Проект по эконометрике

Гедонистическая ценовая функция для брендовых сумок

Подготовили Джанбекова Алина, Красногорова Лилия, Эшмеев Павел

Заранее оговоримся, что проверка всех гипотез в нашем исследовании будет проходить на 5% уровне значимости.

Переменные и их источник

Источником данных послужил сайт https://oskelly.ru/catalog/zhenskoe/sumki
Все переменные получаем из парсинга страницы товара в разделе детали, после чего убираем лишние и заполняем пропуски

Детали —

MM6 MAISON MARGIELA Черная сумка с короткими ручками из искусственной кожи

Размер	INT U
Раздел	Женское
Категория	Сумки с короткими ручками
Бренд	MM6 MAISON MARGIELA
Материал сумок	Искусственная кожа
Цвет	Черный
Длина ручки	Средние ручки
Состояние товара	Новое с биркой
Продавец	Частный продавец
Oskelly ID	3412678

Переменные для анализа:
Категориальные:

- Размер - размер сумки (категориальный признак) </
  Если не был указан размер, то INT U - универсальный
- Категория - тоут/через плечо/с короткими ручками/рюкзаки/клатчи/аксессуары для сумок (категориальный признак)
  Не было пропущенных данных
- Бренд - бренд сумки (категориальный признак)
  Не было пропущенных данных
- Материал сумок (категориальный признак)
  Не было пропущенных данных

- Цвет - цвет сумки (категориальный признак)
  Не было пропущенных данных
- Длина ручки (категориальный признак)
  Не было пропущенных данных
- Состояние товара (категориальный признак)
  Не было пропущенных данных
- Продавец (категориальный признак)
  Не было пропущенных данных
- Модель (категориальный признак)
  Если не была указана, то заполнили Ordinary
- Ценовая категория (в зависимости от бренда на основе графика распределения цен в зависимости от бренда)
  Экстремально люксовые - 'BVLGARI', 'HERMES PRE-OWNED'
  Люксовые - 'BALENCIAGA', 'BURBERRY', 'GOYARD' и тд (в коде можно посмотреть)
  Доступный люкс - 'VALENTINO', 'DOLCE&GABBANA', 'JACQUEMUS' и тд (в коде можно посмотреть)
  Дорогие - 'MARC JACOBS', 'VERSACE', 'A.P.C.' и тд (в коде можно посмотреть)
  Доступные - остальные
- Метод производства (было проведено исследование, какие бренды как делают свои сумки)
  Сделанные вручную - 'HERMES PRE-OWNED', 'CHRISTIAN DIOR PRE-OWNED', 'BOTTEGA VENETA', 'LOEWE', 'FENDI'
  Завершенные вручную - 'GUCCI','SAINT LAURENT', 'JACQUEMUS'
  Произведенные на фабрике - остальные

Бинарные:
Для них, если указано "Да" на сайте, то указываем 1
Иначе (пустые) - указываем 0

- Наличие пыльника (бинарный признак)
- Винтаж - является ли сумка винтажной (бинарный признак)
- Наличие коробки (бинарный признак)
- Наличие сертификата (бинарный признак)
- Легенда (бинарный)
  Если модель не Ordinary, то является легендарной моделью бренда

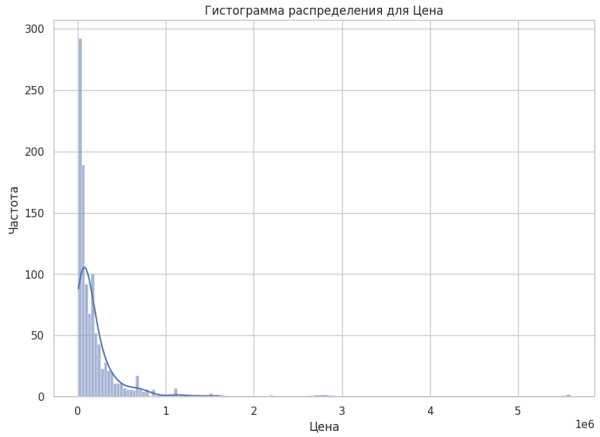
Целевая переменная:

- Цена - цена сумки, рубли.

