

ИССЛЕДОВАНИЕ ВЛИЯНИЯ ХАРАКТЕРИСТИК АВТОМОБИЛЯ НА ЕГО СТОИМОСТЬ: АНАЛИЗ И ПРОГНОЗИРОВАНИЕ С ПОМОЩЬЮ PYTHON

А.А. Миролюбова, Н.В. Смирнова, Д.А. Ворошин

Анастасия Александровна Миролюбова* (ORCID 0000-0003-3785-0538), Наталия Владимировна Смирнова (ORCID 0000-0002-1873-5334), Дмитрий Алексеевич Ворошин
Ивановский государственный химико-технологический университет, пр. Шереметевский, 7, Иваново, 153000, Россия
E-mail: mirolubowa@mail.ru*, nataliasmirnova_@mail.ru, ivtf@bk.ru

Научная статья посвящена исследованию влияния характеристик автомобилей на их стоимость с использованием методов машинного обучения и статического анализа на языке программирования Python. Цель работы состояла в построении прогностической модели, позволяющей выявить ключевые факторы, влияющие на цену транспортных средств, на основании данных о поддержанных автомобилях, размещенных на онлайн-платформе Kaggle. Актуальность данной темы связана не только с растущими возможностями цифровых технологий, но и необходимостью объективной оценки автомобилей на вторичном рынке. Этому рынку характерно ряд особенностей, в том числе низкий уровень стандартизации, высокий уровень децентрализации и большой объём неструктурированных данных. В таких условиях важность объективной, понятной и быстрой оценки стоимости автомобиля значительно возрастает. Исследование проводилось по методике, которая включала шесть этапов: 1) загрузка и предобработка данных; 2) описательный анализ данных; 3) визуализация данных; 4) формулировка и проверка гипотез; 5) корреляционный анализ и 6) прогнозирование. Современные библиотеки Python, такие как Sklearn, Numpy, Pandas, SciPy, Statsmodels, Matplotlib, Seaborn и Plotly предоставляют широкие возможности для анализа больших и разнородных данных, характерных для автомобильного рынка, а также гибкой настройки и интерактивной визуализации. Результаты интеллектуального анализа и прогнозирования стоимости поддержанного автомобиля показали адекватность серии однофакторных линейных регрессионных моделей. Достаточно высокую предсказательную способность продемонстрировали модели, в составе которых физические и технические характеристики авто: объём двигателя (enginesize); снаряжённая масса (curbweight), мощность (horsepower) и ширина автомобиля (carwidth).

Ключевые слова: автомобиль, визуализация данных, вторичный рынок, корреляционный анализ, однофакторная регрессионная модель, описательная статистика, проверка гипотез, прогнозирование стоимости, Python.

INVESTIGATING THE IMPACT OF CAR CHARACTERISTICS ON ITS VALUE: ANALYSIS AND FORECASTING USING PYTHON

A.A. Mirolubova, N.V. Smirnova, D.A. Voroshin

Anastasia A. Mirolubova* (ORCID 0000-0003-3785-0538), Natalia V. Smirnova (ORCID 0000-0002-1873-5334), Dmitry A. Voroshin
Ivanovo State University of Chemistry and Technology, Sheremetevsky ave., 7, Ivanovo, 153000, Russia
E-mail: mirolubowa@mail.ru*, nataliasmirnova_@mail.ru, ivtf@bk.ru

The article is devoted to the study of the influence of car characteristics on their value using machine learning methods and static analysis in the Python programming language. The goal of the work was to build a predictive model that allows for the identification of key factors affecting the price of vehicles, based on data on used cars posted on the online platform Kaggle. The relevance of this topic is related not only to the growing capabilities of digital technologies, but also to the need for an objective assessment of cars in the secondary market. This market is characterized by a number of features, including a low level of standardization, a high level of decentralization, and a large amount of unsystematized data. In such conditions,

the importance of an objective, clear, and quick assessment of the value of a car increases significantly. The study is conducted using a methodology that includes six stages: 1) data loading and preprocessing; 2) descriptive data analysis; 3) data visualization; 4) hypothesis formulation and testing; 5) correlation analysis; and 6) forecasting. Modern Python libraries such as Sklearn, Numpy, Pandas, SciPy, Statsmodels, Matplotlib, Seaborn, and Plotly provide extensive capabilities for analyzing large and heterogeneous data specific to the car market, as well as flexible customization and interactive visualization. The results of data mining and forecasting of the value of a used car have shown the adequacy of a series of one-factor linear regression models. A fairly high predictive ability is demonstrated by models that include the physical and technical characteristics of a car: engine size; curb weight, horsepower and car width.

Keywords: car, data visualization, secondary market, correlation analysis, single-factor regression model, descriptive statistics, hypothesis testing, price forecasting, Python.

Для цитирования:

Миролюбова А.А., Смирнова Н.В., Ворошин Д.А. Исследование влияния характеристик автомобиля на его стоимость: анализ и прогнозирование с помощью Python. *Известия высших учебных заведений. Серия «Экономика, финансы и управление производством» [Ивэкофин]*. 2025. № 04(66). С. 115-127. DOI: 10.6060/ivecofin.2025664.751

For citation:

Mirolubova A.A., Smirnova N.V., Voroshin D.A. Investigating the impact of car characteristics on its value: analysis and forecasting using Python. *Ivecofin*. 2025. N 04(66). P. 115-127. DOI: 10.6060/ivecofin.2025664.751 (in Russian)

ВВЕДЕНИЕ

На современном этапе развития цифровых технологий и анализа данных становится всё более актуальным применение методов машинного обучения и статистического анализа для решения прикладных задач в различных сферах экономики. Одной из таких задач является оценка стоимости автомобилей на вторичном рынке. Покупка и продажа автомобилей – процесс, в котором как продавец, так и покупатель стремятся к объективной оценке рыночной стоимости транспортного средства, исходя из его технических и эксплуатационных характеристик. На сегодняшний день рынок подержанных автомобилей является одним из крупнейших сегментов транспортной отрасли в России. По данным аналитического агентства «Автостат», в 2024 году объёмы продаж подержанных автомобилей в России составили 6 млн 036,6 тыс. легковых автомобилей с пробегом [1]. Это на 6% больше, чем в 2023 году и почти в три раза больше, чем объём продаж новых автомобилей. Такая структура отражает фундаментальные изменения на рынке: население всё чаще делает выбор в пользу более доступного транспорта, предпочитая подержанные автомобили новым. Этот сдвиг связан сразу с несколькими макроэкономическими и социальными факторами: рост цен на новые автомобили; снижение доступности автокредитов; снижение реальных доходов населения. В свою очередь на стоимость подержанного автомобиля влияет его марка, модель, год выпуска, пробег, тип двигателя, коробка передач, наличие повреждений и многие другие

факторы. При этом эти характеристики не оказывают одинакового влияния на цену. Одни могут значительно увеличивать стоимость, другие, – наоборот, её снижать. В таких условиях задача построения модели, способной точно предсказывать цену автомобиля на основе его характеристик, приобретает практическую значимость как для индивидуальных пользователей, так и для компаний, занимающихся покупкой и продажей автомобилей. Поэтому цель данной работы заключалась в исследовании влияния различных характеристик автомобилей на их стоимость и построение прогностической модели с использованием языка программирования Python и методов машинного обучения.

МЕТОДИКА ИССЛЕДОВАНИЯ

Для решения поставленной цели был использован стек технологий (рис. 1). Это язык программирования Python и библиотеки для анализа и прогнозирования данных и машинного обучения. Для программирования использовался бесплатный сервис Google Colab.

Исследовательский опыт авторов [2-4] позволил проводить анализ влияния характеристик автомобиля на его стоимость и её прогнозирование в соответствии с методикой, включающей восемь этапов (рис. 2). Первый этап – это загрузка и предобработка данных. На этапе предобработки данных информация приводится к единому формату, обрабатываются пропущенные значения, корректируется тип данных и при необходимости осуществляется кодирование категориальных признаков [5, 6].

