Министерство образования и науки

федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский университет ИТМО»

Факультет инфокоммуникационных технологий

Отчет по дисциплине: **«Современные инструменты анализа данных»**

Лабораторная работа 3

Выполнил:

–

Проверила:

–

Санкт-Петербург

2023

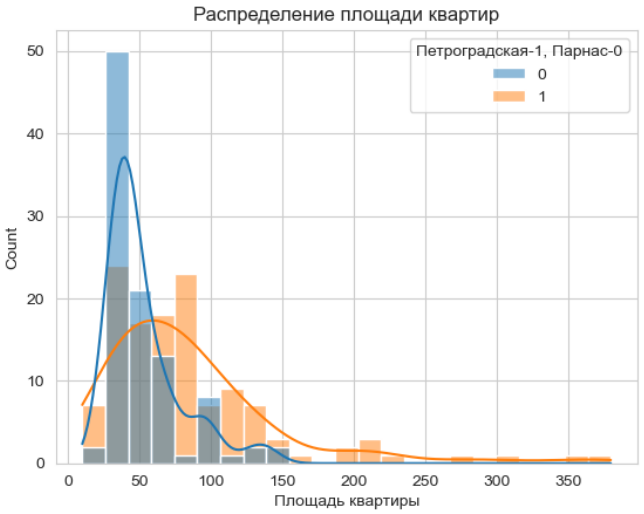
Работа делалась с использованием **python** и последующих модулей для работы с данными и их визуализации: **pandas, numpy, seaborn, matplotlib, statsmodels, sklearn**

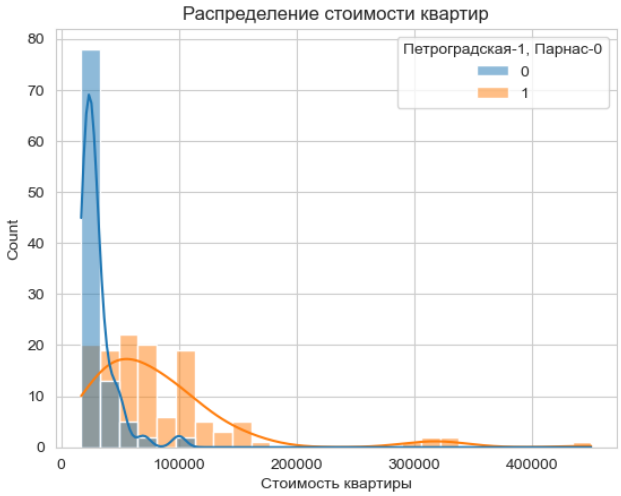
Часть 1. Анализ многомерных данных. Парная и множественная регрессия

1. Указать описательные статистики, графики плотности и гистограммы для каждого района

	Парнас - стоимость кв	Парнас - площадь кв	Петроградская - стоимость кв	Петроградская - площадь кв
count	126.000000	126.000000	100.000000	100.0000
mean	83115.079365	84.742063	29880.000000	52.3300
std	68314.044316	62.512390	14634.282429	25.6598
min	18000.000000	10.000000	17000.000000	23.0000
25%	45000.000000	43.000000	22000.000000	36.0000
50%	70000.000000	70.000000	25000.000000	42.0000
75%	100000.000000	104.500000	32000.000000	60.0000
max	450000.000000	380.000000	100000.000000	140.0000

Статистическое описание данных в зависимости от района





2. Построить модель **парной регрессии** для квартир площадью от 20 до 110 кв.м включительно стоимости от площади. Оценить характеристики построенной модели.

	Стоимость квартиры	Площадь квартиры	Петроградская-1, Парнас-0
Стоимость квартиры	1.000000	0.887266	0.455191
Площадь квартиры	0.887266	1.000000	0.309301
Петроградская-1, Парнас-0	0.455191	0.309301	1.000000