Анализ переменных

Целевая переменная

Цена

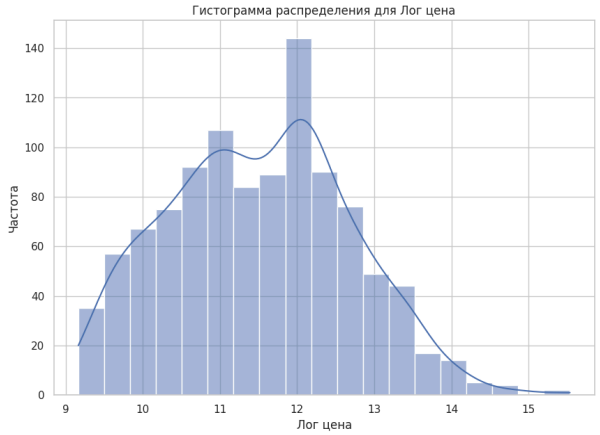


Скошенность: 8.049594378924823
Островершинность: 99.60939372241275

По построенному графику видно, что цены сумок распределены с очень длинным правым хвостом и тяжёлым левым хвостом. Показатель скошенности равный примерно 8 (при значении 0 для нормального распределения) говорит о сильной асимметрии: большинство товаров сконцентрировано внизу ценовой шкалы, но небольшое число лотов стоит в разы дороже и тянет распределение вправо. Островершинность около 100 (при показателе 3 у нормального распределения) означает, что вероятность появления экстремально высоких и низких цен намного выше, чем при нормальном распределении, то есть для выборки присуща высокая частота выбросов. Вероятно распределение прологарифмированной цены будет выглядеть сильно лучше. Взглянем на него.

Логарифм цены

Скошенность: 0.15836535158553108
Островершинность: -0.4665477186972127



Видим, что в данном случае все сильно лучше. Показатель скошенности близкий к 0 говорит о почти симметричном распределении. Показатель -0.47 островершинности указывает на плоское распределение с менее выраженными пиками и более тонкими хвостами по сравнению с нормальным распределением. По данным показателям можно сказать, что распределение близко к нормальному.

И на графике распределения обычных цен, и на графике распределения логарифмированных цен в правом хвосте имеются ненулевые столбцы, растягивающие его. С помощью анализа ящиков с усами убрали выбросы (цена больше 2000000)

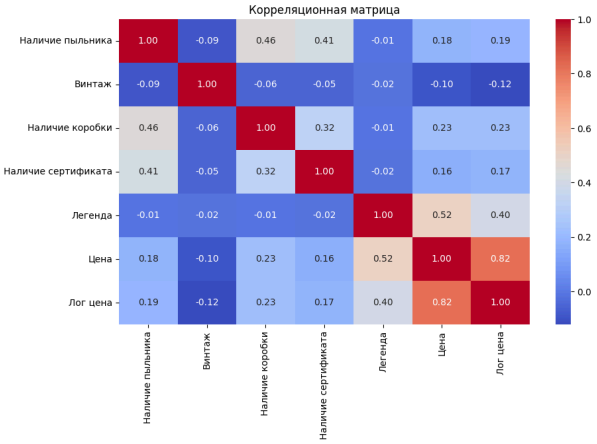
средняя стоимость рассматриваемых моделей составила 202.5 тыс. рублей
стандартное отклонение стоимости довольно велико и составляет примерно 363 тыс. рублей
размах цен сумок довольно большой - от 9 тыс. до 5.6 млн. рублей

Дамми-переменные

	Винтаж	Легенда	Наличие коробки	Наличие пыльника	Наличие сертификата
Минимум	0.000000	0.000000	0.000000	0.000000	0.000000
Максимум	1.000000	1.000000	1.000000	1.000000	1.000000
Среднее	0.067943	0.108134	0.088995	0.292823	0.084211
Стд. отклонение	0.251768	0.310698	0.284873	0.455276	0.277836
Размах	1.000000	1.000000	1.000000	1.000000	1.000000

