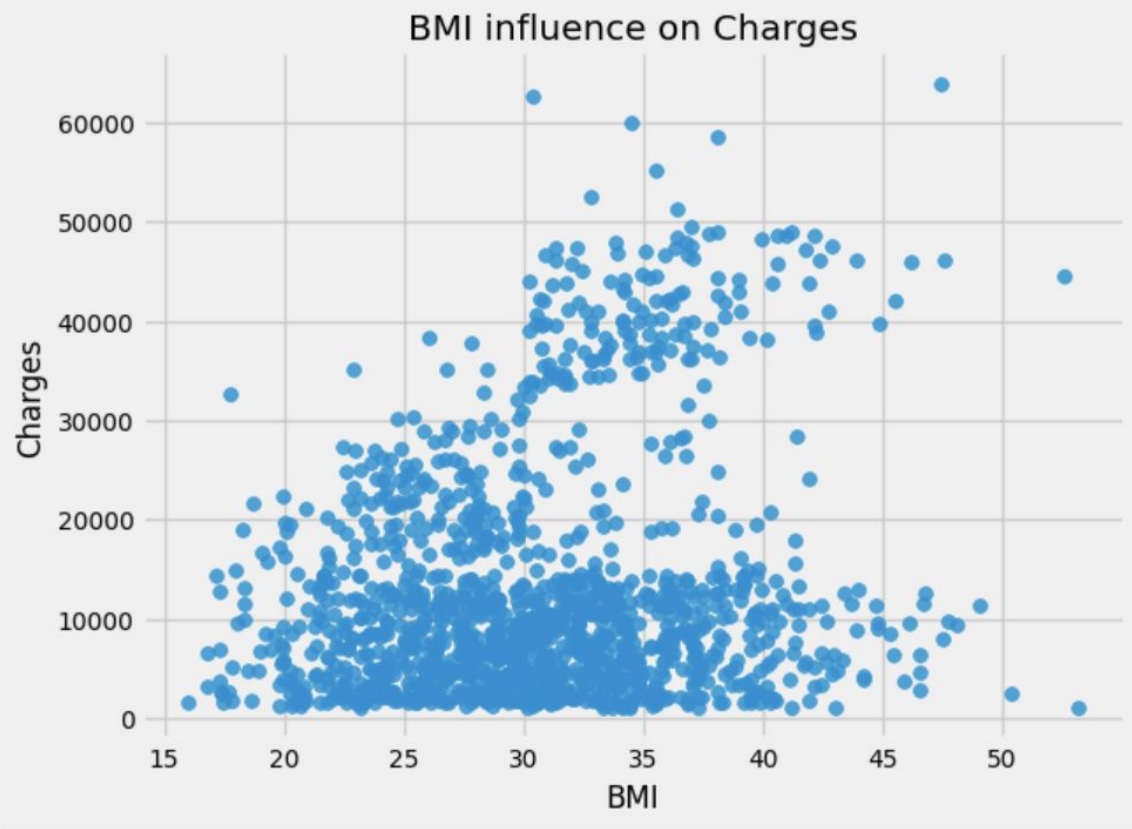
Данные

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