Матрица корреляции признаков

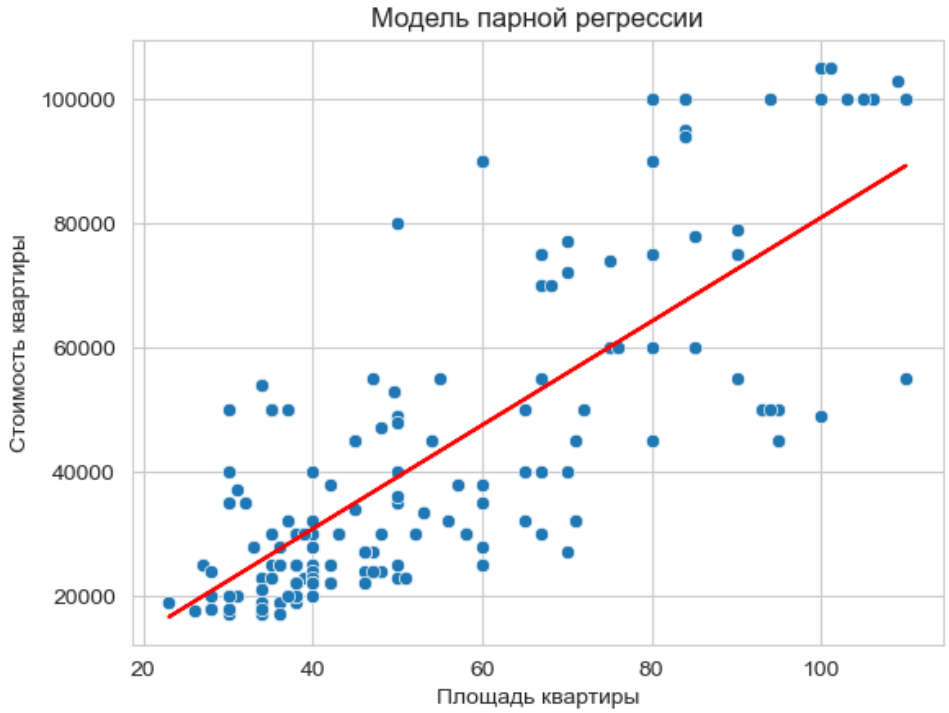
OLS Regression Results

Dep. Variable:	Стоимость квартиры	R-squared:	0.591
Model:	OLS	Adj. R-squared:	0.589
Method:	Least Squares	F-statistic:	269.9
Date:	Wed, 11 Oct 2023	Prob (F-statistic):	4.01e-38
Time:	18:58:08	Log-Likelihood:	-2097.0
No. Observations:	189	AIC:	4198.
Df Residuals:	187	BIC:	4204.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2637.4656	3091.642	-0.853	0.395	-8736.443	3461.512
Площадь квартиры	835.3907	50.851	16.428	0.000	735.075	935.706

Omnibus:	4.598	Durbin-Watson:	0.691
Prob(Omnibus):	0.100	Jarque-Bera (JB):	4.660
Skew:	0.378	Prob(JB):	0.0973
Kurtosis:	2.863	Cond. No.	161.

Результаты построенной парной линейной регрессии



3. Построить модель **множественной регрессии** для квартир площадью от 20 до 110 кв.м включительно стоимости от площади и района.Оценить характеристики построенной модели.

OLS Regression Results

Dep. Variable:	Стоимость квартиры	R-squared:	0.801
Model:	OLS	Adj. R-squared:	0.799
Method:	Least Squares	F-statistic:	374.7
Date:	Wed, 11 Oct 2023	Prob (F-statistic):	5.80e-66
Time:	18:58:12	Log-Likelihood:	-2028.8
No. Observations:	189	AIC:	4064.
Df Residuals:	186	BIC:	4073.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4750.7798	2165.986	-2.193	0.030	-9023.838	-477.722
Площадь квартиры	661.4266	37.641	17.572	0.000	587.169	735.684
Петроградская-1, Парнас-0	2.42e+04	1725.202	14.030	0.000	2.08e+04	2.76e+04

Omnibus:	3.103	Durbin-Watson:	1.088
Prob(Omnibus):	0.212	Jarque-Bera (JB):	3.437
Skew:	0.021	Prob(JB):	0.179
Kurtosis:	3.659	Cond. No.	162.