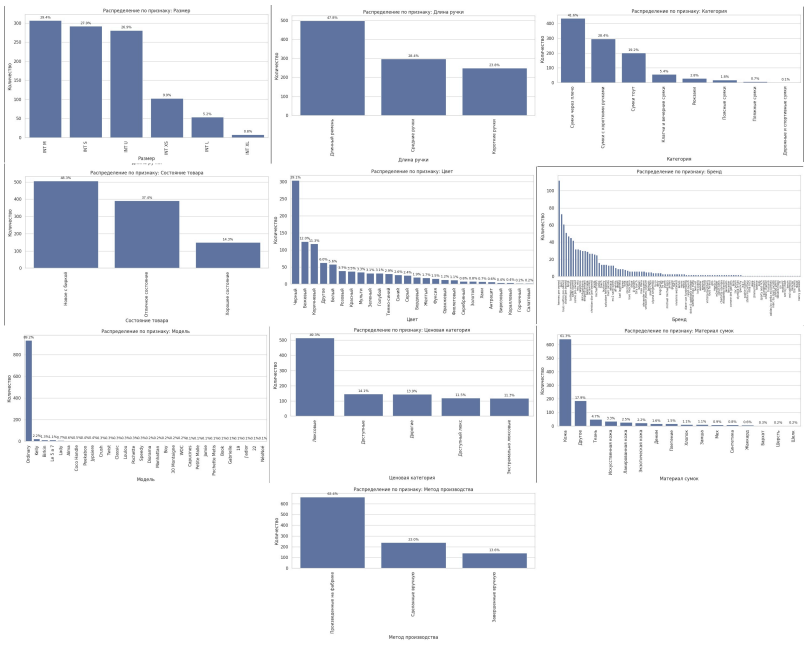
Выводы:

- лишь у 8% сумок из выборки имеется сертификат
- пыльник присутствует примерно у 30% рассматриваемых моделей
- только 9% из всех сумок в выборке поставляются в коробках
- винтажных сумок всего 6% от всей выборки



Из численных переменных лучше всего с ценой коррелирует наличие легенды. Другие параметры коррелируют со стоимостью довольно слабо.

Категориальные переменные



Почти для всех категориальных переменных видно, что большая часть данных приходится лишь на несколько категорий, даже если их довольно много для данного параметра. Логичной идеей является объединение классов для их более равномерного распределения по выборке, что позволило бы сделать оценки моделей более устойчивыми. Но важно также учитывать, что при склейке разных по средней цене групп мы искусственно повышаем шум внутри переменной, что впоследствии уменьшает долю объяснённой дисперсии. Кроме того при объединении отличающихся по цене категорий неучтённая разница в их цене перекладывается на ошибку модели, что увеличивает гетероскедастичность.

Отдельно можно выделить характеристику "Модель" - почти все объекты имеют показатель "Ordinary" по данному параметру. Кажется, что не имеет особого смысла использовать данную характеристику для построения моделей, так как она не будет вносить какой-либо значимой дополнительной информации об объектах. Убираем ее из выборки.

Для других характеристик, в которых хотелось бы провести некоторые склейки разных классов, с помощью ящиков с усами для разных категорий относительно их стоимости и посмотрим возможно ли теоретически объединить категории: у многих классов значения средних различаются, из-за чего их объединение статистически ухудшит результаты моделей.

Из того, что теоретически возможно слить можно выделить размеры сумок L и XL - судя по ящикам с усами их параметры схожи. Проверим это с помощью t-теста Уэлча и U-теста Манна-Уитни (использовать будем логарифм цены, так как ее распределение ближе к нормальному и не скошено в одну сторону).

Levene: p-value = 0.5926
Welch t-test: t = 0.184, p-value = 0.8570
Mann-Whitney U-test: U = 209, p-value = 0.8914

Нет статистически значимой разницы средних лог-цен ($p > 0.05$). Размеры L и XL можно считать однородными по цене, поэтому мы их объединяем.

Эконометрические модели

Для использования категориальных характеристик в дальнейших моделях, преобразуем их с помощью One Hot Encoder.

