Галушкин Т. Р., Нурмухаметов А. Л.
Костин А. А., Мартыненко Д. Ю.

“Книжный” анализ данных

Постановка


Книжные магазины
Читай-город
Книжный лабиринт
Молодая гвардия
Исследователь + Эксперт
Исследователь + Эксперт
Исследователь + Эксперт

Сбор данных

Категории

Популярные бестселлеры

Классика

Научно-популярная литература

Параметры

Цена

Количество страниц

Визуальная привлекательность (1-5)

Качество обложки (1-5)

Качество переплёта/корешка (1-5)

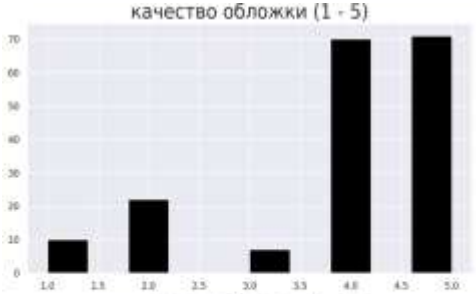
Качество печати (1-5)

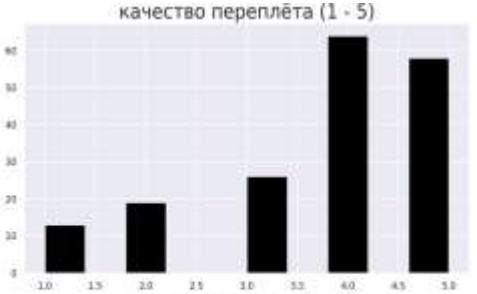
Качество бумаги (1-5)

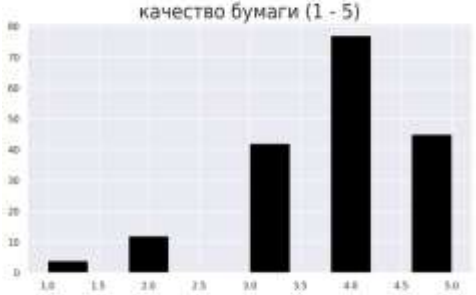
Формат книги (0-3)

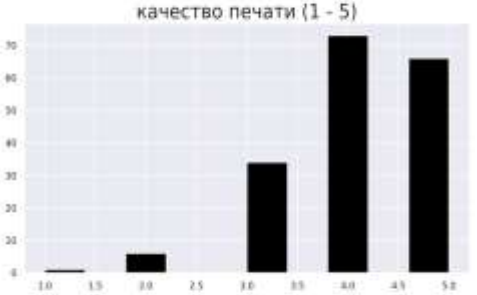
Размер книги (0-3)

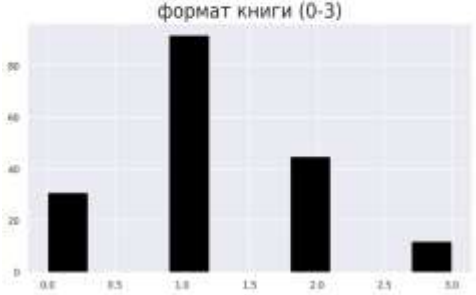
Распределение основных параметров

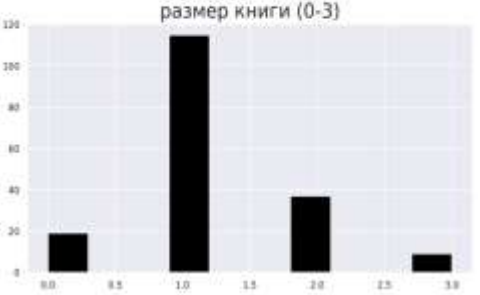




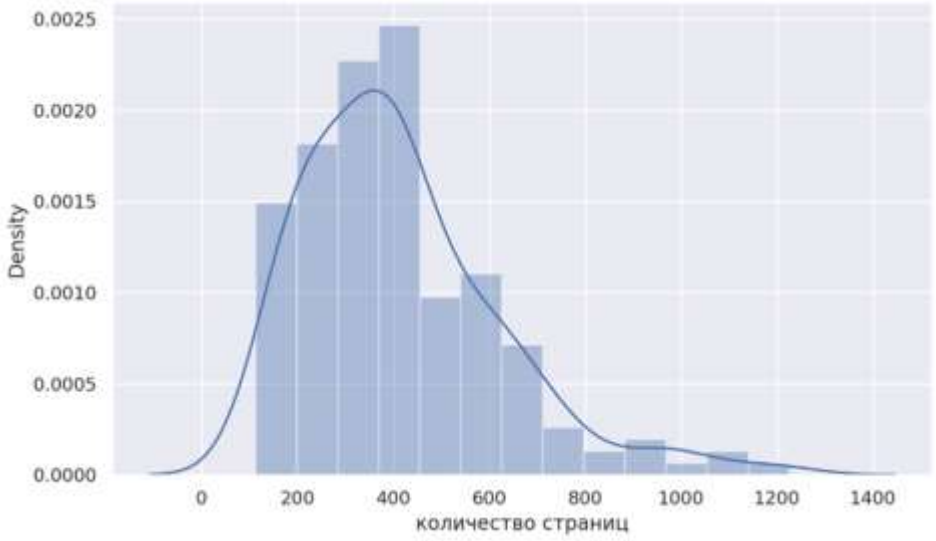


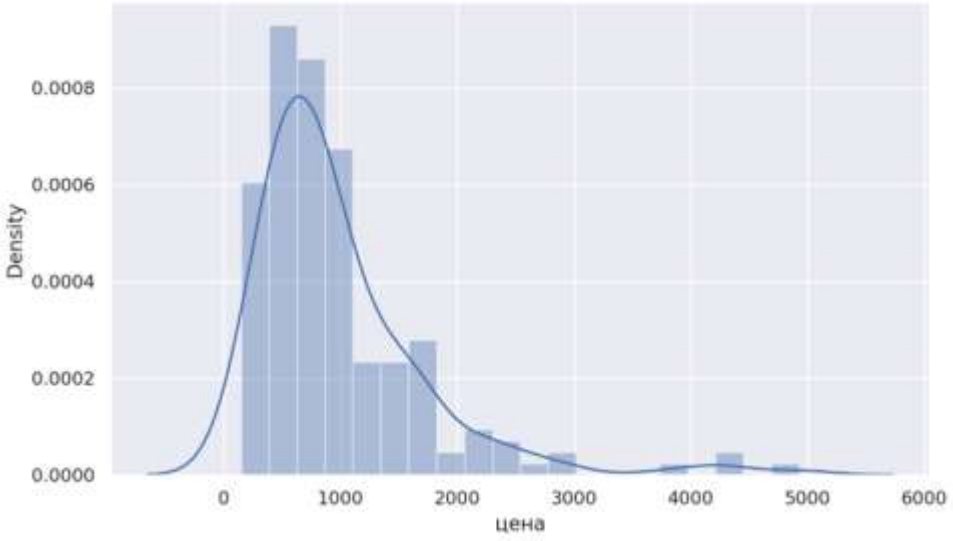




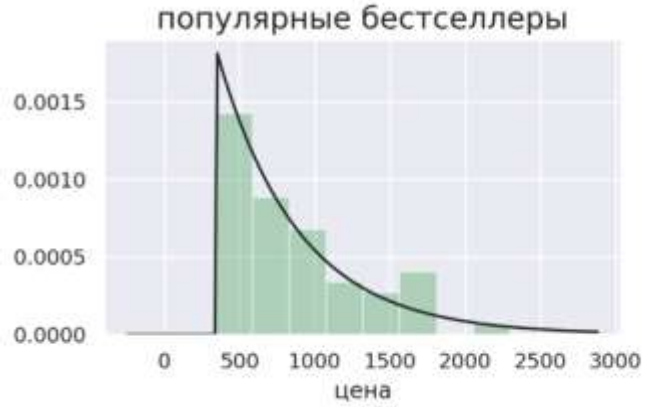


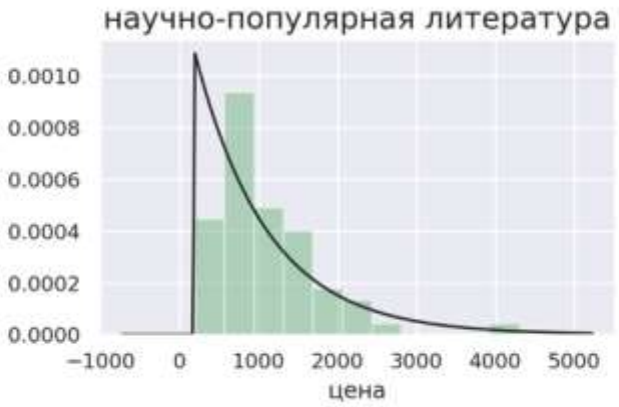
Распределение цены и количества страниц

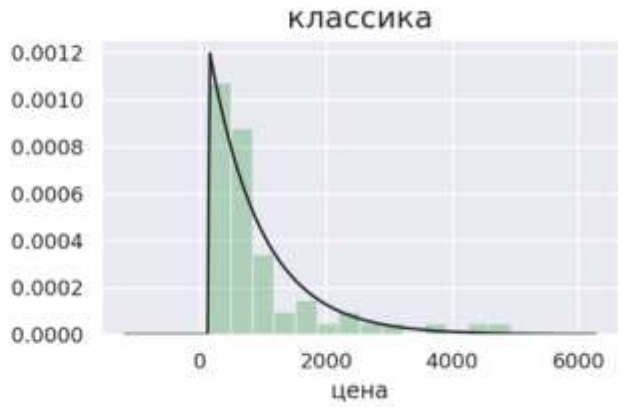




Распределение цены по категориям







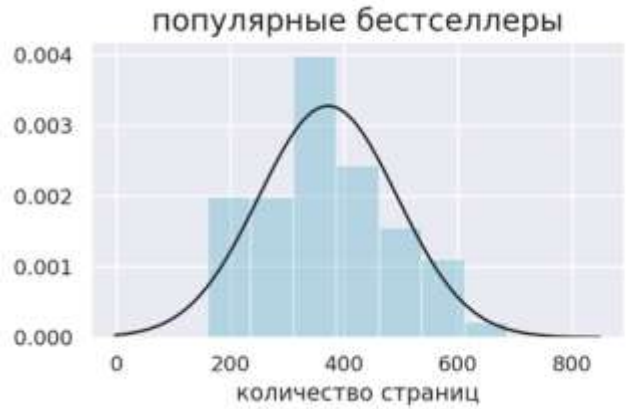
Проверка на экспоненциальность

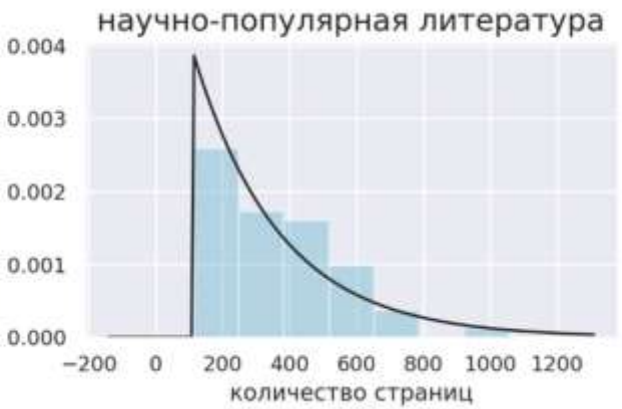
Популярные бестселлеры: statistic = 6.04, critical value = 1.32