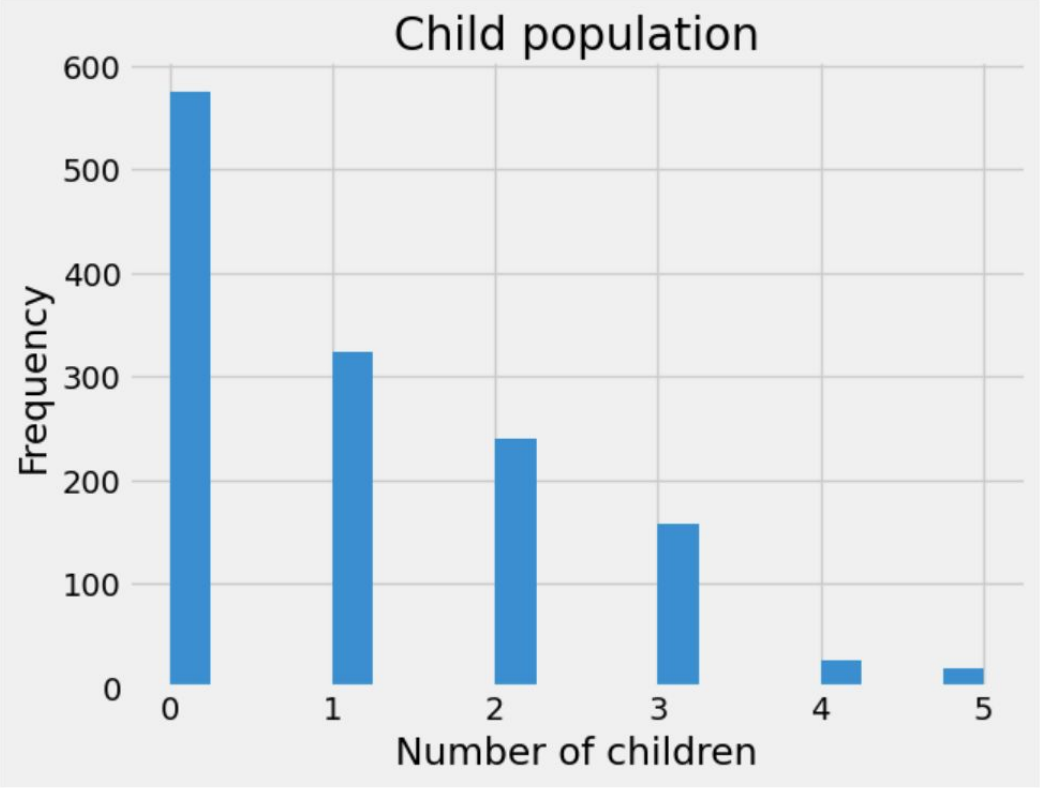
	sex	smoker	region
count	1338	1338	1338
unique	2	2	4
top	male	no	southeast
freq	676	1064	364