Результаты построенной множественной линейной регрессии

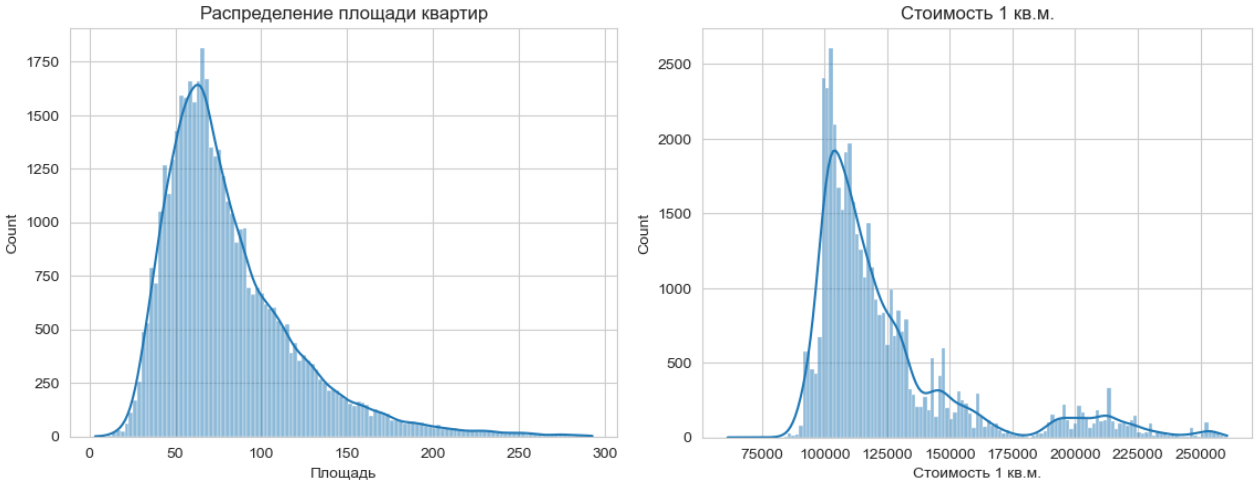
Часть 2. Анализ многомерных данных. Множественная регрессия

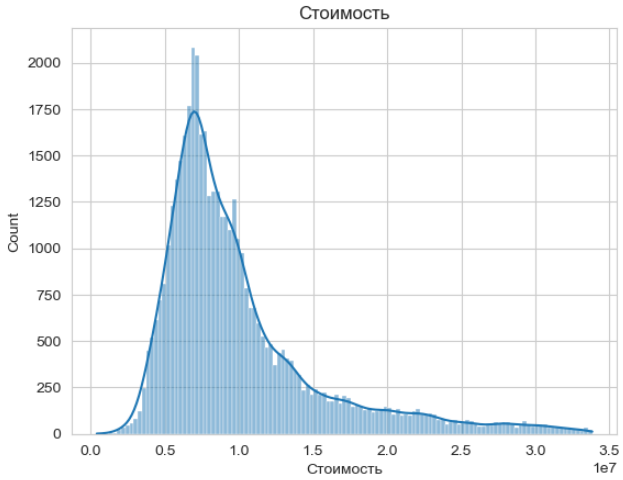
1. Скорректировать выборку случайным образом, чтобы осталось минимум 5000 записей.

Из выборки были удалены выбросы по всем непрерывным признакам с использованием интервала [mean – 3* std; mean + 3 * std] и выбраны случайным образом 5000 записей.

	Площадь	Стоимость 1 кв.м.	Стоимость
count	44035.000000	44035.000000	4.403500e+04
mean	81.370508	124625.932105	1.004983e+07
std	38.888211	32352.222625	5.480365e+06
min	3.300000	61266.670000	4.095822e+05
25%	54.900000	103689.685000	6.588206e+06
50%	71.600000	113406.300000	8.395278e+06
75%	98.400000	130333.160000	1.150167e+07
max	292.700000	260715.100000	3.383975e+07

Статистика по непрерывным данным после преобразования





2. Построить модель множественной регрессии, где в качестве зависимой переменной выступает стоимость квартиры. Для построения модели преобразовать переменную “Тип дома” в фиктивные переменные. Коэффициенты подобрать самостоятельно. Оценить качество модели. Учесть, что коэффициенты модели должны быть статистически значимы