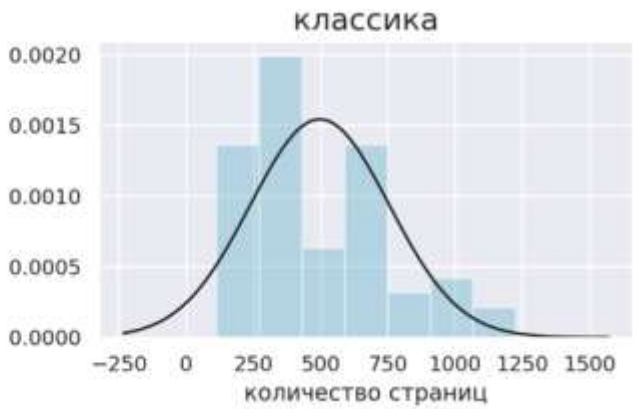
Научно-популярная литература: statistic = 4.7, critical value = 1.32

Классика: statistic = 1.61, critical value = 1.32

Распределение количества страниц по категориям







Проверка на нормальность

Популярные бестселлеры: statistic = 0.28, critical value = 0.743

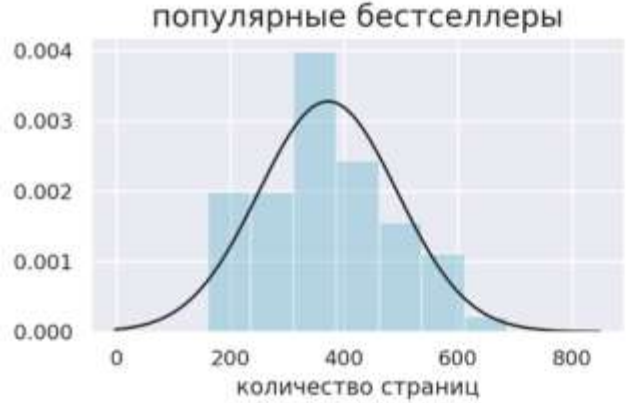
Проверка на экспоненциальность

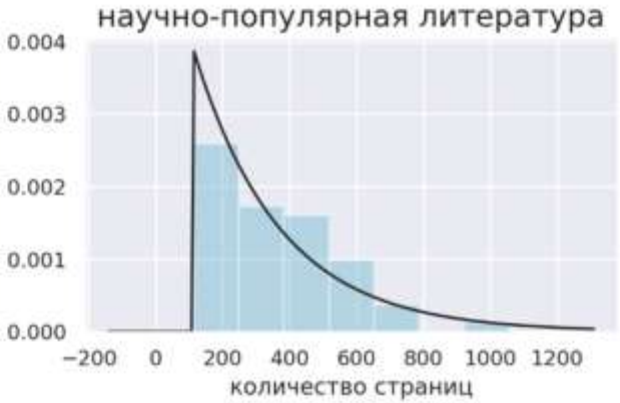
Научно-популярная литература: statistic = 7.12, critical value = 1.33

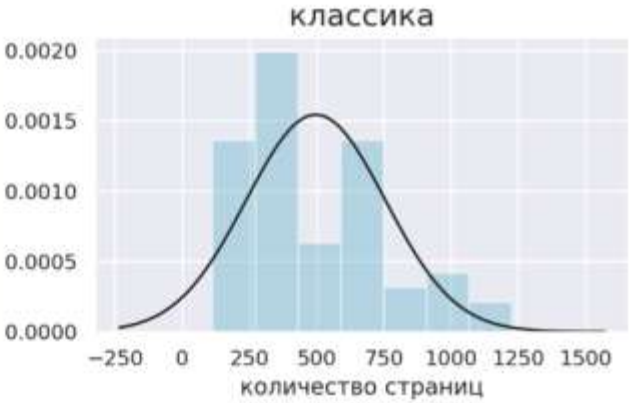
Проверка на нормальность

Классика: statistic = 1.1, critical value = 0.74

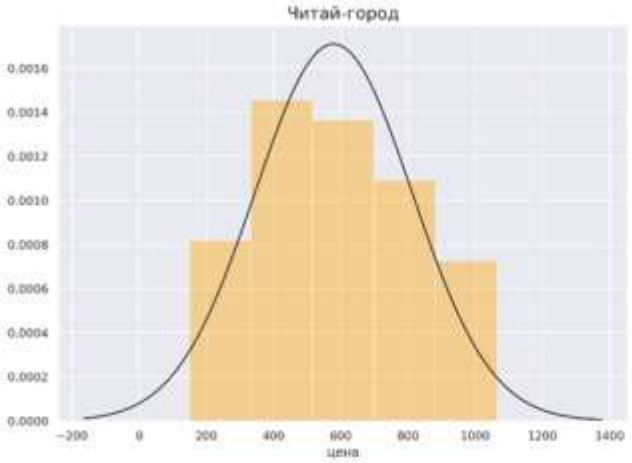
Распределение количества страниц по категориям

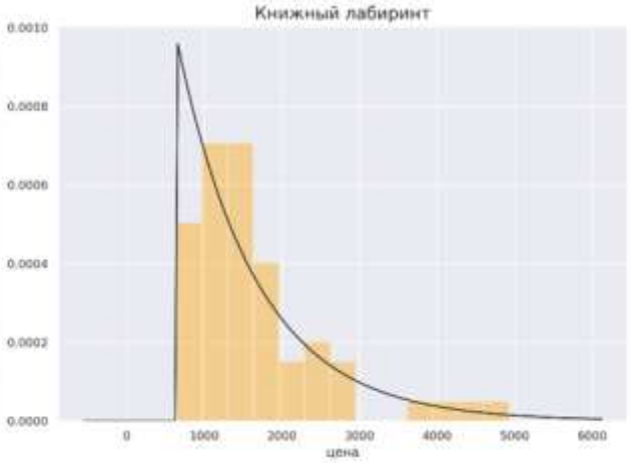


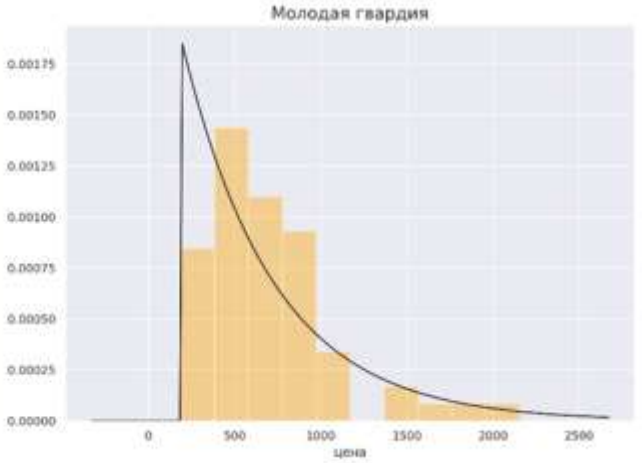




Распределение цены по магазинам







Проверка на нормальность

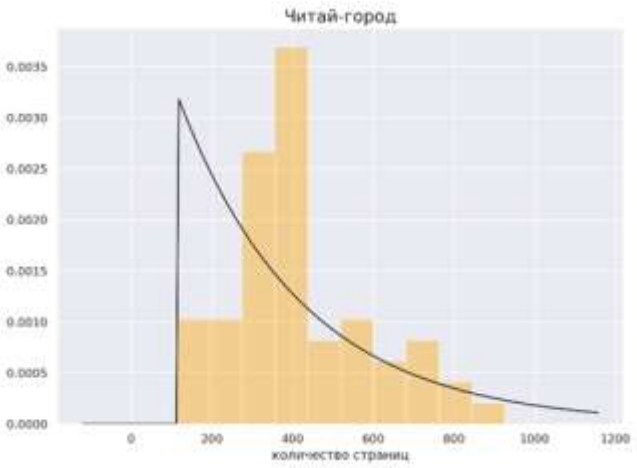
Читай-город: statistic = 0.36, critical value = 0.74

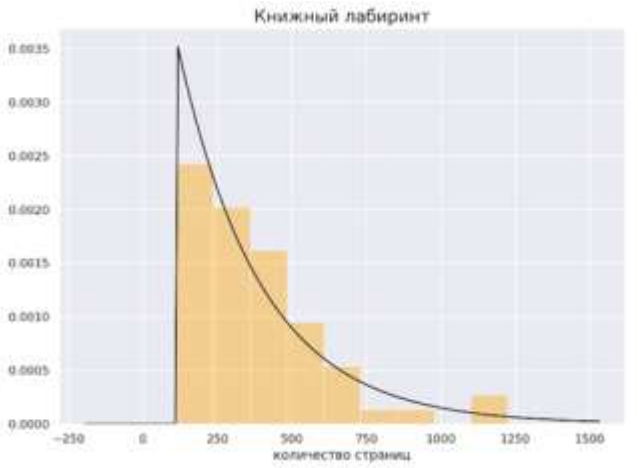
Проверка на экспоненциальность

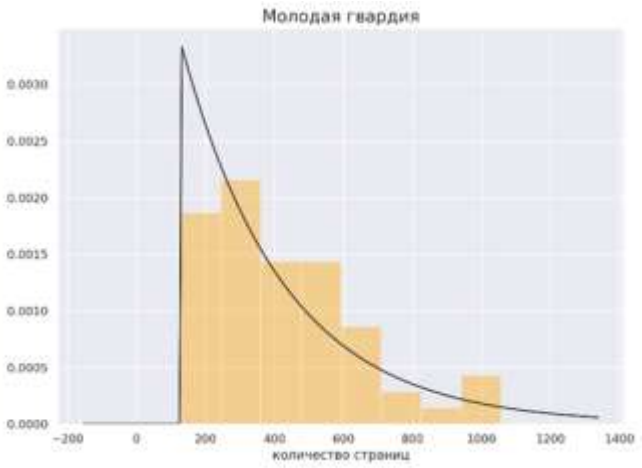
Книжный лабиринт: statistic = 9.1, critical value = 1.33

Молодая гвардия: statistic = 7.46, critical value = 1.33

Распределение количества страниц по магазинам







Проверка на нормальность:

Читай-город: statistic = 1.23, critical value = 0.74

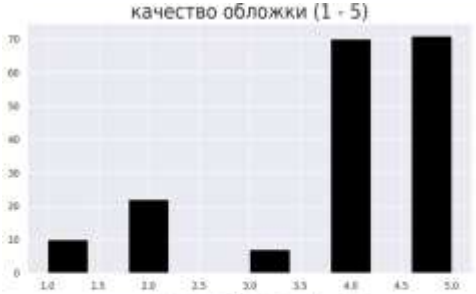
Проверка на экспоненциальность

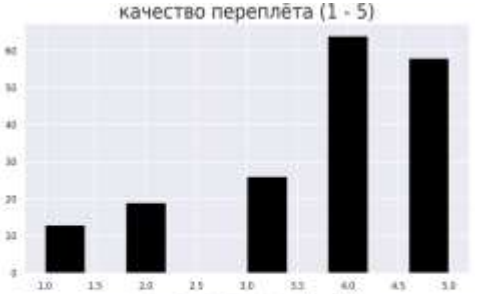
Читай-город: statistic = 10.16, critical value = 1.33

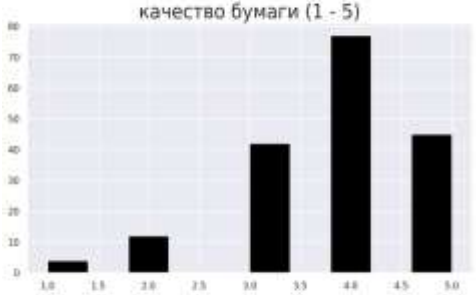
Книжный лабиринт: statistic = 6.35, critical value = 1.33

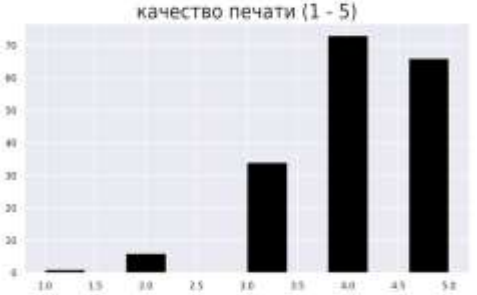
Молодая гвардия: statistic = 7.7, critical value = 1.33

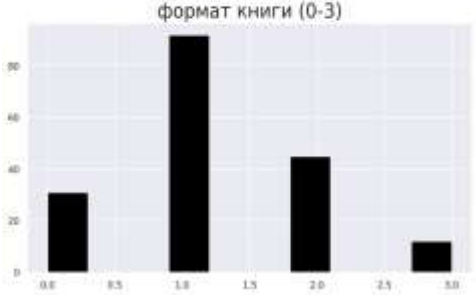
Распределение основных параметров

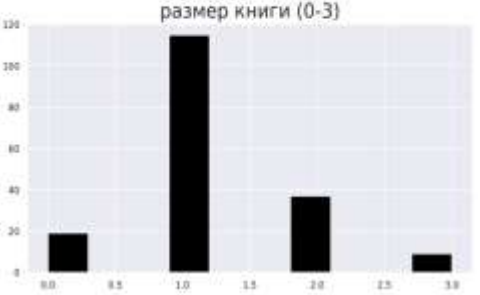








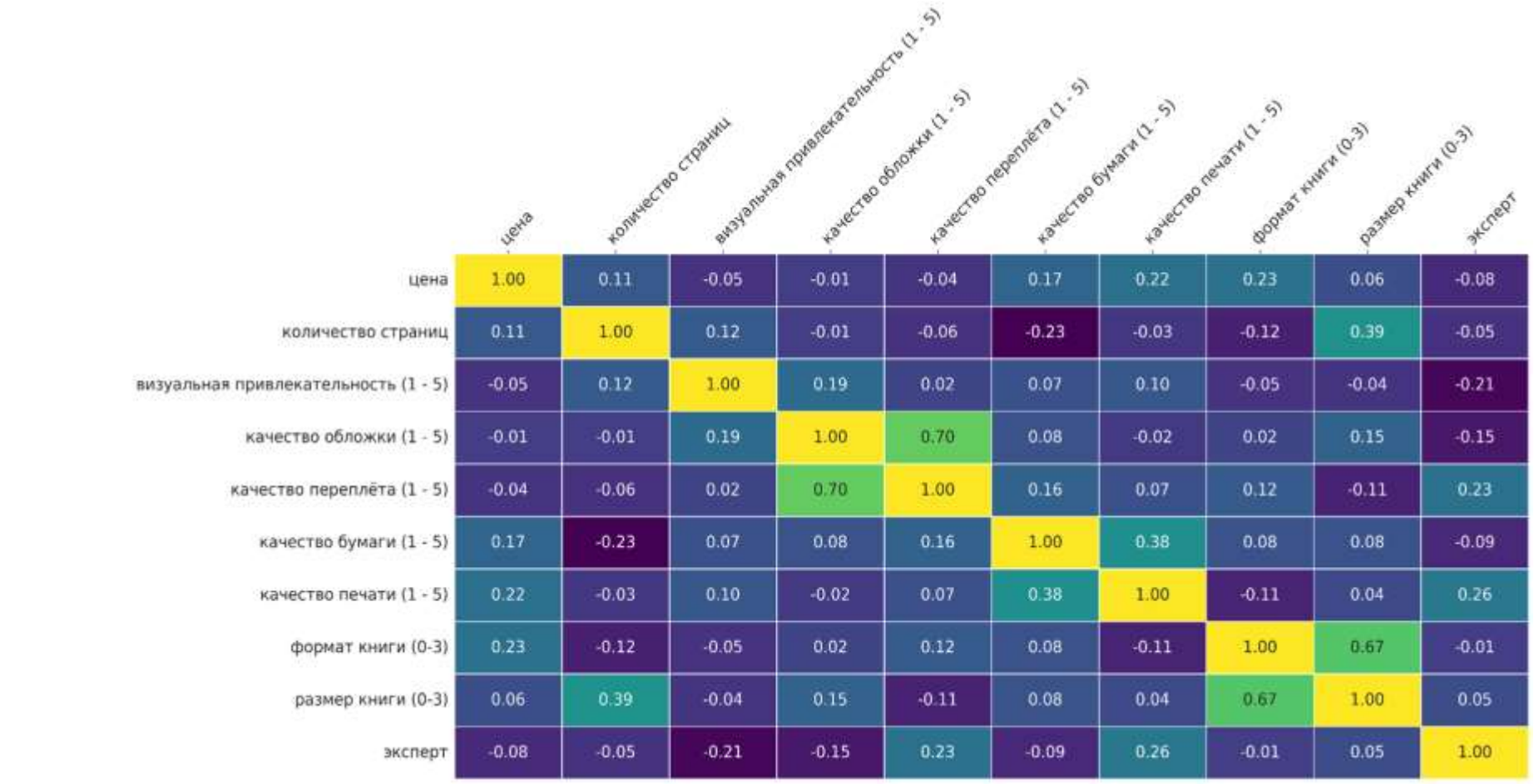




Общие корреляции

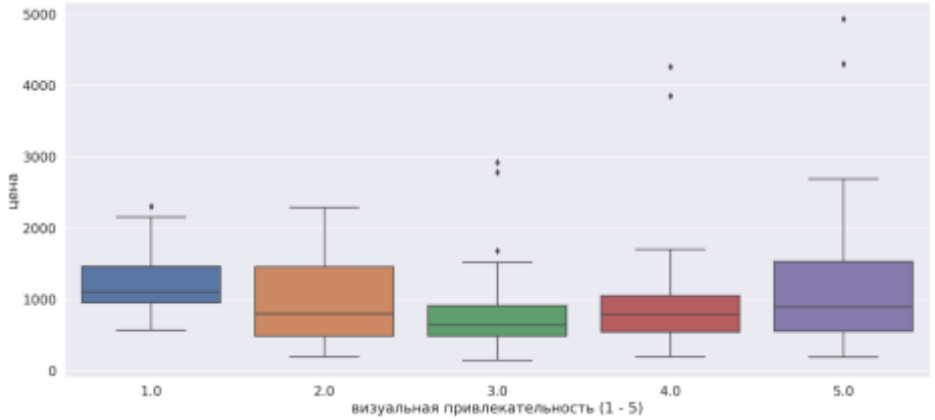
	цена	количество страниц	визуальная привлекательность (1 - 5)	качество обложки (1 - 5)	качество переплёта (1 - 5)	качество бумаги (1 - 5)	качество печати (1 - 5)	формат книги (0-3)	размер книги (0-3)	эксперт
цена	1.00	0.16	0.06	0.24	0.20	0.39	0.34	0.49	0.45	-0.01
количество страниц	0.16	1.00	0.06	-0.02	-0.12	-0.17	-0.10	0.19	0.39	-0.10
визуальная привлекательность (1 - 5)	0.06	0.06	1.00	0.34	0.26	0.21	0.15	0.06	0.06	-0.18
качество обложки (1 - 5)	0.24	-0.02	0.34	1.00	0.78	0.48	0.29	0.40	0.35	0.01
качество переплёта (1 - 5)	0.20	-0.12	0.26	0.78	1.00	0.51	0.35	0.34	0.25	0.17
качество бумаги (1 - 5)	0.39	-0.17	0.21	0.48	0.51	1.00	0.54	0.38	0.31	0.06
качество печати (1 - 5)	0.34	-0.10	0.15	0.29	0.35	0.54	1.00	0.19	0.19	0.25
формат книги (0-3)	0.49	0.19	0.06	0.40	0.34	0.38	0.19	1.00	0.77	0.03
размер книги (0-3)	0.45	0.39	0.06	0.35	0.25	0.31	0.19	0.77	1.00	0.02
эксперт	-0.01	-0.10	-0.18	0.01	0.17	0.06	0.25	0.03	0.02	1.00