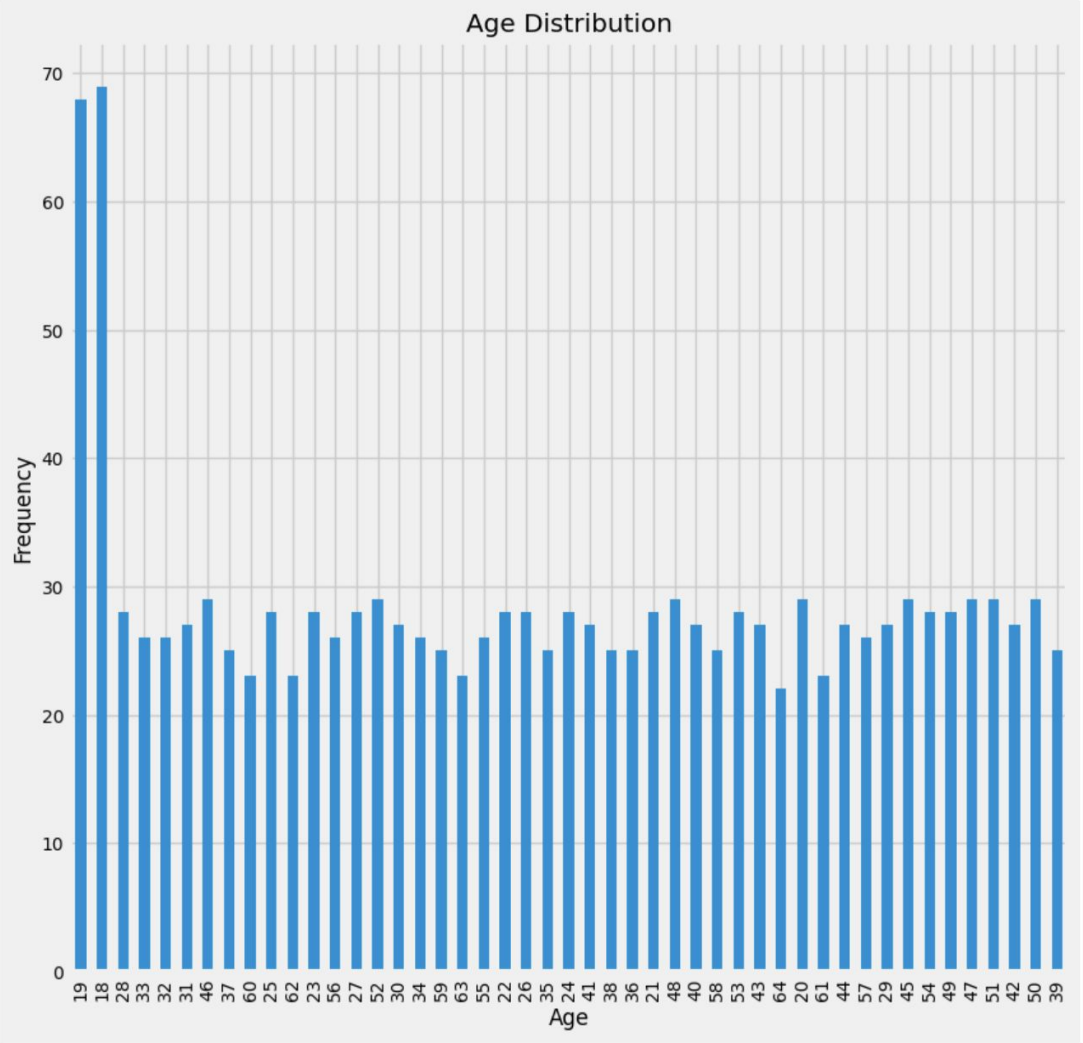
Данные


BMI dependence on Age
BMI
Age
50
45
40
35
30
25
20
15
20
30
40
50
60


BMI influence on Charges
Charges
BMI
60000
50000
40000
30000
20000
10000
0
15
20
25
30
35
40
45
50

Данные


Child population
Frequency
600
500
400
300
200
100
0
0
1
2
3
4
5
Number of children


Age Distribution
Frequency
70
60
50
40
30
20
10
0
19 18 28 33 32 31 46 37 60 25 62 23 56 27 52 30 34 59 63 55 22 26 35 24 41 38 36 21 48 40 58 53 43 64 20 61 44 57 29 45 54 49 47 51 42 50 39
Age

Данные


60000
50000
40000
30000
20000
10000
0
Charges
Southwest
Southeast
Northwest
Northeast
Region
sex
female
male

Интерпретация и значимость оцененных коэффициентов регрессии

	Коэффициент	P-value
const	-11938.538576	0.000
age	256.856353	0.000
bmi	339.193454	0.000
children	475.500545	0.001
sex_male	-131.314359	0.693
smoker_yes	23848.534542	0..000
region_northwest	-352.963899	0.459
region_southeast	-1035.022049	0.031
region_southwest	-960.050991	0.045