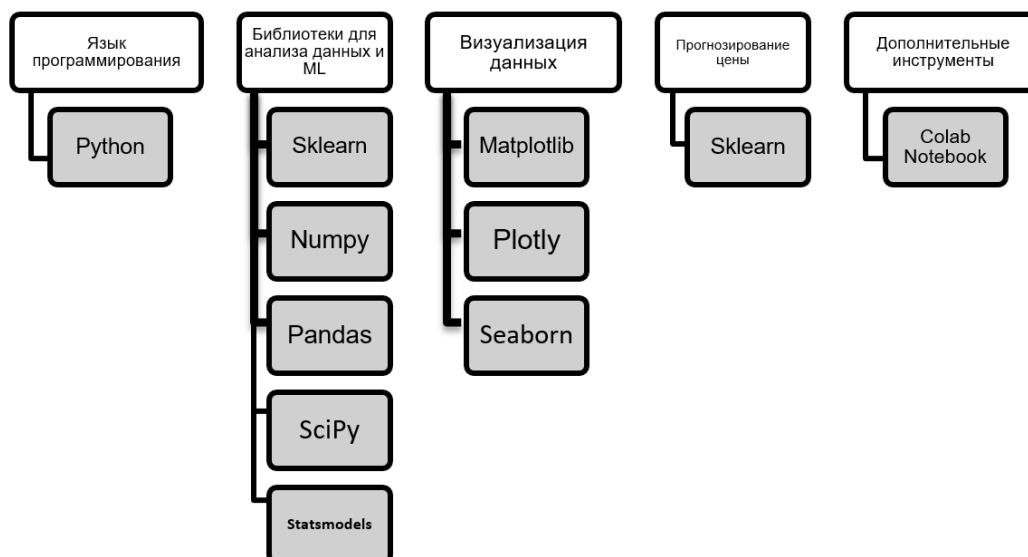


Рисунок 1. Стэк технологий для исследования влияния характеристик автомобиля на его стоимость и прогнозирование

Figure 1. Technology stack for researching the impact of car characteristics on its cost and forecasting



Рисунок 2. Методика исследования влияния характеристик автомобиля на его стоимость: анализ и прогнозирование с помощью Python

Figure 2. Research methodology for studying the impact of car characteristics on its cost: analysis and forecasting using Python

Второй этап – описательная статистика – позволяет оценить основные характеристики каждого признака. На данном этапе исследуются базовые метрики. К ним относятся количество наблюдений, среднее значение, медиана, мода, стандартное отклонение, дисперсия, минимум, максимум, а также более сложные показатели, такие как квартильный размах и коэффициенты асимметрии и эксцесса [7-11]. Применительно к автомобильному рынку описательная статистика позволяет какова средняя цена автомобилей в зависимости от года выпуска, выявить, какие пробоги являются типичными, а какие – выбросами, оценить типичную мощность двигателей или объемы продаж по сегментам. Это дает возможность заранее исключить некорректные данные и сформировать более чистую и структурированную выборку для дальнейшего анализа.

Третий этап – визуализации данных – позволяет понять внутреннюю структуру данных, выявить закономерности, а также эффективно донести результаты исследования [12]. Визуализация помогает идентифицировать аномалии, выбросы, тенденции и кластеризацию данных. Наиболее часто применяемыми инструментами визуализации являются гистограммы и boxplot; диаграммы рассеяния; тепловые карты корреляции; многомерные графики, позволяющие оценить взаимодействие между несколькими признаками одновременно.

Четвертый этап предполагает формулировку и проверку гипотез о данных на основе предварительного анализа и результатов визуализации. Проверка действительности гипотез проводится на основе статистических методов. Сюда входит проверка нулевой и альтернативной гипотезы, выбор подходящего теста, установка уровня значимости и интерпретация результатов [13].

На пятом этапе после получения обобщенных характеристик выборки применяется корреляционный анализ. Это метод, позволяющий определить степень связи между признаками. Наиболее распространённый в данной области – коэффициент корреляции Пирсона, предназначенный для оценки линейной зависимости между количественными переменными. Применение корреляционного анализа на автомобильных данных позволяет выявить, например, что цена автомобиля обратно пропорциональна его пробегу (при прочих равных), положительно коррелирует с годом выпуска, мощностью двигателя или объемом. Также возможна оценка мультиколлинеарности между признаками. Например, если объем двигателя и мощность имеют высокую корреляцию, один из признаков может быть исключён из модели во избежание искажения результата.

Шестой этап заключается в прогнозировании или построении модели, позволяющей оценить не только будущую стоимость конкретного автомобиля, но и выявить закономерности, определяющие поведение цен в зависимости от сово-

купности факторов. В рамках анализа автомобильного рынка применяют как традиционные, так и современные подходы к прогнозированию. К традиционным можно отнести линейную и множественную регрессию [14], эконометрические модели (например, модели ARIMA, VAR), трендовый анализ и скользящие средние. Современные же методы включают алгоритмы машинного обучения. Это деревья решений, случайный лес (Random Forest), градиентный бустинг (Gradient Boosting), нейронные сети и другие. Последние особенно эффективны при наличии большого объема разнородных данных, что характерно для автомобильного рынка.

РЕАЛИЗАЦИЯ МЕТОДИКИ И РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

В качестве источника данных использовался датасет, доступный на платформе Kaggle [15]. Данные представлены в формате CSV и загружаются в рабочее окружение при помощи стандартных средств библиотеки Pandas [16] и библиотеки NumPy [17].

Тип данных: числовой	Symboling – страховой риск
	Wheelbase – колёсная база
	Carlength, carwidth, carheight – габариты
	Curbweight – снаряжённая масса
	Compressionratio – степень сжатия
	Enginesize – объём двигателя
	Horsepower – мощность
	Peakrpm – максимальные обороты
	Citympg, highwaympg – расход топлива (город/шоссе)
	Price – цена автомобиля
Тип данных: категориальный	CarName – название автомобиля
	Fueltype – тип топлива
	Aspiration – тип двигателя (turbo/std)
	Carbody – тип кузова (sedan, hatchback,)
	Doornumber – количество дверей (two/four)
	Enginetype – тип двигателя (ohc, ohcv, etc.)
	Cylindernumber – число цилиндров
	Fuelsystem – тип топливной системы
	Brand – производитель автомобиля (извлечён из CarName)

Рисунок 3. Структура набора данных для исследования вторичного автомобильного рынка
Figure 3. Data set structure for the study of the used car market

Для первичного знакомства с данными были выведены первые пять строк датасета с помощью метода `df.head()` и функции `display()`. С помощью метода `df.info()` была выведена информация о структуре данных анализируемого датасета. В датасете содержится 205 записей, индексируемых от 0 до 204; 26 столбцов, каждый из которых имеет определенный тип данных: 8 столбцов имеют тип `int64` (целочисленные значения), включая такие признаки, как `car_ID`, `curbweight` и `horsepower`; 8 столбцов имеют тип `float64` (числа с плавающей запятой), включая такие характеристики, как `wheelbase`, `carwidth` и `price`; 10 столбцов имеют тип

`object` (текстовые или категориальные данные), такие как `CarName`, `fueltype` и `doornumber` (рис.). Метод `info()` подтвердил, что в датасете нет пропущенных значений – каждый столбец содержит 205 ненулевых записей. Для определения уникальных значений использовался метод `unique()`. Столбец «`car_ID`» был удален, так как он является уникальным идентификатором для каждого автомобиля и не содержит информации, которая могла бы быть полезной для построения модели предсказания цен на автомобили. Так же чтобы убедиться, что датасет не содержит дублирующих записей, была выполнена проверка на дубликаты с помощью метода

uplicated()). Результат проверки показал, что в датасете нет дублирующих строк.

С помощью метода `df.describe` были получены таблицы описательной статистики для числовых (рис.4) и категориальных признаков (рис.5). Индекс страхового риска автомобиля, `symboling`, принимает значения от -2 до 3. Среднее значение составляет 0,83, что свидетельствует о преобладании автомобилей с нейтральным или умеренным уровнем риска. Минимальное значение (-2) указывает на наименьший риск, а максимальное (3) – на наибольший. Возможные значения количества дверей, `doornumber`, от 2 до 4. Преобладают модели с четырьмя дверями, две двери имеют только 25% всех автомобилей. Значения длины колёсной базы, `wheelbase`, варьируются от 86,6 до 120,9, среднее – 98,76. Более длинная колесная база характерна для полноразмерных автомобилей, обеспечивая больший комфорт и устойчивость на дороге.