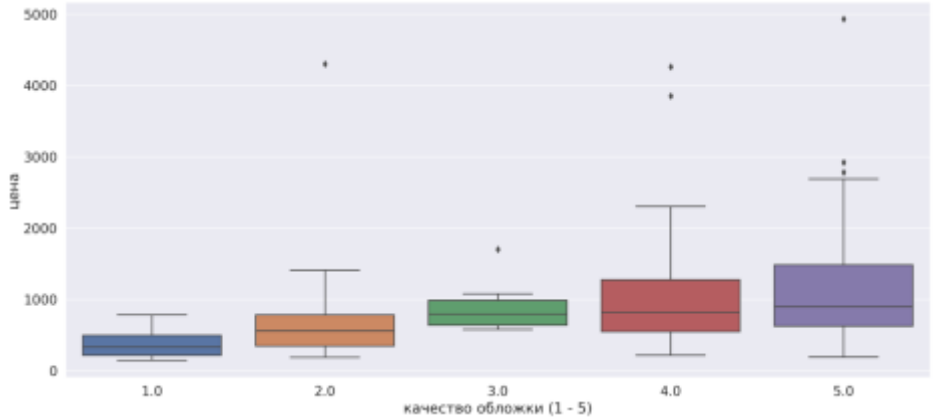
Частные корреляции

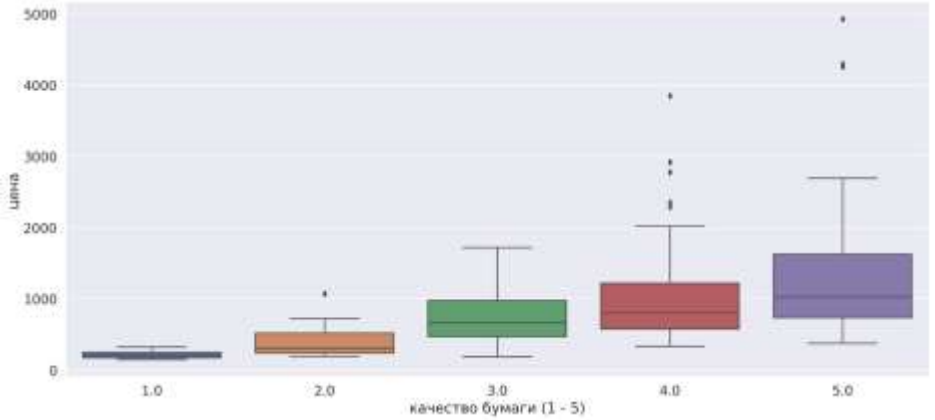


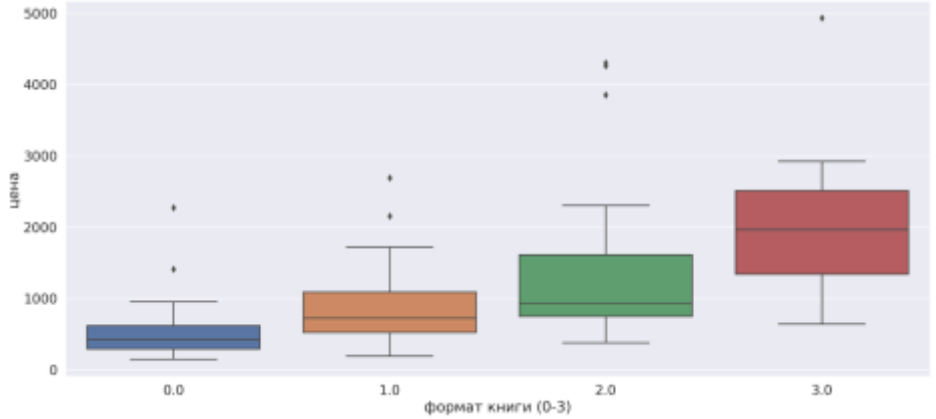
	цена	количество страниц	визуальная привлекательность (1 - 5)	качество обложки (1 - 5)	качество переплёта (1 - 5)	качество бумаги (1 - 5)	качество печати (1 - 5)	формат книги (0-3)	размер книги (0-3)	эксперт
цена	1.00	0.11	-0.05	-0.01	-0.04	0.17	0.22	0.23	0.06	-0.08
количество страниц	0.11	1.00	0.12	-0.01	-0.06	-0.23	-0.03	-0.12	0.39	-0.05
визуальная привлекательность (1 - 5)	-0.05	0.12	1.00	0.19	0.02	0.07	0.10	-0.05	-0.04	-0.21
качество обложки (1 - 5)	-0.01	-0.01	0.19	1.00	0.70	0.08	-0.02	0.02	0.15	-0.15
качество переплёта (1 - 5)	-0.04	-0.06	0.02	0.70	1.00	0.16	0.07	0.12	-0.11	0.23
качество бумаги (1 - 5)	0.17	-0.23	0.07	0.08	0.16	1.00	0.38	0.08	0.08	-0.09
качество печати (1 - 5)	0.22	-0.03	0.10	-0.02	0.07	0.38	1.00	-0.11	0.04	0.26
формат книги (0-3)	0.23	-0.12	-0.05	0.02	0.12	0.08	-0.11	1.00	0.67	-0.01
размер книги (0-3)	0.06	0.39	-0.04	0.15	-0.11	0.08	0.04	0.67	1.00	0.05
эксперт	-0.08	-0.05	-0.21	-0.15	0.23	-0.09	0.26	-0.01	0.05	1.00

Относительные показатели цены по категориям









Двухфакторный анализ цены

	sum_sq	df	F	PR(>F)
C(shop)	41981804.8	2.0	67.923662	1.946737e-22
C(book_type)	1652733.1	2.0	2.674008	7.185930e-02
C(shop):C(book_type)	5285373.4	4.0	4.275685	2.532652e-03
Residual	52845270.9	171.0	NaN	NaN

По факторам: “Магазин” и “Категория книг”

Двухфакторный анализ цены

	sum_sq	df	F	PR(>F)
C(book_type)	7.869265e+05	2.0	0.959391	3.852028e-01
C(book_format)	4.517368e+06	3.0	3.671603	1.345107e-02
C(book_format):C(book_type)	2.628518e+07	6.0	10.681962	4.716992e-10
Residual	6.930991e+07	169.0	NaN	NaN

По факторам: “Формат книги” и “Категория книг”

Lasso регрессия с alpha: 0.5 и 100

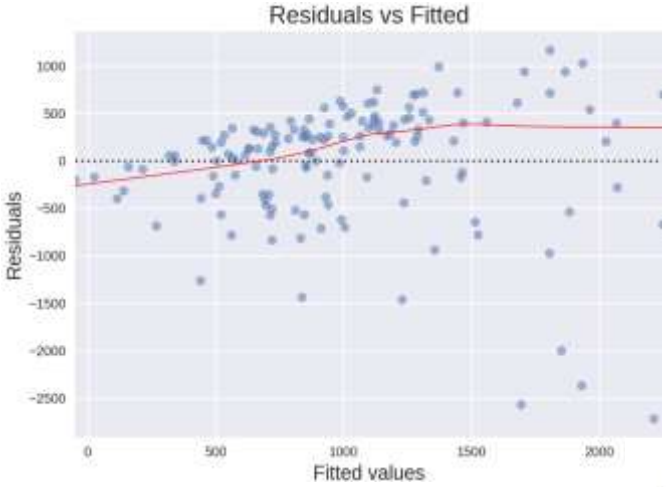
Коэффициенты при alpha = 0.5:

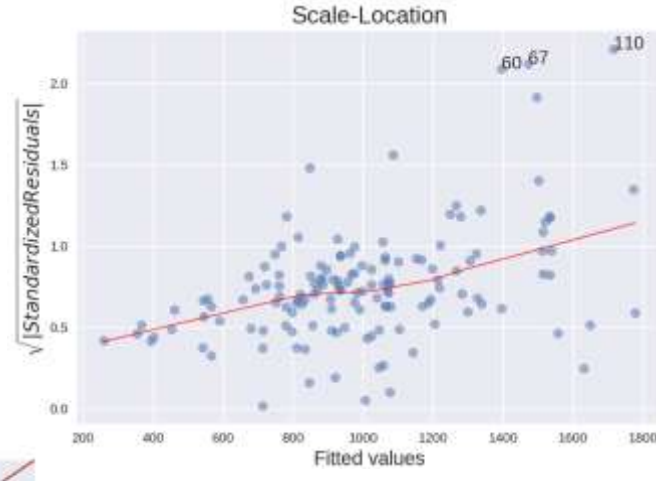
категория книг = 94.71
количество страниц = 73.53
визуальная привлекательность = -6.83
качество обложки = 16.51
качество переплёта = -65.43
качество бумаги = 184.36
качество печати = 193.41
формат книги = 316.14
размер книги = 75.61

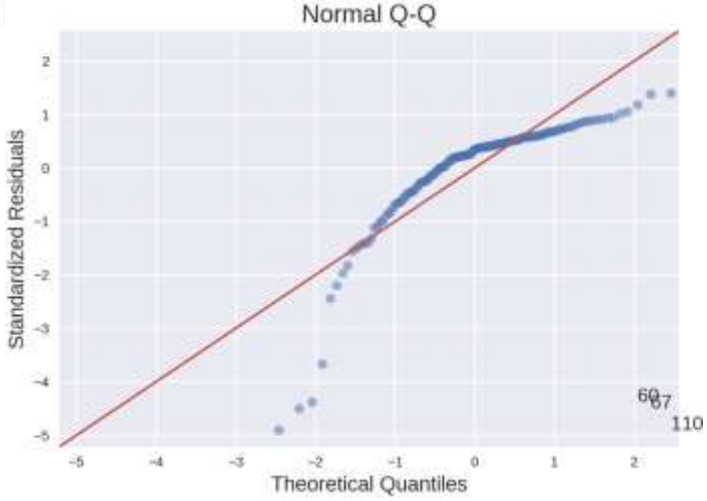
Коэффициенты при alpha = 100:

категория книг = 0.0
количество страниц = 56.88
визуальная привлекательность = 0.0
качество обложки = 0.0
качество переплёта = 0.0
качество бумаги = 127.61
качество печати = 75.04
формат книги = 237.71
размер книги = 0.0