Метрики качества построенной модели регрессии

OLS Regression Results

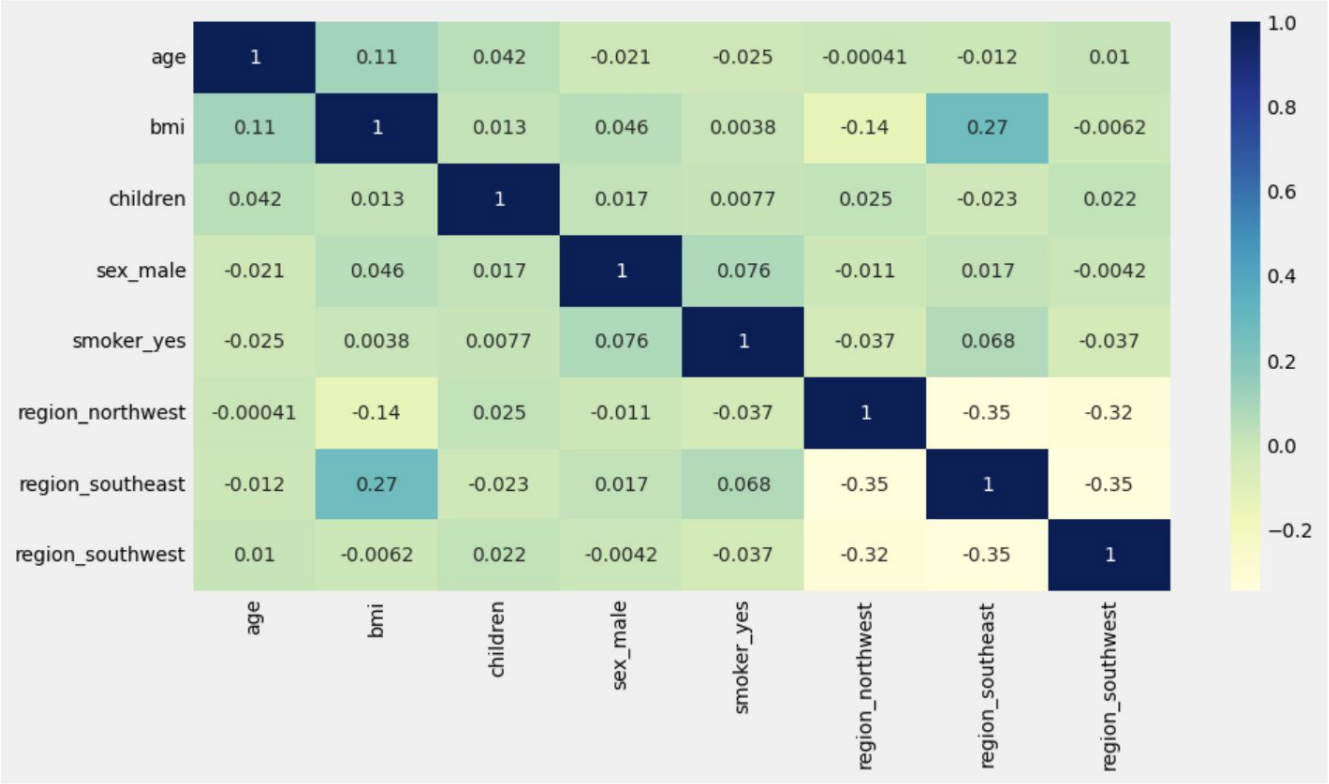
Dep. Variable:	charges	R-squared:	0.751
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	500.8
Date:	Mon, 04 Dec 2023	Prob (F-statistic):	0.00
Time:	03:49:00	Log-Likelihood:	-13548.
No. Observations:	1338	AIC:	2.711e+04
Df Residuals:	1329	BIC:	2.716e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.194e+04	987.819	-12.086	0.000	-1.39e+04	-1e+04
age	256.8564	11.899	21.587	0.000	233.514	280.199
bmi	339.1935	28.599	11.860	0.000	283.088	395.298
children	475.5005	137.804	3.451	0.001	205.163	745.838
sex_male	-131.3144	332.945	-0.394	0.693	-784.470	521.842
smoker_yes	2.385e+04	413.153	57.723	0.000	2.3e+04	2.47e+04
region_northwest	-352.9639	476.276	-0.741	0.459	-1287.298	581.370
region_southeast	-1035.0220	478.692	-2.162	0.031	-1974.097	-95.947
region_southwest	-960.0510	477.933	-2.009	0.045	-1897.636	-22.466

Omnibus:	300.366	Durbin-Watson:	2.088
Prob(Omnibus):	0.000	Jarque-Bera (JB):	718.887
Skew:	1.211	Prob(JB):	7.86e-157
Kurtosis:	5.651	Cond. No.	311.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