Линейная регрессия

OLS Regression Results			
Dep. Variable:	Цена	R-squared:	0.566
Model:	OLS	Adj. R-squared:	0.486
Method:	Least Squares	F-statistic:	7.096
Date:	Wed, 07 May 2025	Prob (F-statistic):	5.07e-85
Time:	21:32:46	Log-Likelihood:	-14507.
No. Observations:	1051	AIC:	2.934e+04
Df Residuals:	887	BIC:	3.015e+04
Df Model:	163		
Covariance Type:	nonrobust		

Статистически значимыми оказались
Бинарный: Винтаж, Наличие сертификата, Легенда
Категориальные: Бренд CHANEL PRE-OWNED, Бренд CHRISTIAN DIOR PRE-OWNED, Бренд HERMES PRE-OWNED,
Экзотическая кожа,
Короткие ручки, Экстремально люксовые

Отрицательно влияют на цену: Винтаж (в среднем уменьшает стоимость на -102900 рублей), Бренд CHRISTIAN DIOR PRE-OWNED(-111400), Короткие ручки (-54590)

Положительно влияют на цену: Наличие сертификата (в среднем увеличивает стоимость на 93870 рублей), Легенда (в среднем увеличивает стоимость 360100), Бренд CHANEL PRE-OWNED (131200), Бренд HERMES PRE-OWNED (303500), Экзотическая кожа (573700), Экстремально люксовые (352800)

Мультиколлинеарность: есть показатели $VIF > 10$ следовательно, присутствует

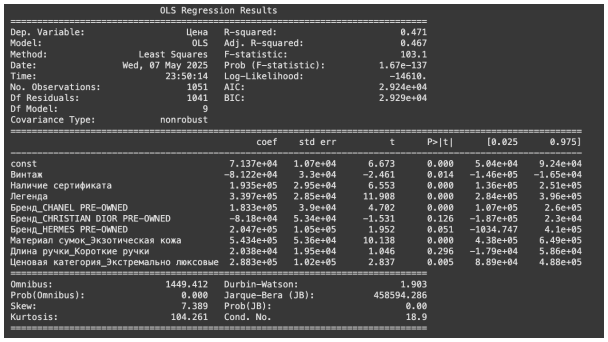
МГК

Изначально берем все главные компоненты и оцениваем на них линейную регрессию. Незначимые главные компоненты удаляем, остальные оставляем (значимых осталось только 45).

OLS Regression Results			
Dep. Variable:	Цена	R-squared:	0.522
Model:	OLS	Adj. R-squared:	0.501
Method:	Least Squares	F-statistic:	24.38
Date:	Wed, 07 May 2025	Prob (F-statistic):	6.28e-130
Time:	23:04:58	Log-Likelihood:	-14558.
No. Observations:	1051	AIC:	2.921e+04
Df Residuals:	1005	BIC:	2.944e+04
Df Model:	45		
Covariance Type:	nonrobust		

Получилось, что все параметры в новом параметрическом пространстве имеют $VIF < 10$, то есть мультиколлинеарность в такой выборке отсутствует. При этом $Adj.R^2$, AIC и BIC у модели со значимыми главными компонентами лучше чем у первоначальной модели линейной регрессии ($0.501 > 0.486$, $2921 < 2934$, $2944 < 3015$ соответственно). К сожалению, при использовании подобного рода данных интерпретировать результаты модели не представляется возможным, поэтому попробуем рассмотреть модель линейной регрессии только со значимыми характеристиками.