Датасет содержит автомобили со средними значениями длины – 174, ширины – 66 и высоты – 54. Стандартное отклонение по длине автомобиля, равное 12, можно рассматривать как относительно низкое. Это указывает на то, что данные не сильно рассеяны и имеют тенденцию кластеризоваться вокруг нескольких конкретных значений.

описательная статистика для числовых признаков
`df.describe().T`

	count	mean	std	min	25%	50%	75%	max
<code>symboling</code>	205.0	0.834146	1.245307	-2.00	0.00	1.00	2.00	3.00
<code>doornumber</code>	205.0	3.121951	0.994966	2.00	2.00	4.00	4.00	4.00
<code>wheelbase</code>	205.0	98.756585	6.021776	86.60	94.50	97.00	102.40	120.90
<code>carlength</code>	205.0	174.049268	12.337289	141.10	166.30	173.20	183.10	208.10
<code>carwidth</code>	205.0	65.907805	2.145204	60.30	64.10	65.50	66.90	72.30
<code>carheight</code>	205.0	53.724878	2.443522	47.80	52.00	54.10	55.50	59.80
<code>curbweight</code>	205.0	2555.565854	520.680204	1488.00	2145.00	2414.00	2935.00	4066.00
<code>cylindernumber</code>	205.0	4.380488	1.080854	2.00	4.00	4.00	4.00	12.00
<code>enginesize</code>	205.0	126.907317	41.642693	61.00	97.00	120.00	141.00	326.00
<code>boreratio</code>	205.0	3.329756	0.270844	2.54	3.15	3.31	3.58	3.94
<code>stroke</code>	205.0	3.255415	0.313597	2.07	3.11	3.29	3.41	4.17
<code>compressionratio</code>	205.0	10.142537	3.972040	7.00	8.60	9.20	9.40	23.00
<code>horsepower</code>	205.0	104.117073	39.544167	48.00	70.00	95.00	116.00	288.00
<code>peakrpm</code>	205.0	5125.121951	476.985643	4150.00	4800.00	5200.00	5500.00	6600.00
<code>citympg</code>	205.0	25.219512	6.542142	13.00	19.00	24.00	30.00	49.00
<code>highwaympg</code>	205.0	30.751220	6.886443	16.00	25.00	30.00	34.00	54.00
<code>price</code>	205.0	13276.710571	7988.852332	5118.00	7788.00	10295.00	16503.00	45400.00

Рисунок 4. Описательная статистика для числовых признаков

Figure 4. Descriptive statistics for numerical features

Диапазон значений снаряжённой массы автомобиля, `curbweight`, измеренной в фунтах составляет от 1 488 до 4 066 при среднем значении 2 555,57. Вес тесно связан с мощностью, типом кузова и уровнем оснащения автомобиля. Среднее количество цилиндров в двигателе, `cylindernumber`, равно 4,38 при диапазоне от 3 до 12. Преобладают 4-х цилиндровые двигатели, что

характерно для массового сегмента. Модели с 6,8 или более цилиндрами, относятся к более мощным и дорогим категориям. Размер двигателя, `enginesize`, анализируемых автомобилей находится в пределах от 61 до 326. Четверть авто (25%) имеют двигатель 97 или меньше; 50% – 120 или меньше и 75% – 141 или меньше. Отношение диаметра цилиндра к ходу поршня, `boreratio`, составляет от 2,54 до 3,94 при среднем значении 3,33. Этот показатель влияет на геометрию двигателя и может отражать различия в конструкции между экономичными и спортивными моделями. Среднее значение коэффициента сжатия двигателя, `compressionratio`, – 10,14. Высокие значения могут свидетельствовать о наличии турбонаддува или об особенностях конструкции, направленных на повышение эффективности. Важный показатель производительности автомобиля – мощность двигателя (`horsepower`), – напрямую влияет на его цену и рыночное позиционирование. Находится в диапазоне – от 48 до 288 в лошадиных силах. Диапазон значений топливной экономичности в городских условиях, `citympg` (миль на галлон), составляет от 13 до 49. Более высокие значения указывают на большую эффективность, характерную для гибридных или малолитражных автомобилей. И, наконец, средняя стоимость автомобиля (`price`) равна 13 276,71\$, цена находится в диапазоне от 5 118\$ до 45 400\$, медиана – 10 295\$. Наличие дорогих моделей смещает распределение вправо, увеличивая среднее значение. Диапазон цен подтверждает разнородность выборки.

Результаты описательной статистики категориальных признаков показывают, что всего представлено 147 уникальных наименований авто из 205 наблюдений, что подтверждает широкий ассортимент автомобилей. Часто встречающаяся модель – «toyota corona»; преобладающий тип топлива – gas (бензин).

описательная статистика для категориальных признаков
`df.describe(include='object').T`

	count	unique	top	freq
<code>CarName</code>	205	147	toyota corona	6
<code>fueltype</code>	205	2	gas	185
<code>aspiration</code>	205	2	std	168
<code>carbody</code>	205	5	sedan	96
<code>drivewheel</code>	205	3	fwd	120
<code>enginelocation</code>	205	2	front	202
<code>enginetype</code>	205	7	ohc	148
<code>fuelsystem</code>	205	8	mpfi	94

Рисунок 5. Описательная статистика для категориальных признаков

Figure 5. Descriptive statistics for categorical features

Система подачи воздуха характеризуется (стандартным впуском (std) и турбонаддувом (turbo). В 168 случаях используется стандартная система, что типично для массовых автомобилей. В датасете представлены пять типов кузова. Наиболее распространён – sedan (96 наблюдений), что соответствует популярности данного типа на рынке. Присутствуют авто с передним (fwd) и задним (rwd) приводом колес, а также полноприводные (4wd). Передний привод встречается чаще всего в 120 случаях. Почти все автомобили имеют переднее расположение двигателя (202 случая), что объясняется распространённой компоновкой передний привод – передний двигатель. Всего выделено 7 уникальных типов двигателей, из которых наиболее часто встречается ohc (верхнеклапанный двигатель) – 148 случаев. Автомобили имеют 8 различных типов систем подачи топлива, fuelsystem, наиболее популярная – mpfi (многоточечный впрыск топлива) с частотой 94.

Для более точной и содержательной визуализации данных была проведена логическая классификация признаков по функциональным группам. Это позволило выделить смысловые блоки, отражающие различные аспекты характеристик автомобилей:

1. основные физические характеристики автомобиля;
2. типы и конструкционные особенности;
3. показатели работы двигателя;
4. технические характеристики двигателя;
5. финансовые показатели.

Для анализа физических параметров автомобилей была выполнена визуализация пяти числовых признаков, отражающих габаритные размеры и массу транспортных средств: wheelbase, carlength, carwidth, carheight, curbweight. Визуализация реализована с помощью библиотеки matplotlib, а именно – с использованием метода plot() для построения линейных графиков с маркерами. Для каждого признака строился отдельный подграфик (subplot) (рис.6). Визуализация основных физических характеристик автомобиля позволяет наглядно оценить диапазон, разброс и потенциальные выбросы по каждому из физических признаков. Например, по графику curbweight можно увидеть значительные различия в массе автомобилей, что может быть связано с различиями в конструкции, классе и типе автомобиля. Аналогично, значения carlength и wheelbase варьируются в широком диапазоне, что указывает на присутствие в выборке как компактных, так и полноразмерных моделей.