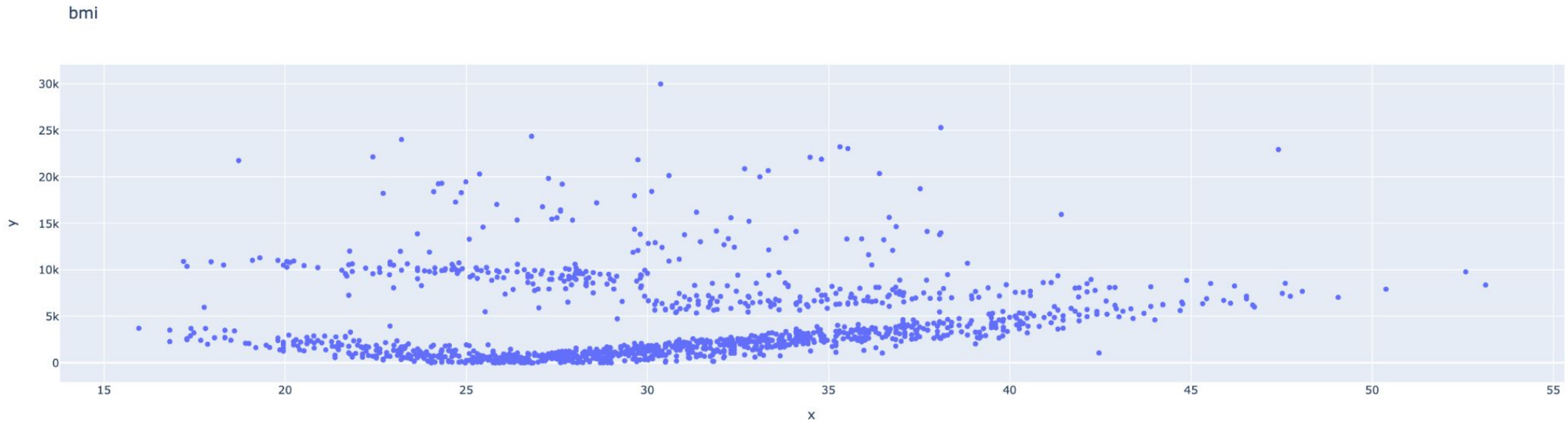
Мультиколлинеарность



age	bmi	children	sex_male	smoker_yes	region_northwest	region_southeast	region_southwest
1.016822	1.10663	1.004011	1.0089	1.012074	1.518823	1.65223	1.529411

Гетероскедастичность

lm	lm_pvalue	fvalue	f_pvalue
121.743601	1.446718e-22	16.628612	1.145606e-23



Автокорреляция ошибок

lmfloat	lmpvalfloat	fvalfloat	lmfloat
12.000303	0.285036	1.193696	0.290497

0.285036>0.05 ⇒ не отвергаем нулевую гипотезу об отсутствии автокорреляции.

Другие наборы переменных

OLS Regression Results

Dep. Variable:	charges	**R-squared:**	0.754
Model:	OLS	**Adj. R-squared:**	0.752
Method:	Least Squares	**F-statistic:**	406.7
Date:	Mon, 04 Dec 2023	**Prob (F-statistic):**	0.00
Time:	10:59:09	**Log-Likelihood:**	-13539.
No. Observations:	1338	**AIC:**	2.710e+04
Df Residuals:	1327	**BIC:**	2.716e+04
Df Model:	10		
Covariance Type:	nonrobust		

логарифмированные признаки (age, bmi)

OLS Regression Results

Dep. Variable:	charges	**R-squared:**	0.858
Model:	OLS	**Adj. R-squared:**	0.846
Method:	Least Squares	**F-statistic:**	74.47
Date:	Mon, 04 Dec 2023	**Prob (F-statistic):**	0.00
Time:	10:59:09	**Log-Likelihood:**	-13174.
No. Observations:	1338	**AIC:**	2.655e+04
Df Residuals:	1237	**BIC:**	2.707e+04
Df Model:	100		
Covariance Type:	nonrobust		