Линейная регрессия со статистически значимыми



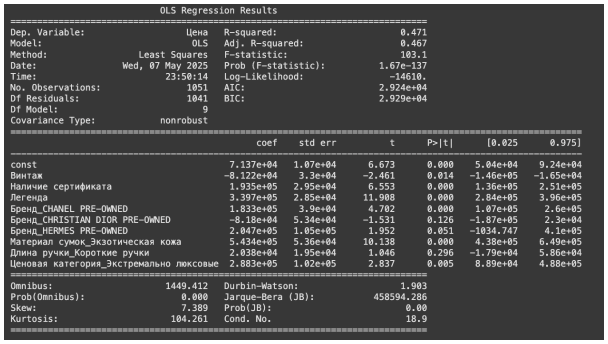
OLS Regression Results

Dep. Variable:	Цена	R-squared:	0.471
Model:	OLS	Adj. R-squared:	0.467
Method:	Least Squares	F-statistic:	103.1
Date:	Wed, 07 May 2025	Prob (F-statistic):	1.67e-137
Time:	23:50:14	Log-Likelihood:	-14610.
No. Observations:	1051	AIC:	2.924e+04
Df Residuals:	1041	BIC:	2.929e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.137e+04	1.07e+04	6.673	0.000	5.04e+04	9.24e+04
Винтаж	-8.122e+04	3.3e+04	-2.461	0.014	-1.46e+05	-1.65e+04
Наличие сертификата	1.935e+05	2.95e+04	6.553	0.000	1.36e+05	2.51e+05
Легенда	3.397e+05	2.85e+04	11.908	0.000	2.84e+05	3.96e+05
Бренд_CHANEL PRE-OWNED	1.833e+05	3.9e+04	4.702	0.000	1.07e+05	2.6e+05
Бренд_CHRISTIAN DIOR PRE-OWNED	-8.18e+04	5.34e+04	-1.531	0.126	-1.87e+05	2.3e+04
Бренд_HERMES PRE-OWNED	2.047e+05	1.05e+05	1.952	0.051	-1034.747	4.1e+05
Материал сумок_Экзотическая кожа	5.434e+05	5.36e+04	10.138	0.000	4.38e+05	6.49e+05
Длина ручки_Короткие ручки	2.038e+04	1.95e+04	1.046	0.296	-1.79e+04	5.86e+04
Ценовая категория_Экстремально люксовые	2.883e+05	1.02e+05	2.837	0.005	8.89e+04	4.88e+05

Omnibus:	1449.412	Durbin-Watson:	1.903
Prob(Omnibus):	0.000	Jarque-Bera (JB):	458594.286
Skew:	7.389	Prob(JB):	0.00
Kurtosis:	104.261	Cond. No.	18.9

Статистически значимые: const, Винтаж (цена в среднем -81220 рублей), Наличие сертификата (+193500), Легенда (+339700), Бренд CHANEL PRE-OWNED (+183300), Экзотическая кожа (+543400), Экстремально люксовые (+288300)
Вновь уберем статистически не значимые переменные



OLS Regression Results

Dep. Variable:	Цена	R-squared:	0.471
Model:	OLS	Adj. R-squared:	0.467
Method:	Least Squares	F-statistic:	103.1
Date:	Wed, 07 May 2025	Prob (F-statistic):	1.67e-137
Time:	23:50:14	Log-Likelihood:	-14610.
No. Observations:	1051	AIC:	2.924e+04
Df Residuals:	1041	BIC:	2.929e+04
Df Model:	9		
Covariance Type:	nonrobust		

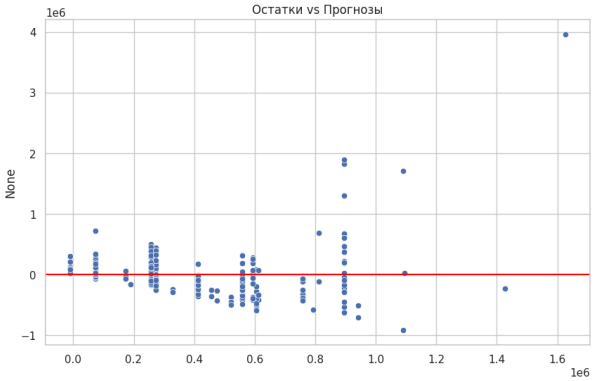
	coef	std err	t	P>|t|	[0.025	0.975]
const	7.137e+04	1.07e+04	6.673	0.000	5.04e+04	9.24e+04
Винтаж	-8.122e+04	3.3e+04	-2.461	0.014	-1.46e+05	-1.65e+04
Наличие сертификата	1.935e+05	2.95e+04	6.553	0.000	1.36e+05	2.51e+05
Легенда	3.397e+05	2.85e+04	11.908	0.000	2.84e+05	3.96e+05
Бренд_CHANEL PRE-OWNED	1.833e+05	3.9e+04	4.702	0.000	1.07e+05	2.6e+05
Бренд_CHRISTIAN DIOR PRE-OWNED	-8.18e+04	5.34e+04	-1.531	0.126	-1.87e+05	2.3e+04
Бренд_HERMES PRE-OWNED	2.047e+05	1.05e+05	1.952	0.051	-1034.747	4.1e+05
Материал сумок_Экзотическая кожа	5.434e+05	5.36e+04	10.138	0.000	4.38e+05	6.49e+05
Длина ручки_Короткие ручки	2.038e+04	1.95e+04	1.046	0.296	-1.79e+04	5.86e+04
Ценовая категория_Экстремально люксовые	2.883e+05	1.02e+05	2.837	0.005	8.89e+04	4.88e+05