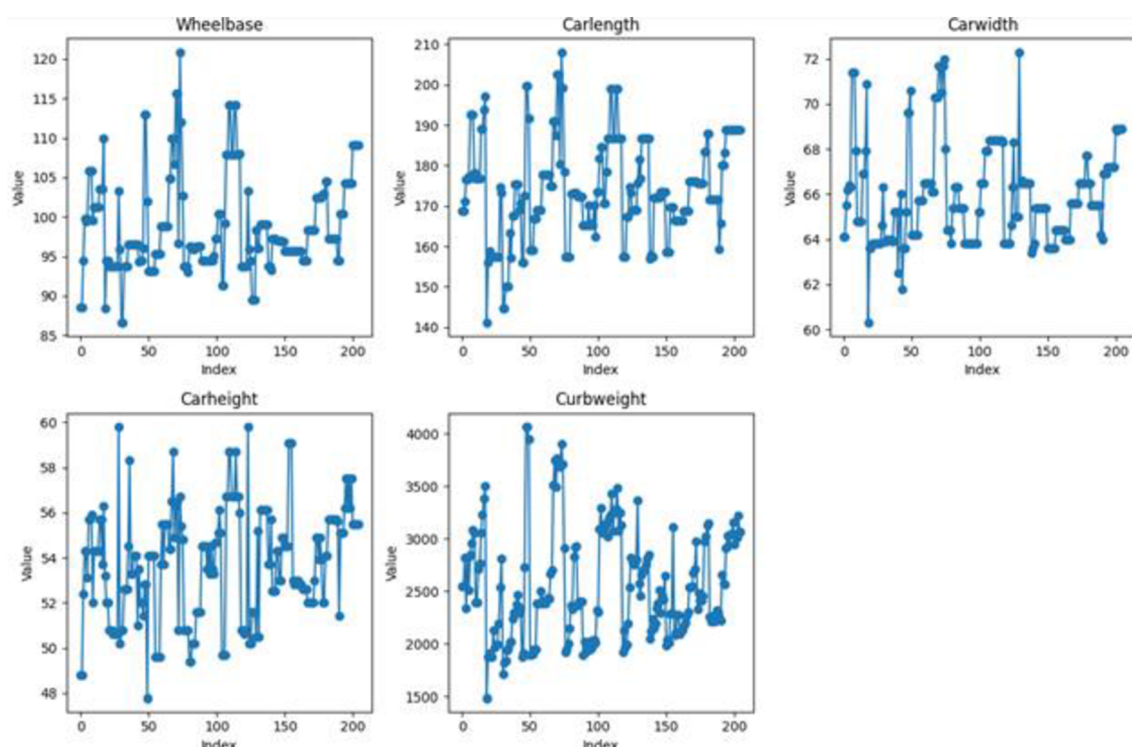


Рисунок 6. Визуализация основных физических характеристик автомобиля
Figure 6. Visualization of the main physical characteristics of a car

Для оценки состава выборки по производителям автомобилей и визуализации структуры конструкционных особенностей была выполнена предварительная обработка текстового признака

CarName, содержащего полное наименование автомобиля. Из полного наименования автомобиля извлекается первая строка, соответствующая бренду и подсчитывается частота появления каждого

бренда в выборке. Столбчатая диаграмма показывает 28 различных брендов автомобилей (рис.7). Значительное количество наблюдений приходится на бренд Toyota (более 30 автомобилей), а также на Nissan, Mazda, Honda и Mitsubishi, что указывает на доминирование японских автопроизводителей в

наборе данных. Наибольшую долю в выборке занимают автомобили с кузовом типа sedan (более 45%). Далее следуют hatchback и wagon. Менее распространены convertible и hardtop, каждые из которых занимают менее 10%. Для этого использовался метод bar библиотека Plotly.

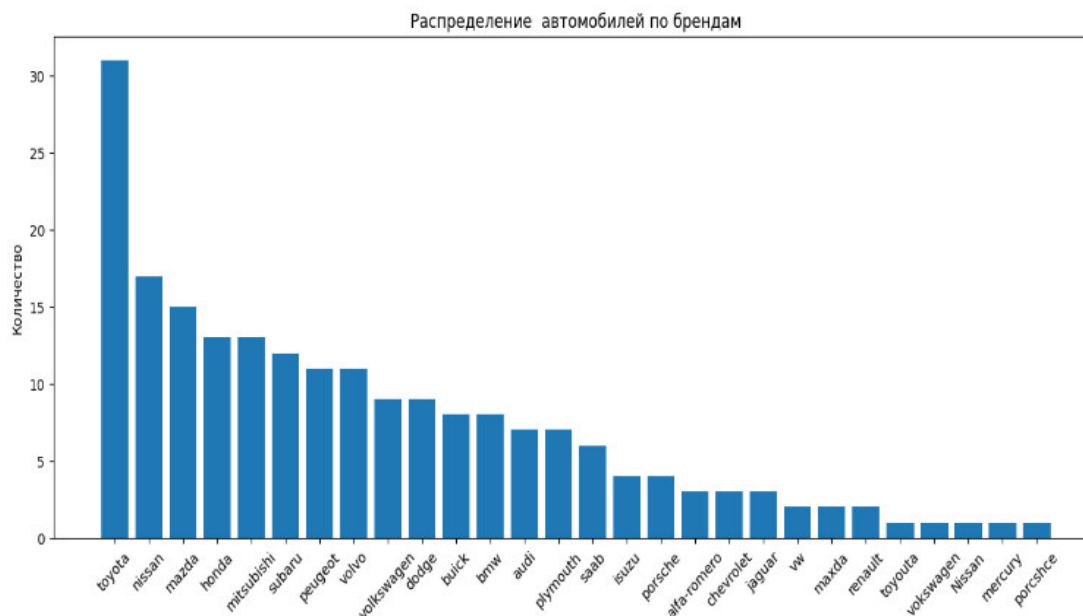


Рисунок 7. Распределение автомобилей по брендам
Figure 7. Distribution of cars by brands

Для анализа конструкционных особенностей автомобилей была построена круговая диаграмма, отражающая распределение выборки по типу кузова (carbody). Это позволяет оценить, какие кузовные конфигурации доминируют в датасете и, следовательно, могут существенно влиять на итоговую стоимость автомобиля. На диаграмме (рис. 8) видно, что наибольшую долю в выборке занимают автомобили с кузовом типа sedan (более 45%). Далее следуют hatchback и wagon. Менее распространены convertible и hardtop, каждые из которых занимают менее 10%.

С целью анализа конструктивных и топливных характеристик автомобилей были построены круговые диаграммы, отображающие долю различных типов двигателей, видов топлива и топливных систем в выборке (рис. 9). Наиболее распространённый тип – ohc (верхнеклапанный), на него приходится более 70% автомобилей. Это объясняется его универсальностью и распространённостью в автомобилях среднего класса. Абсолютное доминирование бензиновых двигателей (gas) – около 90%. Дизельные модели представлены незначительно, что характерно для легкового сегмента в США. Абсолютное доминирование бензиновых двигателей (gas) – около 90%. Наибольшую долю занимает система mpfi (многоточечный впрыск топлива). Другие типы

(например, 2bbl, 1bbl) встречаются реже и связаны с менее современными конструкциями.

График анализа средней цены по брендам позволяет визуально оценить ценовую политику различных производителей и подтвердить наличие различий между массовым и премиальным сегментами (рис. 10). Диаграмма показывает, что наибольшую среднюю цену имеют автомобили брендов BMW, Porsche и Jaguar, что соответствует их позиционированию как премиальных. В то же время бренды Subaru, Chevrolet и Dodge демонстрируют самые низкие средние цены, что указывает на их принадлежность к бюджетному сегменту. Для этого использовался метод barplot библиотеки Seaborn.

Для анализа показателей работы двигателя был построен график рассеяния, отражающий соотношение между переменными price и highwaympg, price и его расходом топлива в городских условиях citympg. Анализ графиков рассеяния демонстрирует схожую закономерность (рис. 11). В обоих случаях наблюдается обратная зависимость: при увеличении топливной экономичности (то есть при росте значения MPG) цена автомобиля, как правило, снижается. Это означает, что более дорогие автомобили чаще всего оказываются менее экономичными как в городском режиме, так и на шоссе.