OLS Regression Results

Dep. Variable:	Стоимость	R-squared:	0.964
Model:	OLS	Adj. R-squared:	0.964
Method:	Least Squares	F-statistic:	8415.
Date:	Wed, 11 Oct 2023	Prob (F-statistic):	0.00
Time:	18:58:27	Log-Likelihood:	-76412.
No. Observations:	5000	AIC:	1.529e+05
Df Residuals:	4983	BIC:	1.530e+05
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.362e+06	4.74e+05	9.211	0.000	3.43e+06	5.29e+06
Площадь	1.099e+05	499.032	220.260	0.000	1.09e+05	1.11e+05
Стоимость 1 кв.м.	13.6958	2.575	5.318	0.000	8.647	18.744
Этаж	5.168e+04	3.35e+04	1.545	0.122	-1.39e+04	1.17e+05
Зона	-1.236e+06	4.4e+04	-28.073	0.000	-1.32e+06	-1.15e+06
Тип дома_БИЗНЕС	6.461e+06	2.38e+05	27.186	0.000	6e+06	6.93e+06
Тип дома_ИНД	-7.113e+05	1.41e+05	-5.033	0.000	-9.88e+05	-4.34e+05
Тип дома_КИРП_5эт	-1.11e+06	1.34e+05	-8.288	0.000	-1.37e+06	-8.47e+05
Тип дома_КИРП_ТИП	-8.356e+05	1.04e+05	-8.006	0.000	-1.04e+06	-6.31e+05
Тип дома_КИРП_совр	1.844e+05	8.31e+04	2.221	0.026	2.16e+04	3.47e+05
Тип дома_КОНСТР	-1.548e+06	1.19e+05	-13.058	0.000	-1.78e+06	-1.32e+06
Тип дома_МАЛЭТ	-5.244e+05	4.45e+05	-1.179	0.239	-1.4e+06	3.48e+05
Тип дома_ПАН_5эт	-2.01e+06	3.32e+05	-6.055	0.000	-2.66e+06	-1.36e+06
Тип дома_ПАН_совр	-3.316e+05	3.73e+05	-0.888	0.375	-1.06e+06	4e+05
Тип дома_СТАЛ	-4.566e+05	7.11e+04	-6.422	0.000	-5.96e+05	-3.17e+05
Тип дома_СФ	-1.658e+06	7.17e+04	-23.109	0.000	-1.8e+06	-1.52e+06
Тип дома_СФ_КР	-1.019e+06	7.04e+04	-14.472	0.000	-1.16e+06	-8.81e+05
Тип дома_ЭЛИТА	7.921e+06	3.14e+05	25.208	0.000	7.3e+06	8.54e+06

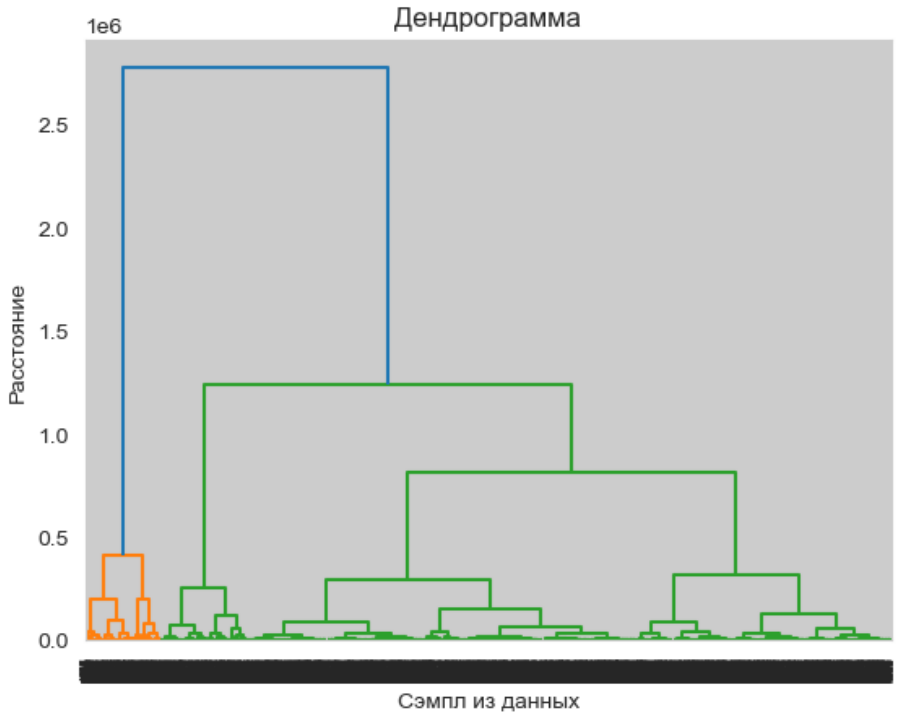
Omnibus:	760.778	Durbin-Watson:	2.046
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12418.833
Skew:	0.095	Prob(JB):	0.00
Kurtosis:	10.718	Cond. No.	2.86e+20

Выше представлены результаты множественной линейной регрессии. Категориальные признаки были закодированы с использованием one-hot encoding.

Часть 3. Анализ многомерных данных. Кластеризация

Для выбранных данных провести кластеризацию каждым из методов. Переменные подобрать самостоятельно из количественных. Оценить характеристики построенной модели и выводы по результатам.

- метод к-средних
- древовидная классификация



Дендограмма на основе древовидной классификации данных

Для продолжения поиска оптимального количества кластеров (хотя из полученной дендограммы можно сделать вывод о диапазоне 2-5 кластеров) была проведена оценка разбиения на разное кол-во кластером с учетом silhouette оценки.

Наилучшее значение получается на **двух кластерах.**