Omnibus:	1449.412	Durbin-Watson:	1.903
Prob(Omnibus):	0.000	Jarque-Bera (JB):	458594.286
Skew:	7.389	Prob(JB):	0.00
Kurtosis:	104.261	Cond. No.	18.9

Как видим, здесь уже все коэффициенты значимы. При этом $Adj.R^2$ все же хуже чем у самой полной модели, но AIC и BIC меньше. Теперь проверим получившуюся модель на мультиколлинеарность.

Мультиколлинеарность в данной модели отсутствует (VIF < 10).

Гетероскедастичность:



Кажется, на графике не наблюдается зависимость разброса остатков от прогнозных значений и средний уровень остатков во всех частях графика не равен нулю, то есть вероятнее в модели не наблюдается гетероскедастичность. При этом похоже, средний уровень остатков зависит от прогнозных значений, что говорит о возможно неверно выбранной функциональной форме. Проверим наши догадки с помощью теста Уайта и Бройша-Пагана.

- **White test: p-value: 7.345950734159816e-198**
  Гомоскедастичность модели отвергается
- **P-value теста Бройша-Пагана: 2.719457670529733e-35**
  Гомоскедастичность модели отвергается
  То есть все же гетероскедастичность в нашей модели присутствует. Для её устранения воспользуемся попрравкой с использованием оценок в форме Davidson и MacKinnon (HC3) для дисперсии коэффициентов. Используется именно эта форма, так как в нашей выборке пристутствуют выбросы по стоимости, которые мы видели при первичном анализе датасета.

OLS Regression Results

Dep. Variable:	Цена	R-squared:	0.468
Model:	OLS	Adj. R-squared:	0.465
Method:	Least Squares	F-statistic:	36.23
Date:	Thu, 08 May 2025	Prob (F-statistic):	5.52e-40
Time:	01:58:47	Log-Likelihood:	-14614.
No. Observations:	1051	AIC:	2.924e+04
Df Residuals:	1044	BIC:	2.928e+04
Df Model:	6		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
const	7.393e+04	1.19e+04	6.220	0.000	5.06e+04	9.72e+04
Винтаж	-8.327e+04	2.51e+04	-3.315	0.001	-1.33e+05	-3.4e+04
Наличие сертификата	1.997e+05	6.64e+04	3.005	0.003	6.95e+04	3.3e+05
Легенда	3.374e+05	4.92e+04	6.852	0.000	2.41e+05	4.34e+05
Бренд_CHANEL PRE-OWNED	1.832e+05	3.39e+04	5.399	0.000	1.17e+05	2.5e+05
Материал сумок_Экзотическая кожа	5.306e+05	2.38e+05	2.227	0.026	6.37e+04	9.97e+05
Ценовая категория_Экстремально люксовые	4.848e+05	4.42e+04	10.963	0.000	3.98e+05	5.71e+05

Omnibus:	1441.002	Durbin-Watson:	1.899
Prob(Omnibus):	0.000	Jarque-Bera (JB):	454146.500
Skew:	7.305	Prob(JB):	0.00
Kurtosis:	103.783	Cond. No.	6.66

White test: p-value: 7.345950734159816e-198
Гомоскедастичность модели отвергается

Итог