Структура автомобилей по типу кузова (carbody)

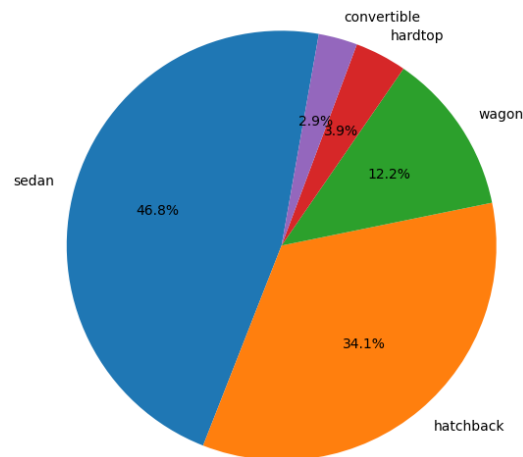


Рисунок 8. Круговая диаграмма структуры автомобилей по типам кузова
Figure 8. Pie chart of car structure by body configuration

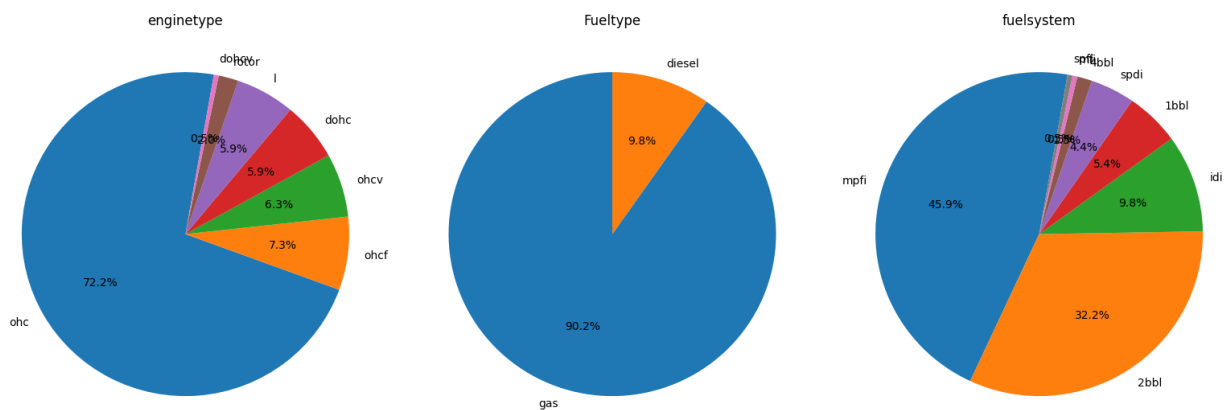


Рисунок 9. Круговая диаграмма распределения автомобилей по типу двигателя, виду топлива и топливных систем
Figure 9. Pie chart of vehicle distribution by engine type, fuel type, and fuel system

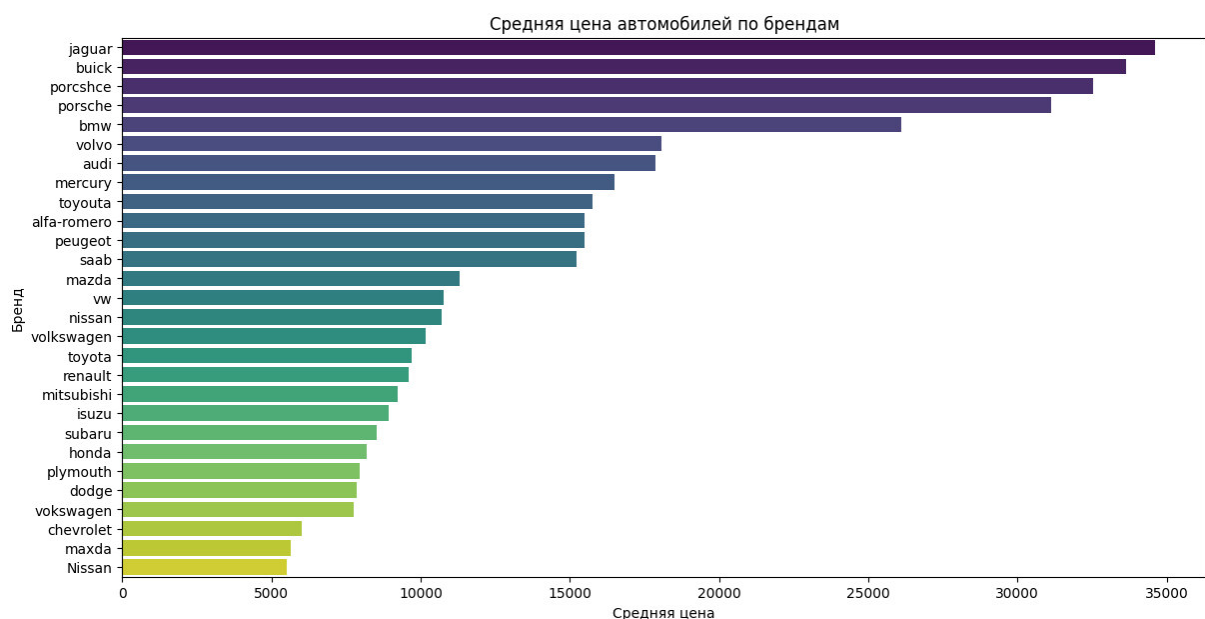


Рисунок 10. Распределение средней цены по брендам автомобилей
Figure 10. Distribution of average prices by car brand

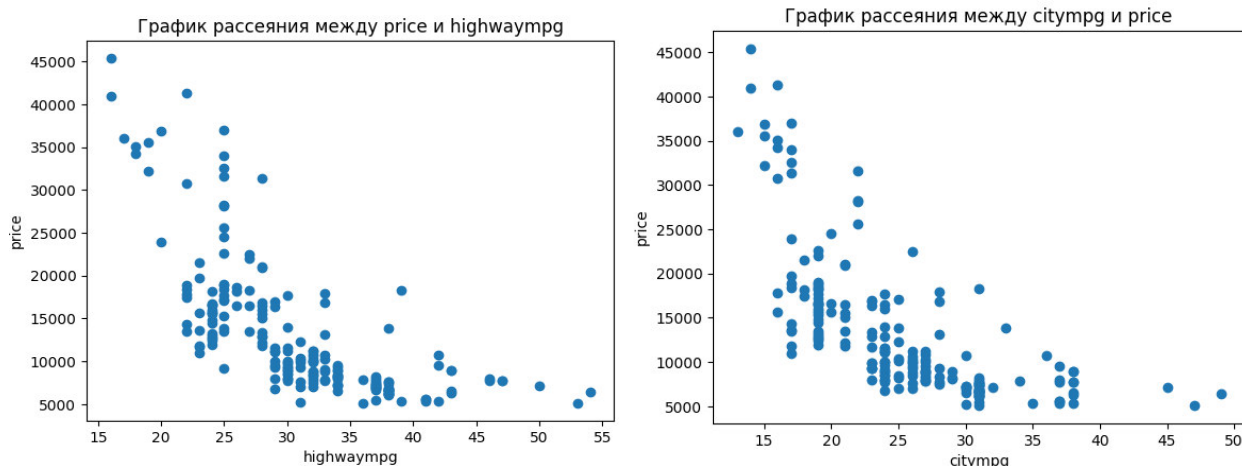


Рисунок 11. Визуализация показателей работы двигателя
Figure 11. Visualization of engine performance indicators

Для визуального анализа распределения цен автомобилей было построено два графика: линейный график значений Price и гистограмма значений столбца Price. Линейный график позволяет визуально отследить изменение цен по каждой записи в датасете, что удобно для выявления выбросов – особенно дорогих автомобилей. Он показывает, что в выборке присутствуют резкие скачки цены, что может указывать на премиальные модели. Гистограмма, в свою очередь, даёт более обобщённую картину распределения: она демонстрирует, что наибольшее количество автомобилей сосредоточено в ценовом диапазоне до 15 000 долларов, а высокие значения цены встречаются значительно реже (рис. 12). Такая картина распределения цен вполне соответствует действительности. Стоимость автомобиля зависит от множества факторов, включая бренд, модель, технические характеристики, год выпуска, комплектацию, состояние и другие параметры. Поэтому наличие как доступных, так и пре-

миальных автомобилей в выборке отражает реальную структуру рынка и объясняет широкий разброс значений переменной Price.