Lasso регрессия с alpha: 0.5 и 100







VALIDATION

ABSOLUTE_ERROR

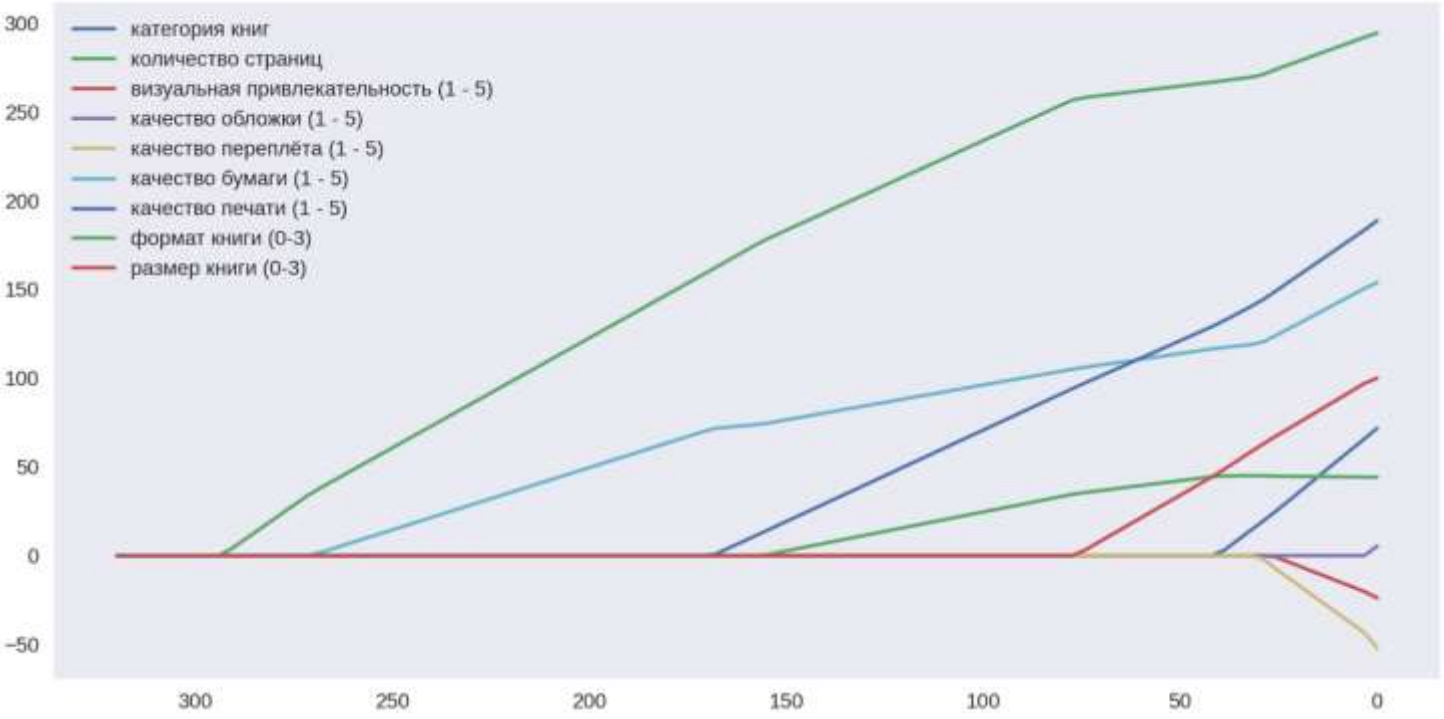
ALPHA=0.5: 417

VALIDATION

ABSOLUTE_ERROR

ALPHA=100: 400

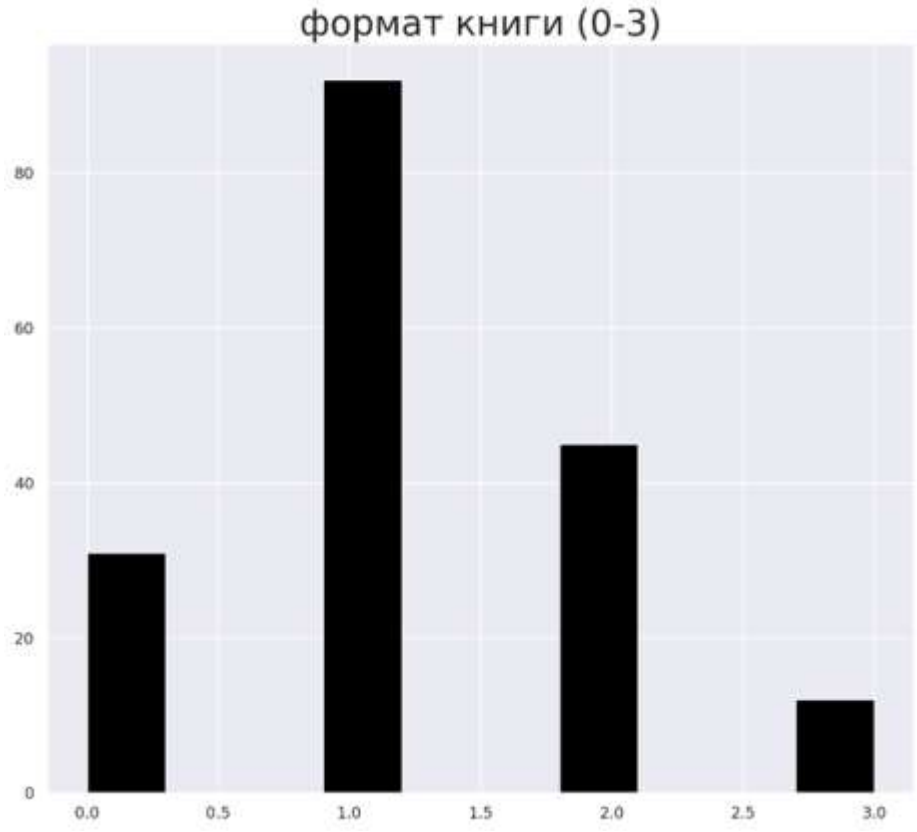
Lasso регрессия с alpha от 0 до 300

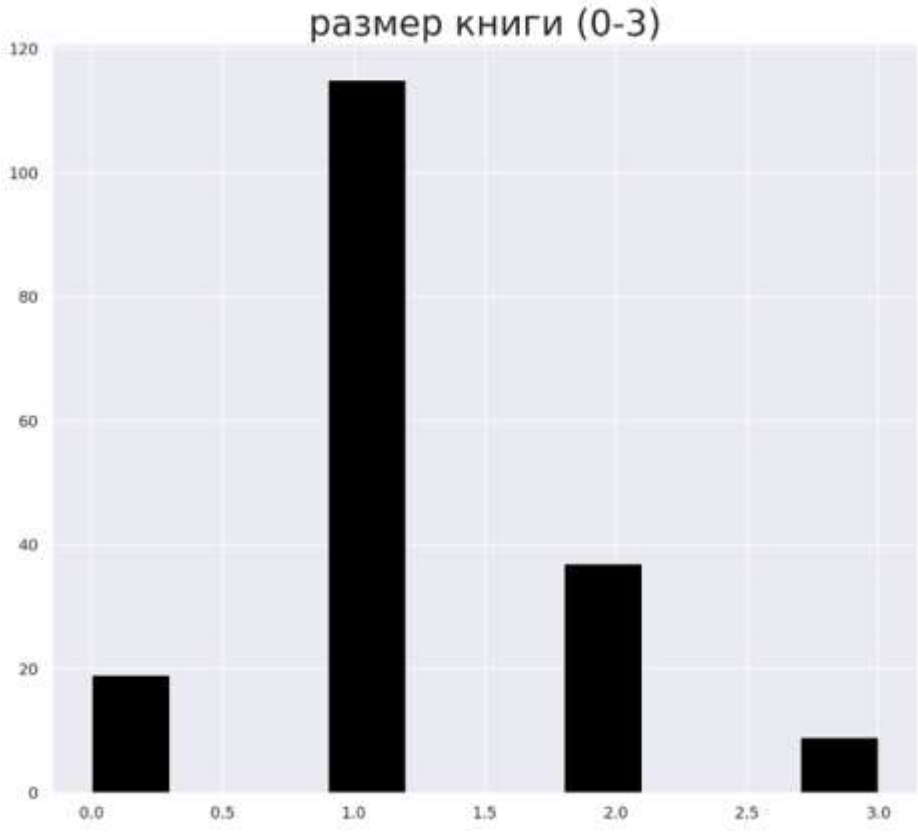


Важные предикторы:
Качество бумаги
Качество печати
Формат книги
Количество страниц

Зависимость значения коэффициентов от величины штрафа

Lasso регрессия с alpha = 0.5





Преобразование Бокса-Кокса по всем коэффициентам

Коэффициенты после преобразования:

категория книг = 0.0004

количество страниц = 0.0977

визуальная привлекательность = -0.1467

качество обложки = 0.03126

качество переплёта = 0.03440

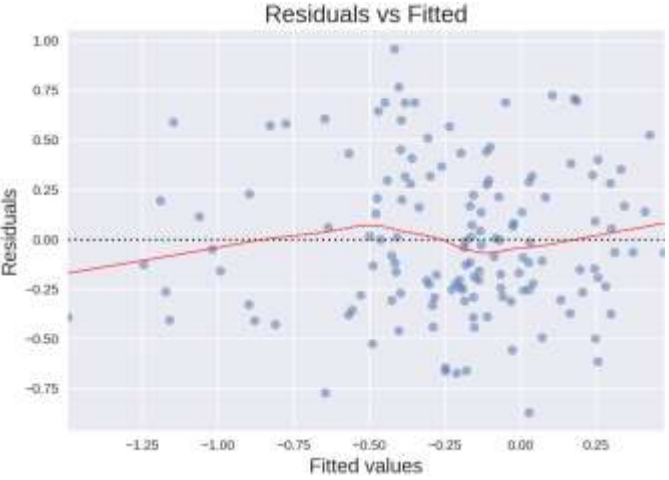
качество бумаги = 0.5113

качество печати = 0.3517

формат книги = 0.380

размер книги =0.1415

Преобразование Бокса-Кокса по всем коэффициентам

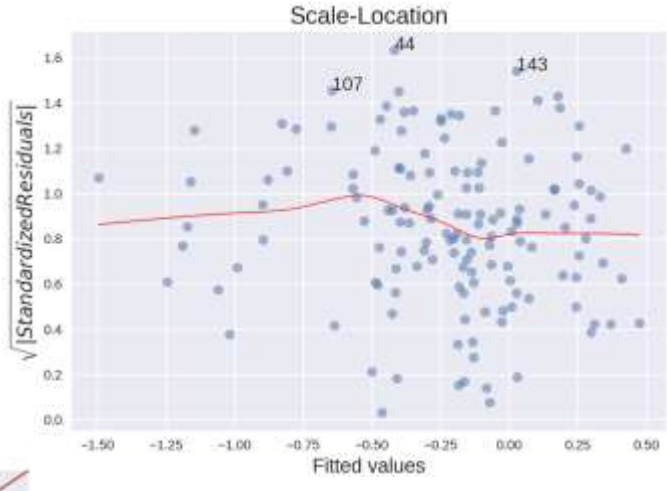


PVALUE:

JARQUE-BERE: 0.17

het_breuschpagan: 0.12

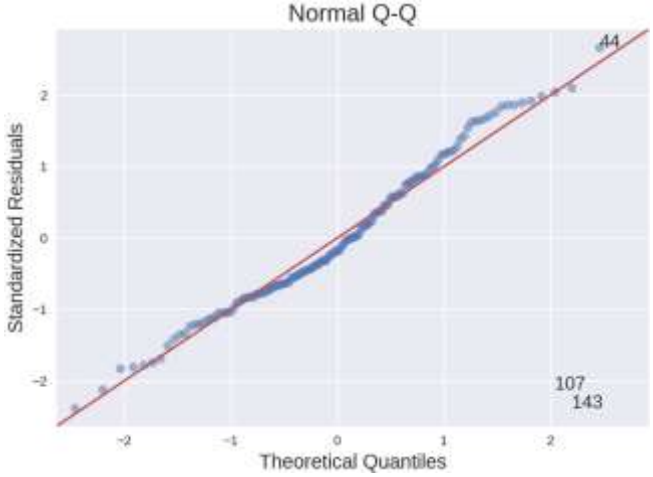
breusch_godfrey: 0.8



VALIDATION:

ABSOLUTE_ERROR

ALL COEFFS: 337



VALIDATION:

ABSOLUTE_ERROR

PRIME COEFFS: 357

Преобразование Бокса-Кокса по важным коэффициентам

Коэффициенты после преобразования:

количество страниц = 0.58

качество бумаги = 0.61

качество печати = -0.02

формат книги = 0.21

Lasso регрессия с магазинами с alpha: 0.5 и 150

Коэффициенты при alpha=0.5:

категория книг = 77.61
количество страниц = 79.24
визуальная привлекательность = -5.96
качество обложки = 29.51
качество переплёта = -85.39
качество бумаги = 205.41
качество печати = 91.31
формат книги = 244.06
размер книги = 62.31

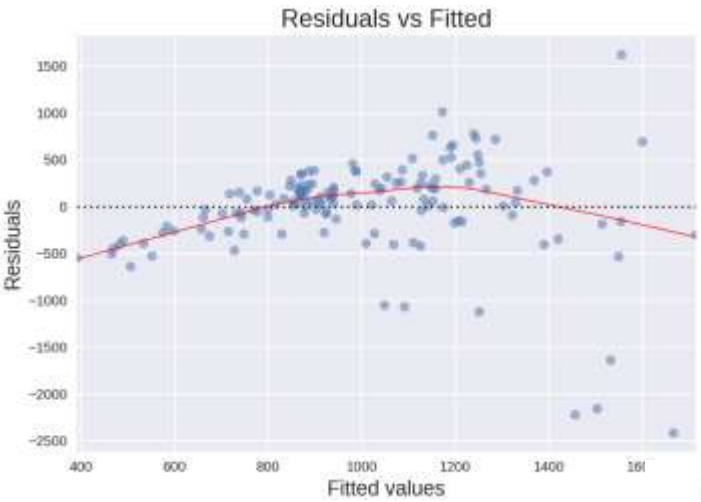
Читай-город = -42.95
Книжный лабиринт = 893.38
Молодая гвардия = 0.0

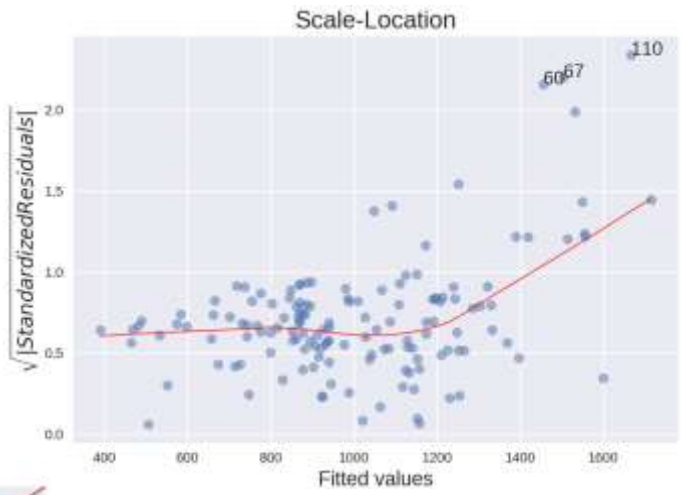
Коэффициенты при alpha=150:

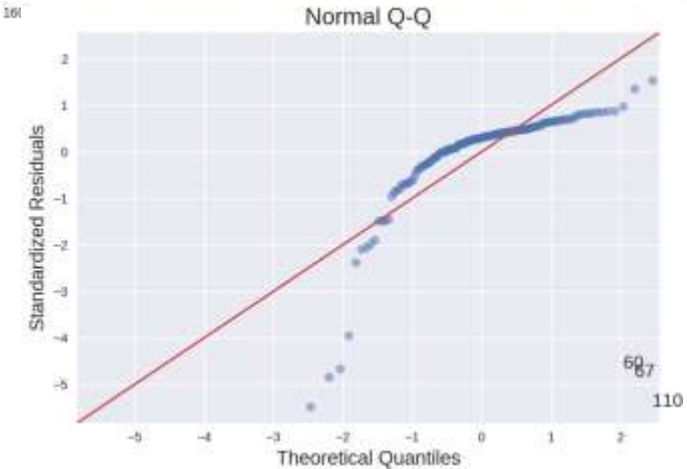
категория книг = 0.0
количество страниц = 34.81
визуальная привлекательность = 0.0
качество обложки = 0.0
качество переплёта = 0.0
качество бумаги = 110.31
качество печати = 0.0
формат книги = 159.91
размер книги = 0.0

Читай-город = 0.0
Книжный лабиринт = 303.47
Молодая гвардия = 0.0

Lasso регрессия с добавлением магазинов



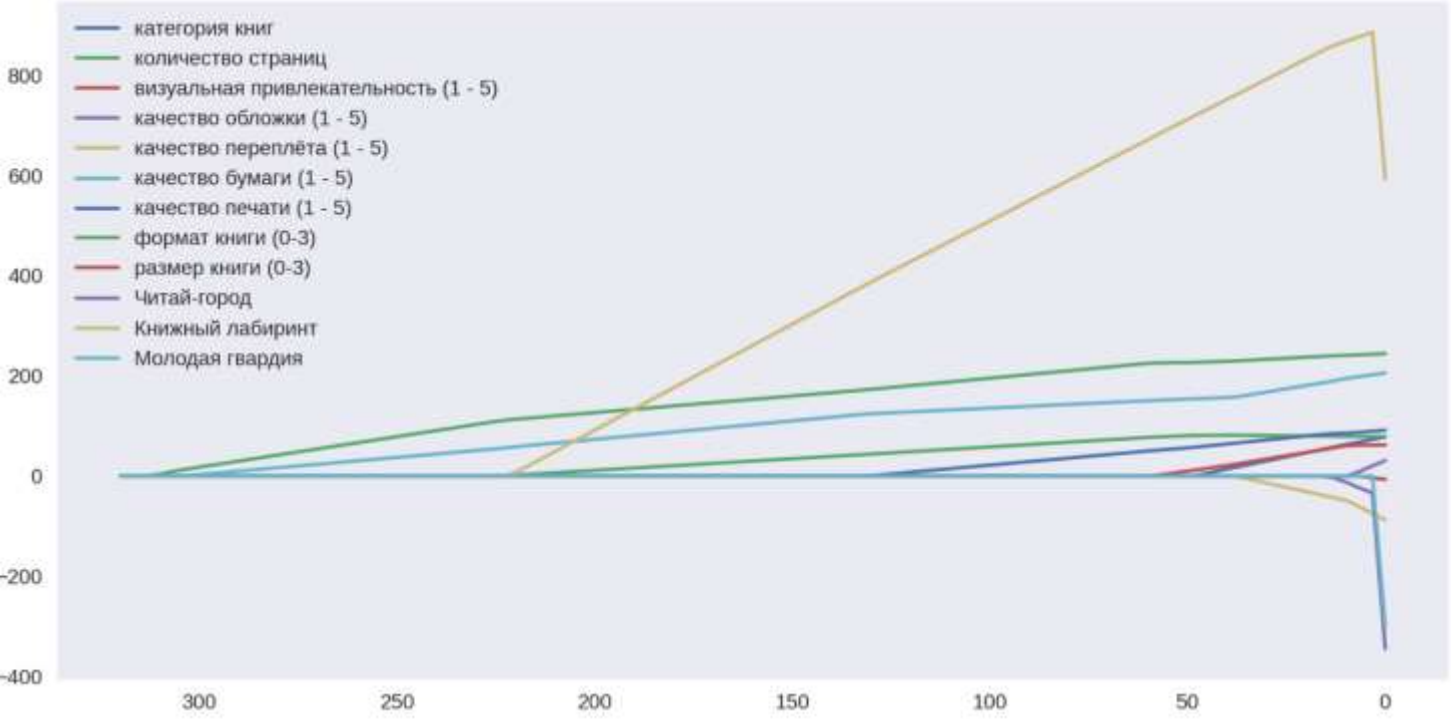




VALIDATION:
ABSOLUTE_ERROR
ALL COEFFS: 288

VALIDATION:
ABSOLUTE_ERROR
PRIME COEFFS: 266

Lasso регрессия с добавлением магазинов



Важные предикторы:
Качество бумаги
Качество печати
Формат книги
Количество страниц
Книжный лабиринт

Зависимость значения коэффициентов от величины штрафа

Преобразование Бокса-Кокса по всем коэффициентам с магазинами

Коэффициенты после преобразования:

категория книг = -5.45e-12

количество страниц = 0.38

визуальная привлекательность = -0.109

качество обложки = 0.02

качество переплёта = -6.47e-05

качество бумаги = 0.257

качество печати = 2.56e-15

формат книги = 0.211

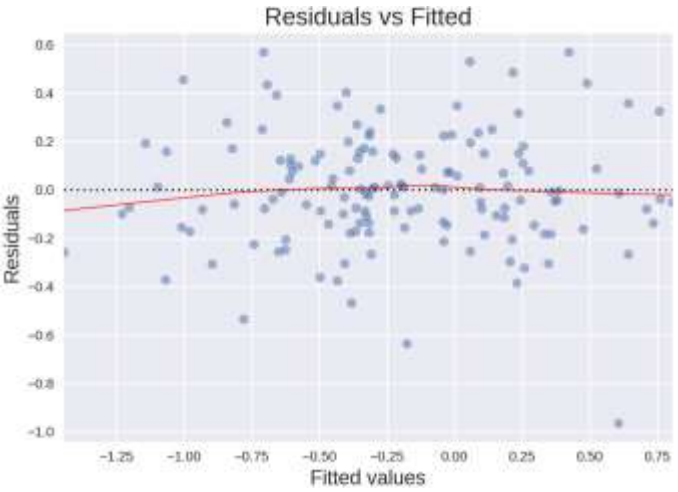
размер книги = 9.16e-09

Читай-город = -0.57

Книжный лабиринт = 0.39

Молодая гвардия = -0.47

Преобразование Бокса-Кокса по всем коэффициентам с магазинами

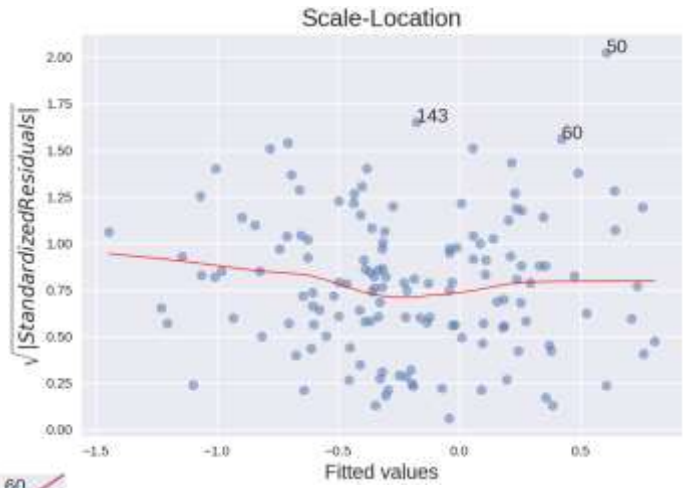


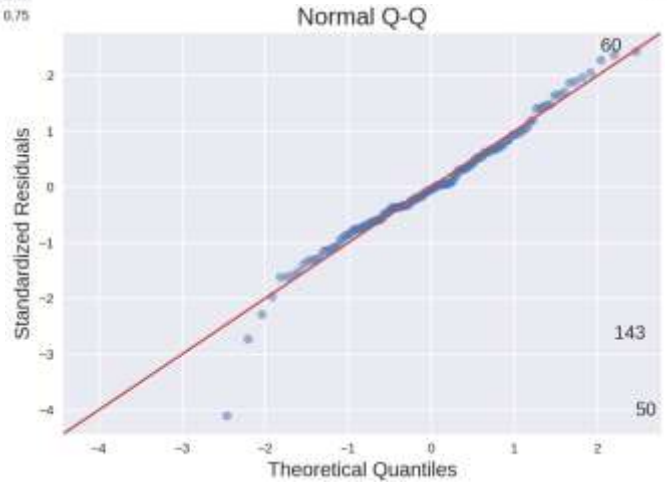
PVALUE:

JARQUE-BERE: 0.003

het_breuschpagan: 0.66

breusch_godfrey: 0.32





VALIDATION:

ABSOLUTE_ERROR

ALL COEFFS: 213

VALIDATION:

ABSOLUTE_ERROR

PRIME COEFFS: 210

Преобразование Бокса-Кокса по важным коэффициентам с магазинами

Коэффициенты после преобразования:

количество страниц = 0.012

качество бумаги = 0.48

качество печати = 221

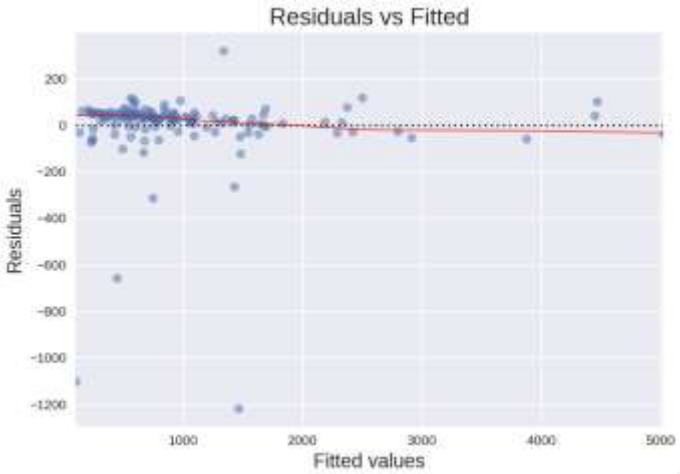
формат книги = 0.27

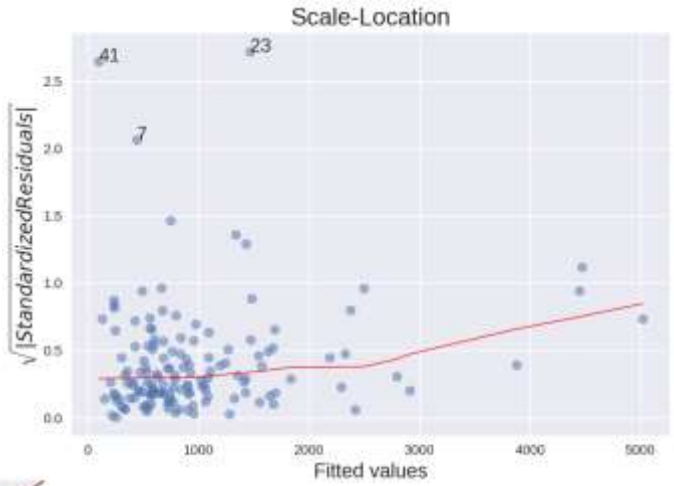
Читай-город = 131

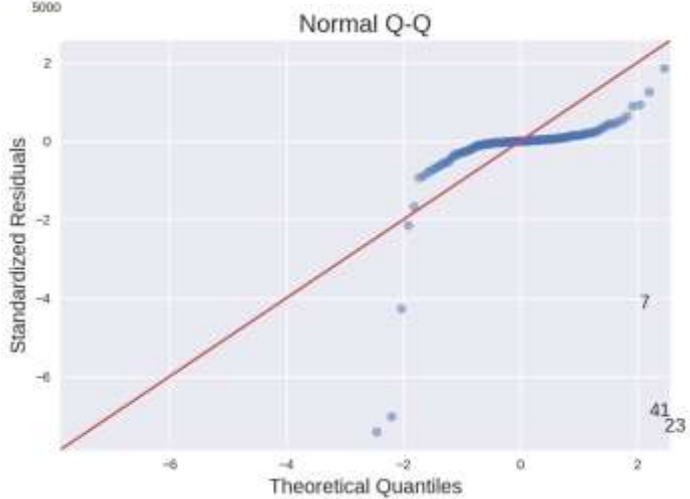
Книжный лабиринт = 132

Молодая гвардия = 131

Непараметрическая модель







VALIDATION:

ABSOLUTE_ERROR

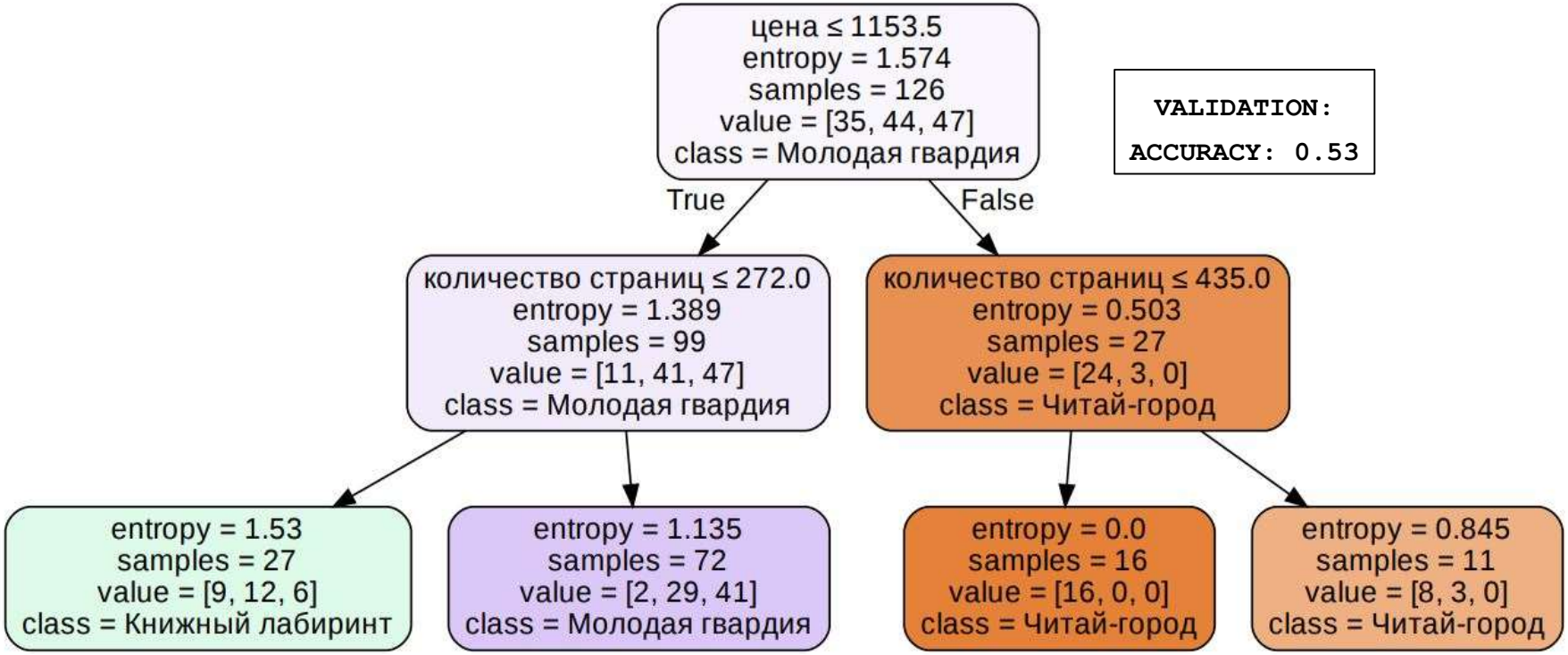
PRIME COEFFS: 384

TRAIN:

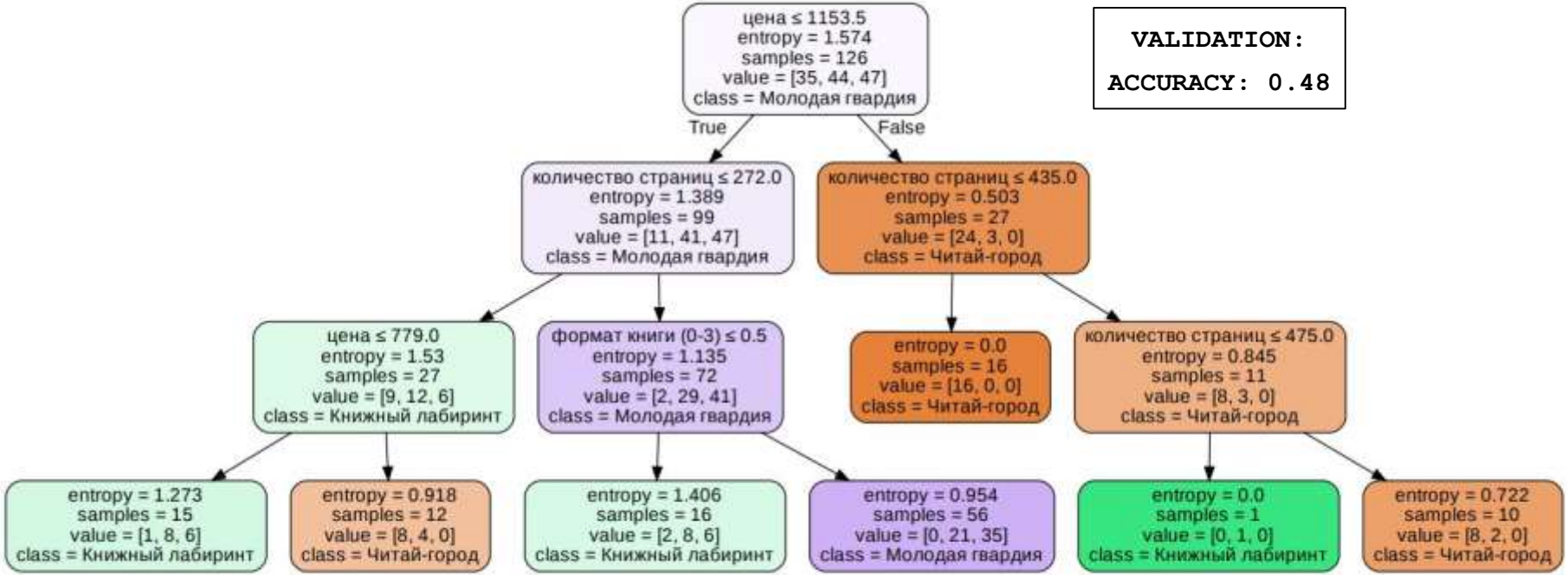
ABSOLUTE_ERROR

PRIME COEFFS: 56

Дерево решений глубины два



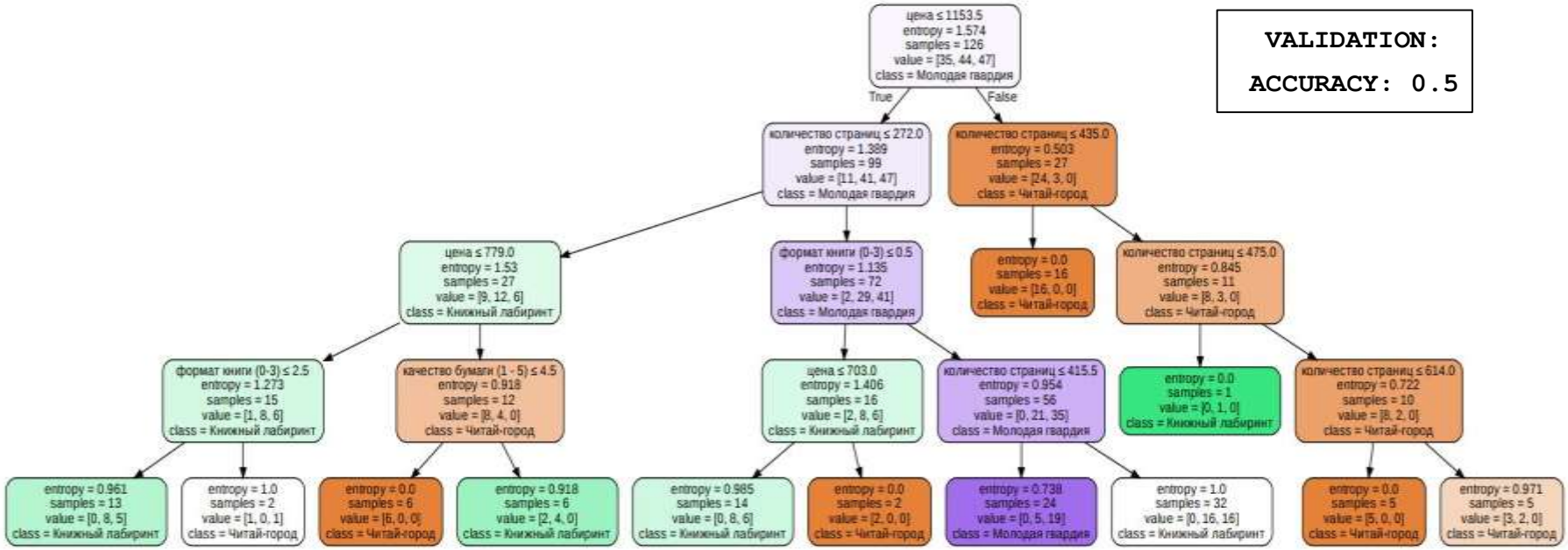
Дерево решений глубины три



VALIDATION:

ACCURACY: 0.48

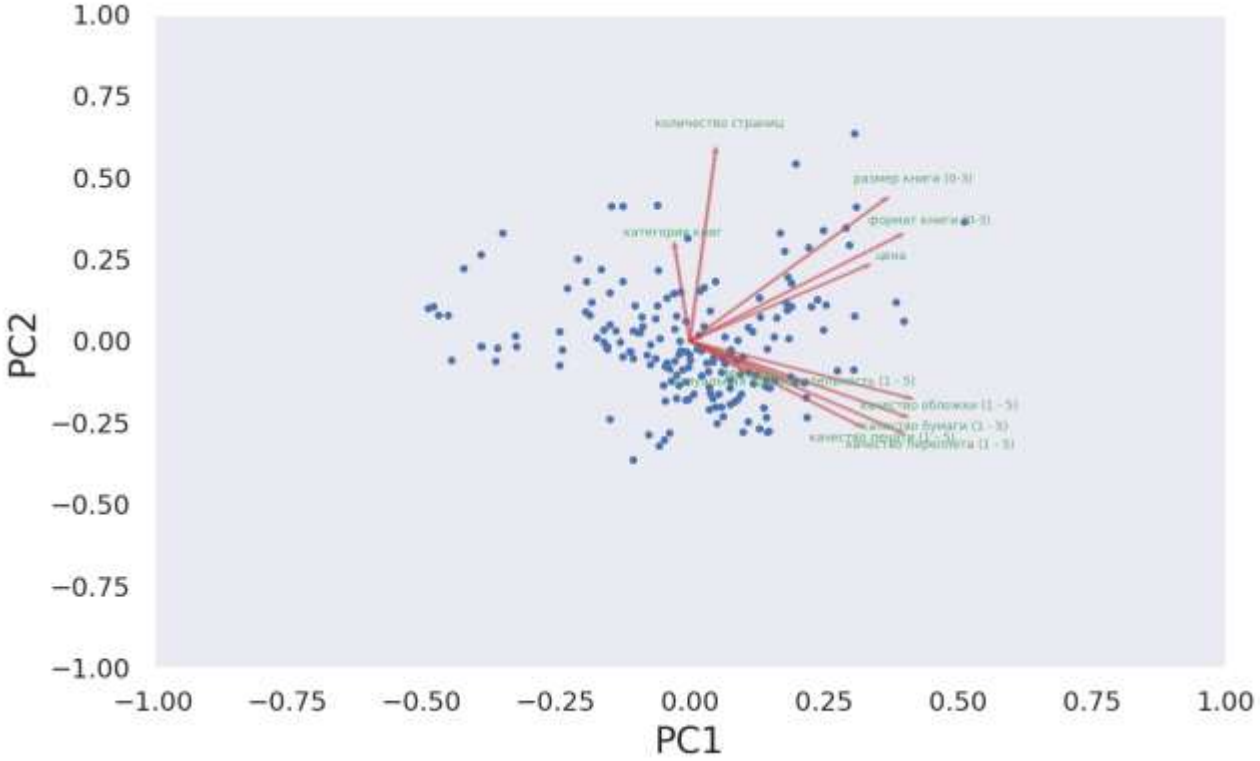
Дерево решений глубины четыре



VALIDATION:

ACCURACY: 0.5

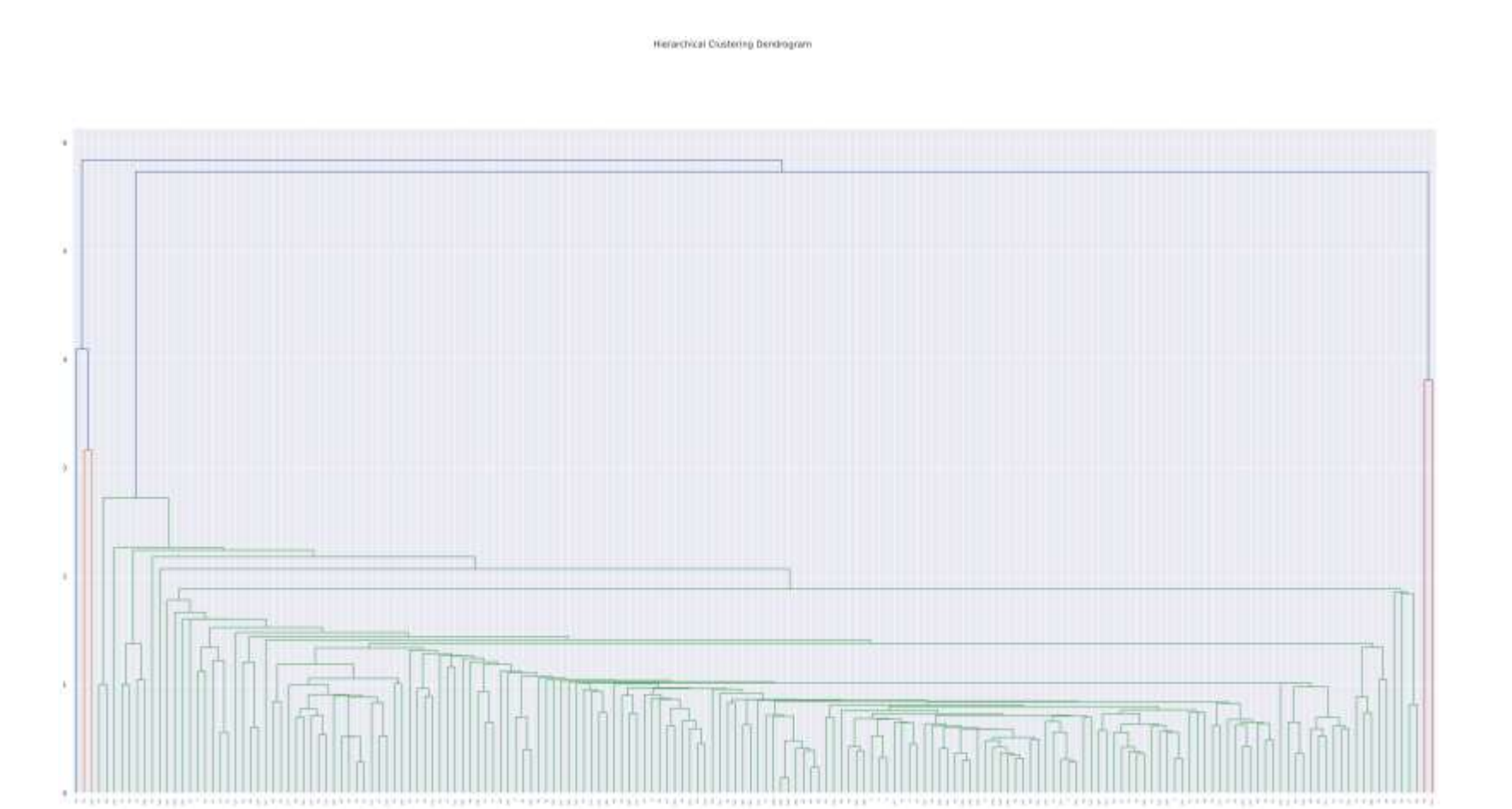
Снижение размерности



Объяснённая дисперсия:

```
PC1 = 0.47
PC2 = 0.19
PC3 = 0.13
Sum = 0.79
```

Дендрограмма по главным компонентам



Кластеризация