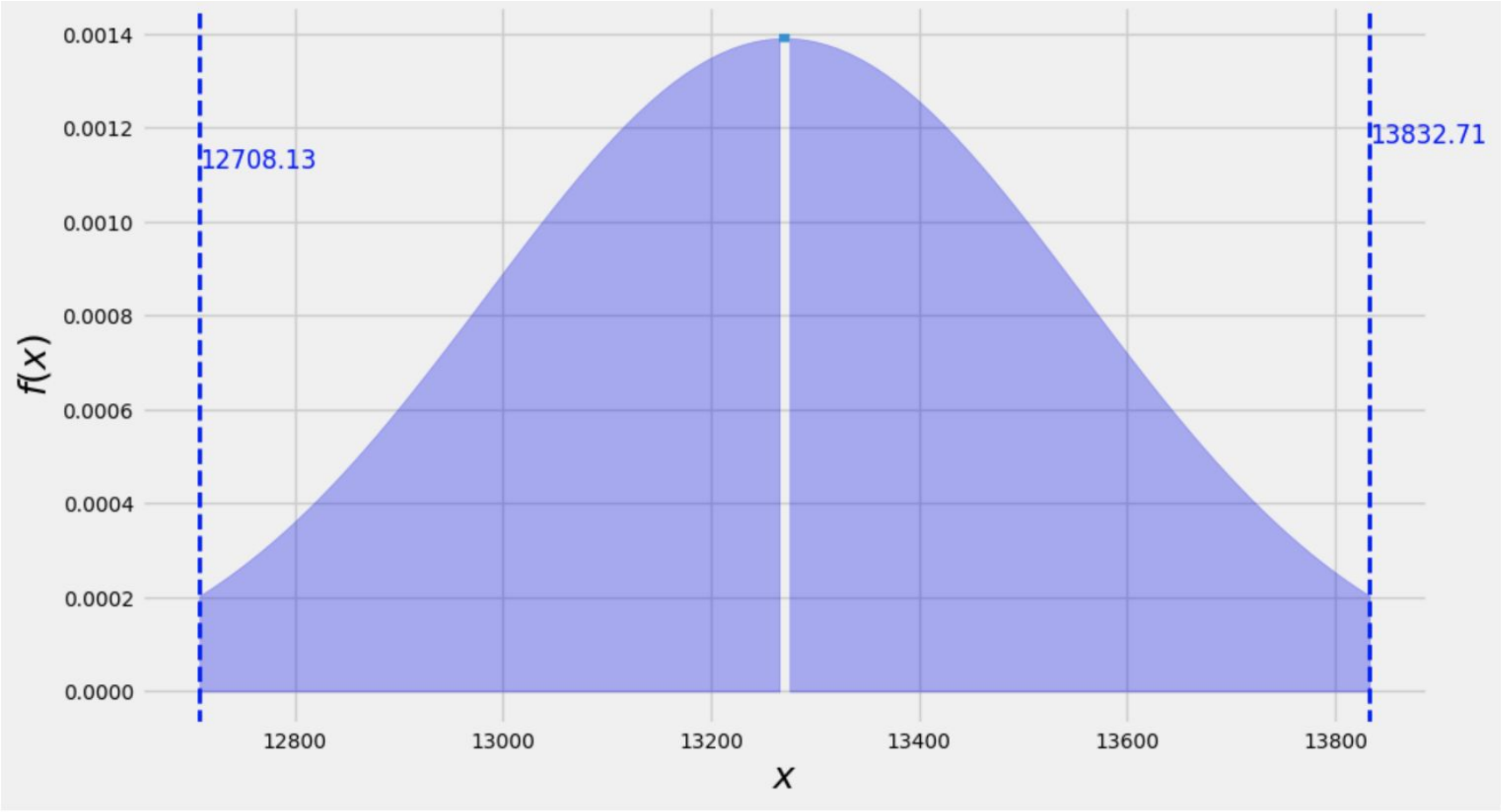
полиномиальные признаки с возведением в степень

OLS Regression Results

Dep. Variable:	charges	**R-squared:**	0.845
Model:	OLS	**Adj. R-squared:**	0.841
Method:	Least Squares	**F-statistic:**	214.7
Date:	Mon, 04 Dec 2023	**Prob (F-statistic):**	0.00
Time:	10:59:10	**Log-Likelihood:**	-13232.
No. Observations:	1338	**AIC:**	2.653e+04
Df Residuals:	1304	**BIC:**	2.671e+04
Df Model:	33		
Covariance Type:	nonrobust		

полиномиальные признаки - попарные произведения

Доверительный интервал



95% асимптотически-нормальный доверительный интервал для предсказания

Заключение:

Нулевая гипотеза о влиянии ИМТ на цену страховки **не отвергается**

Спасибо за внимание!