В ходе исследования были сформулированы и статистически проверены шесть гипотез (рис. 13).

По результатам статистического анализа проверки гипотез были сделаны выводы:

- 1) средние цены на бензиновые и дизельные автомобили не отличаются;
- 2) наличие турбонаддува является фактором, оказывающим влияние на увеличение стоимости автомобиля;
- 3) статистически значимых различий в цене между автомобилями с двумя и четырьмя дверями не обнаружено;
- 4) значимых различий в средней цене между седанами и другими типами автомобилей не выявлено;
- 5) разные системы подачи топлива статистически значимо влияют на среднюю цену автомобиля;
- 6) автомобили с более высоким расходом обычно имеют более низкие цены.

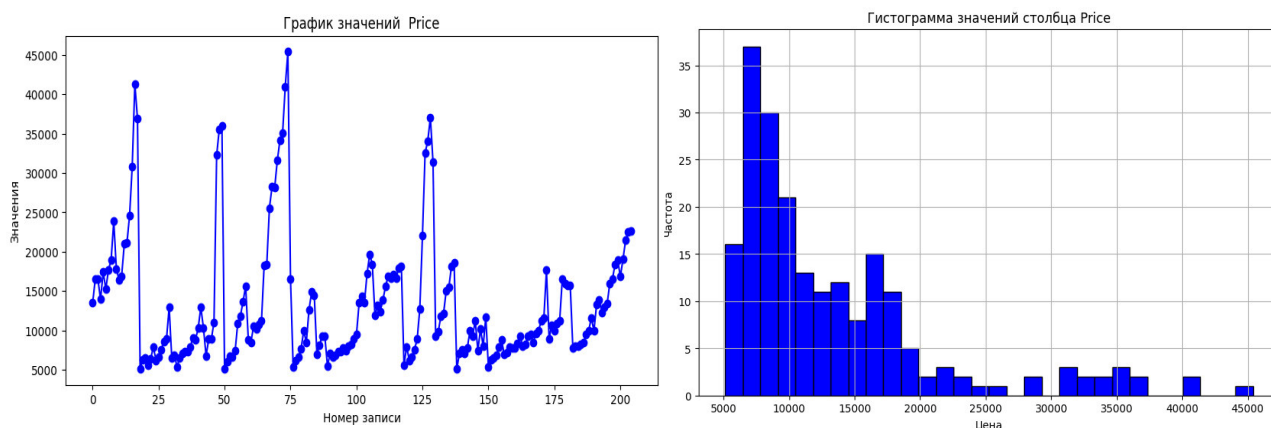


Рисунок 12. Визуальный анализ по показателю Price (Цена)
Figure 12. Visual analysis based on Price

Гипотеза 1: влияние типа топлива (fueltype) на цену автомобиля	•средние цены на бензиновые и дизельные автомобили не отличаются. •t-статистика: -1.514, •P-значение: 0.131
Гипотеза 2: автомобили с турбонаддувом (turbo) имеют более высокую цену, чем автомобили с атмосферным двигателем (std)	•Средняя цена автомобилей с турбонаддувом: 16298.1 •Средняя цена автомобилей с атмосферным двигателем: 12611.2 •t-статистика: 2.576 •p-значение: 0.0107 •p-значение меньше 0.05, значит, есть статистически значимое различие в цене между автомобилями с турбонаддувом и атмосферным двигателем
Гипотеза 3: автомобили с двумя дверями стоят дешевле, чем автомобили с четырьмя дверями	•t-статистика: -0.444 •P-значение: 0.657 •Не удалось отвергнуть нулевую гипотезу: нет статистических доказательств, что автомобили с двумя дверями стоят дешевле.
Гипотеза 4 : средняя цена седанов (sedan) отличается от средней цены других типов кузова (hatchback, wagon, coupe, convertible)	•t-statistic: 1.805, •p-value: 0.072 •Гипотеза не подтверждена: средняя цена седанов статистически значимо не отличается от других типов автомобилей.
Гипотеза 5: разные системы подачи топлива (mpi, mfi, spdi и т.д.) влияют на цену автомобиля	•так как $p(PR(>F)) < 0.5$, то можно утверждать, что разные системы подачи топлива оказывают значительное влияние на цену автомобиля
Гипотеза 6: автомобили с более экономичным расходом топлива (высокие значения citympg и highwaympg) оцениваются выше	•Корреляция между citympg и price: -0.685 •Корреляция между highwaympg и price: -0.697 •Существует сильная обратная корреляция между расходом топлива в городе (citympg) и ценой автомобиля, а также между расходом топлива на шоссе (highwaympg) и ценой автомобиля.

Рисунок 13. Проверка гипотез
Figure 13. Hypothesis testing

Корреляционный анализ проводился на основе расчета коэффициента Пирсона между целевой переменной Price и числовыми признаками (рис.14). Наиболее сильные положительные корреляции были зафиксированы у следующих признаков: ширина кузова (carwidth) – 0,759, снаряжённая масса (curbweight) – 0,835, количество цилиндров (cylindernumber) – 0,718, объём двигателя (enginesize) – 0,874, мощность двигателя (horsepower) – 0,808. С технической точки зрения это объясняется тем, что автомобили с более крупными габаритами и мощными двигателями, как правило, относятся к более дорогим сегментам. Также наблюдаются существенные положительные корреляции со следующими признаками: привод (drivewheel) – 0,578, колёсная база (wheelbase) – 0,578, длина автомобиля (carlength) – 0,683, тип топливной системы (fuelsystem) – 0,527, отношение бореративности (boreratio) – 0,553. Наряду с этим была выявлена устойчивая отрицательная корреляция между расходом топлива и ценой автомобиля, как в условиях городской езды (citympg – 0,685), так и на шоссе (highwaympg – 0,697). Это подтверждает, что более экономичные автомобили

чаще относятся к доступному ценовому сегменту, в то время как премиальные модели зачастую потребляют больше топлива.

Таким образом, корреляционный анализ позволил количественно подтвердить ранее выявленные зависимости и обосновать выбор признаков, которые будут использоваться при построении моделей прогнозирования стоимости автомобиля.

В качестве признаков для множественной линейной регрессии были отобраны 6 признаков, среди них enginesize (размер двигателя), curbweight (снаряжённая масса), horsepower (лошадиные силы), carwidth (ширина автомобиля), citympg (расход топлива в городе), highwaympg (расход топлива на шоссе). Для повышения устойчивости и корректности прогностической модели была проведена оценка мультиколлинеарности между ключевыми признаками (рис.13). Проведённый анализ показал, что все отобранные признаки, выбранные для построения модели прогнозирования цены, демонстрируют высокую степень взаимной корреляции, то есть между ними присутствует выраженная мультиколлинеарность.

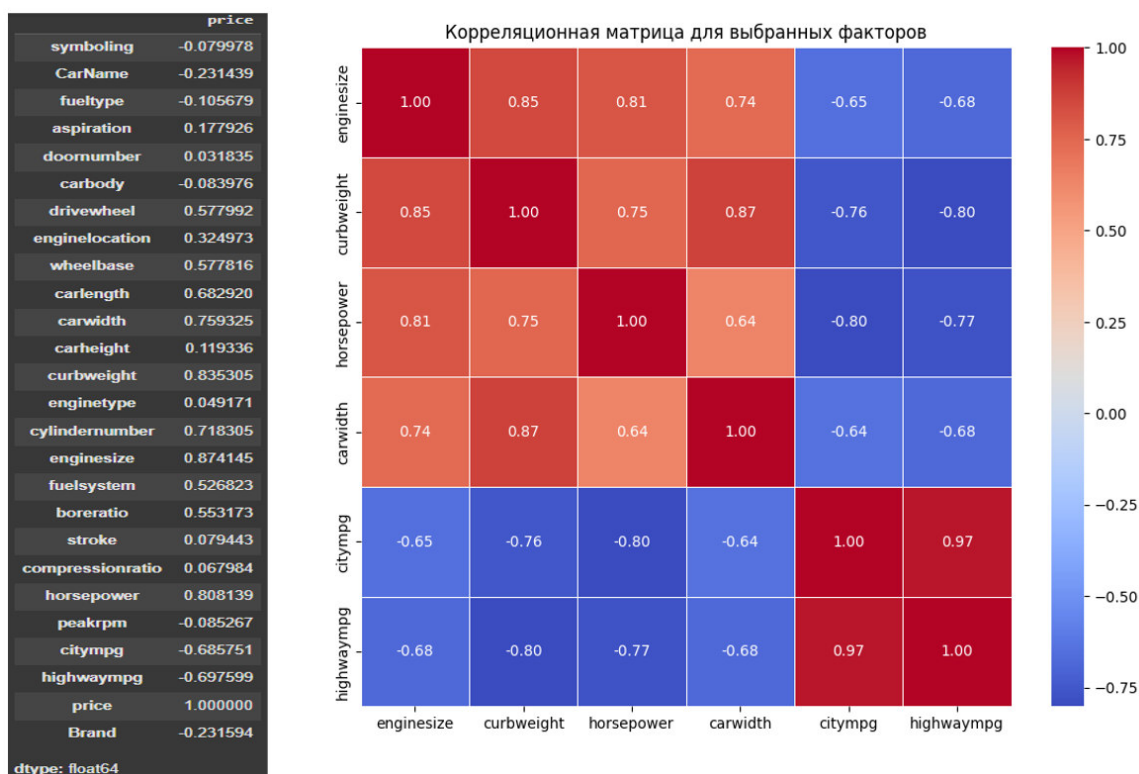


Рисунок 14. Корреляционный анализ взаимосвязей между показателями
Figure 14. Correlation analysis of the relationships between indicators

Поэтому для оценки влияния отдельных факторов на цену автомобиля была реализована серия однофакторных линейных регрессионных моделей [18]. В качестве признаков были выбраны параметры, ранее выявленные как значимые в корреляционном анализе:

- 1) размер двигателя (enginesize);
- 2) снаряжённая масса (curbweight);
- 3) мощность (horsepower);
- 4) ширина автомобиля (carwidth);
- 5) расход топлива в городе (citympg);
- 6) расход топлива на шоссе (highwaympg).

По значениям коэффициента детерминации (R^2), средней квадратичной ошибки между предсказанными и фактическими значениями (MSE) и средней абсолютной ошибки (MAE), отражающих качество моделей, можно определить сильную предсказательную способность первых четырех моделей (таб.).

Таблица 1. Метрики для оценки однофакторных линейных регрессионных моделей
Table 1. Metrics for evaluating single-factor linear regression models

№п/п	Фактор	R^2	MSE	MAE
1.	enginesize	0,80	15465458.96	2748.69
2.	curbweight	0,75	19668972.59	2954.45
3.	horsepower	0,63	29003009.23	3601.91
4.	carwidth	0,64	28398987.75	3131.68
5.	citympg	0,35	51189990.64	5342.41
6.	highwaympg	0,41	46701836.65	5122.79

Первые четыре модели позволяют достаточно точно прогнозировать стоимость автомобиля. Признаки, связанные с топливной экономичностью (citympg и highwaympg), показали более слабую связь с ценой: citympg ($R^2 \approx 0,35$) и highwaympg ($R^2 \approx 0,41$). Это объясняется тем, что экономичность, хотя и важна, уступает влиянию физических параметров, напрямую связанных с классом и мощностью автомобиля.

На графиках визуально подтверждается (рис.15), что модели, основанные на enginesize и curbweight, строят предсказания, максимально приближённые к реальным значениям. Однофакторные регрессионные модели позволяют выявить наиболее значимые индивидуальные признаки, влияющие на цену автомобиля. Наиболее информативными оказались физические и силовые характеристики машины.

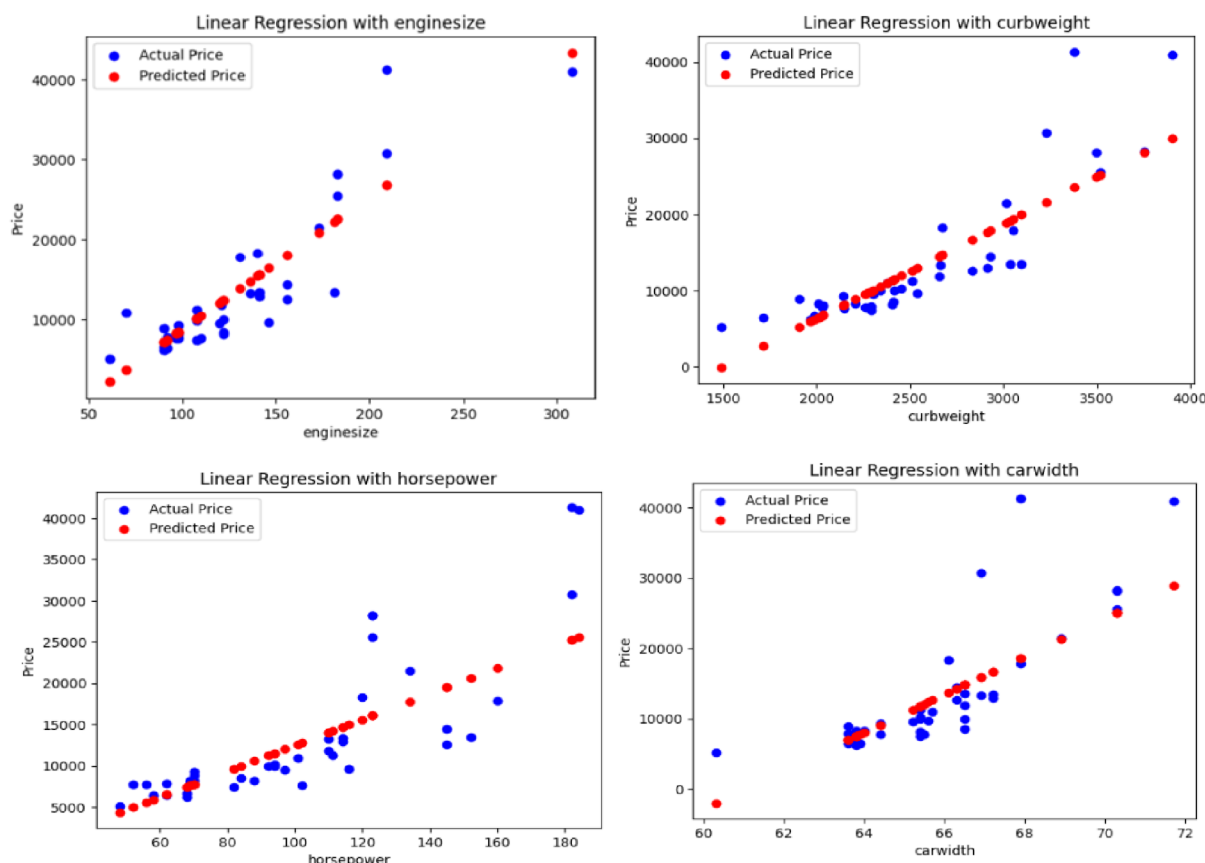


Рисунок 15. Однофакторные регрессионные модели
Figure 15. Single-factor regression models

ЗАКЛЮЧЕНИЕ

Таким образом, прогнозирование цен на вторичном автомобильном рынке, характеризующегося значительным объемом, разнообразием представленных транспортных средств повышает прозрачность рыночных процессов. В условиях цифровизации экономики и повсеместного внедрения аналитических решений, прогнозирование становится неотъемлемой частью эффективного управления как для бизнеса, так и для конечного потребителя. Разработка и внедрение прогностических

моделей на базе Python и соответствующих библиотек анализа данных позволяет не только автоматизировать процесс оценки стоимости автомобилей, но и повысить его точность, надежность и адаптивность к изменяющимся условиям.

Авторы заявляют об отсутствии конфликта интересов.

The authors declare no conflict of interest.

ЛИТЕРАТУРА

1. Рынок легковых автомобилей с пробегом в России в 2024 году и в декабре. <https://www.autostat.ru/press-releases/59289/>.
2. Миролюбова А.А., Сяотун Я., Ворошин Д.А. Анализ макроэкономических показателей Китая с использованием Python. Сборник научных трудов вузов России "Проблемы экономики, финансов и управления производством". 2025. № 56. С. 120-127. EDN ZHZSDW.
3. Миролюбова А.А., Ксенофонтова О.Л., Ворошин Д.А. Опыт интеллектуального анализа и прогнозирования рынка акций. Известия высших учебных заведений. Серия «Экономика, финансы и управление производством» [Ивэкофин]. 2024. № 4(62). С. 55-63. DOI: 10.6060/ivecofin.2024624.702. EDN CQZLPD.
4. Сун Я. Построение генетического алгоритма поиска нейросетевых моделей для прогнозирования курса юаня. В сб. «Дни науки в ИГХТУ» Тезисы докладов Всероссийской школы-конференции молодых ученых. Иваново: ИГХТУ. 2023. С. 564. EDN AHMLJW.

REFERENCES

1. The used car market in Russia in 2024 and in December. <https://www.autostat.ru/press-releases/59289/>. (in Russian).
2. Mirolubova A.A., Xiaotong Ya., Voroshin D.A. Analysis of China's Macroeconomic Indicators Using Python. *Collection of Scientific Papers of Russian Universities "Problems of Economics, Finance, and Production Management"*. 2025. N 56. P. 120-127. EDN ZHZSDW. (in Russian).
3. Mirolubova A.A., Ksenofontova O.L., Voroshin D.A. Experience in data mining and forecasting of the stock market. *Ivecofin*. 2024. N 4(62). P. 55-63. DOI: 10.6060/ivecofin.2024624.702. EDN CQZLPD. (in Russian).
4. Sun Ya. Construction of a genetic algorithm for searching neural network models for predicting the yuan exchange rate. *Materials of the All-Russian School-Conference of Young Scientists «Science Days at ISUCT»*. Ivanovo: ISUCT. 2023. P. 564. EDN AHMLJW. (in Russian).

5. 10 библиотек Python для автоматического разведочного анализа данных. <https://uproger.com/10-bibliotek-python-dlya-avtomaticheskogo-analiza/>.
6. Анализ данных с использованием Python. <https://habr.com/ru/articles/353050/>.
7. **Нильсен Эйлин.** Практический анализ временных рядов: прогнозирование со статистикой и машинное обучение. СПб.: ООО "Диалектика". 2021. 544 с.
8. **Куленцан А.Л., Марчук Н.А., Гвоздев В.В., Ширяев М.Ю., Пузанов А.М.** Статистическое исследование и прогнозирование численности браков и разводов на территории РФ. *Современные наукоемкие технологии. Региональное приложение.* 2023. № 3(75). С. 107-112. DOI: 10.6060/snt.20237503.00014. EDN PRXCJP.
9. **Астраханцева И.А., Герасимов А.С., Смирнова О.П.** Оценка применимости статистических и машинных моделей для прогнозирования инфляции. *Современные наукоемкие технологии. Региональное приложение.* 2024. № 3(79). С. 120-131. DOI: 10.6060/snt.20247903.0019. EDN SJSWAG.
10. **Толкачева О.П.** Экономико-статистический анализ влияния рынка труда на экономическую безопасность страны. *Известия высших учебных заведений. Серия «Экономика, финансы и управление производством» [Ивэкофин].* 2023. № 3(57). С. 59-69. DOI: 10.6060/iveco-fin.2023573.654. EDN QTYLMZ.
11. **Бобков С.П., Астраханцева И.А., Гушин А.А., Бобкова Е.С., Астраханцев Р.Г., Шутов Д.А.** Использование дискретного вероятностного подхода для моделирования проточных трубчатых реакторов. *Известия высших учебных заведений. Серия Химия и химическая технология.* 2025. Т. 68. № 2. С. 96-101. DOI: 10.6060/ivkkt.20256802.7109. EDN OWCXJD.
12. Топ 6 библиотек Python для визуализации: какую и когда лучше использовать? <https://habr.com/ru/companies/otus/articles/558478/>.
13. **Гмурман В.Е.** Руководство к решению задач по теории вероятностей и математической статистике: учеб. пособие. М.: Высшая школа. 2004. 404 с.
14. Тихонов С. Линейная регрессия в Python. <https://education.yandex.ru/handbook/data-analysis/article/pandan-linejnaya-regressiya-v-python>.
15. CarPrice_Assignment.csv. <https://www.kaggle.com/datasets/hellbuoy/car-price-prediction>.
16. Руководство по использованию Python-библиотеки Pandas. http://pandas.geekwriter.ru/getting_started/index.html.
17. Руководство по использованию Python-библиотеки Numpy. <https://pythonru.com/biblioteki/rukovodstvo-po-ispolzovaniju-python-biblioteki-numpy>.
18. Регрессионные модели в Python. <https://blog.skillfactory.ru/regressionnye-modeli-v-python/>.
5. 10 Python libraries for automatic data exploration analysis. <https://uproger.com/10-bibliotek-python-dlya-avtomaticheskogo-analiza/>. (in Russian).
6. Data analysis using Python. <https://habr.com/ru/articles/353050/>. (in Russian).
7. **Nielsen Aileen.** Practical Time Series Analysis: Prediction with Statistics and Machine Learning. St. Petersburg: Dialectics LLC. 2021. 544 p. (in Russian).
8. **Kulenzan A.L., Marchuk N.A., Gvozdev V.V., Shiryayev M.Yu., Puzanov A.M.** Statistical research and forecasting of the number of marriages and divorces in the territory of the Russian Federation. *Modern high technologies. Regional application.* 2023. N 3(75). P. 107-112. DOI: 10.6060/snt.20237503.00014. EDN PRXCJP. (in Russian).
9. **Astrakhantseva I.A., Gerasimov A.S., Smirnova O.P.** Evaluating statistical and machine learning models for inflation forecasting. *Modern high technologies. Regional application.* 2024. N 3 (79). P. 120-131. DOI: 10.6060/snt.20247903.0019. EDN SJSWAG. (in Russian).
10. **Tolkacheva O.P.** Economic and statistical analysis of the labor market impact on the economic security of the country. *Ivecofin.* 2023. N 3(57). P. 59-69. DOI: 10.6060/ivecofin.2023573.654. EDN QTYLMZ. (in Russian).
11. **Bobkov S.P., Astrakhantseva I.A., Gushchin A.A., Bobkova E.S., Astrakhantsev R.G., Shutov D.A.** Using a discrete probabilistic approach for simulating flow tubular reactors. *ChemChemTech.* 2025. Vol. 68. N 2. P. 96-101. DOI: 10.6060/ivkkt.20256802.7109. EDN OWCXJD. (in Russian)
12. Top 6 Python Libraries for Visualization: Which One to Use and When? <https://habr.com/ru/companies/otus/articles/558478/>. (in Russian).
13. **Gmurman V.E.** Guide to Solving Problems in Probability Theory and Mathematical Statistics: textbook. Moscow: Vysshaya shkola. 2004. 404 p. (in Russian).
14. Tikhonov S. Linear Regression in Python. <https://education.yandex.ru/handbook/data-analysis/article/pandan-linejnaya-regressiya-v-python>. (in Russian).
15. CarPrice_Assignment.csv. <https://www.kaggle.com/datasets/hellbuoy/car-price-prediction>.
16. Guide to using Python libraries Pandas. http://pandas.geekwriter.ru/getting_started/index.html. (in Russian).
17. Guide to using the Python library Numpy. <https://pythonru.com/biblioteki/rukovodstvo-po-ispolzovaniju-python-biblioteki-numpy>. (in Russian).
18. Regression models in Python. <https://blog.skillfactory.ru/regressionnye-modeli-v-python/>. (in Russian).

Поступила в редакцию 22.09.2025
Принята к опубликованию 06.10.2025

Received 22.09.2025
Accepted 06.10.2025