

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
Национальный исследовательский технологический университет «МИСИС»
Институт информационных технологий и автоматизированных систем управления
Кафедра Бизнес-информатики и систем управления производством

Практическая работа №4

по дисциплине «Статистические методы анализа данных в принятии решений»
на тему «Статистический анализ данных в принятии решений»

Направление подготовки
38.03.05 Бизнес-информатика
Семестр 4

Выполнил:

Сычиков Владимир Андреевич

(ФИО студента)

ББИ-23-6

(№ группы)

11.03.2025

(дата сдачи)

Подпись:

Проверил:

(ФИО преподавателя)

(оценка)

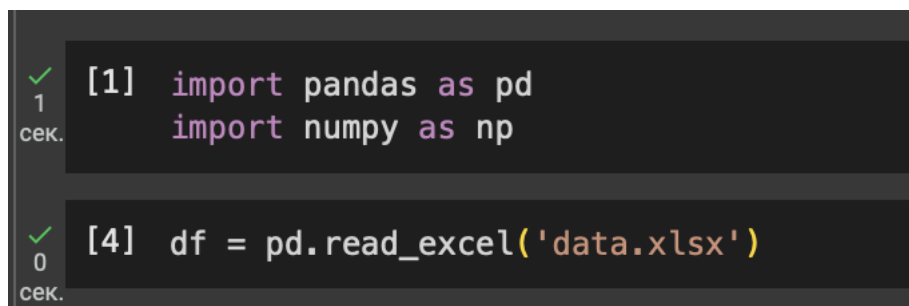
(дата проверки)

Подпись:

Москва – 2025

Ход работы:

Для начала работы я импортировал все необходимые библиотеки(pandas, numpy) с помощью команды `import`, а также задал элиасы для удобства обращения к библиотеке. Далее я подгрузил выборку и записал ее в датафрейм `df`.

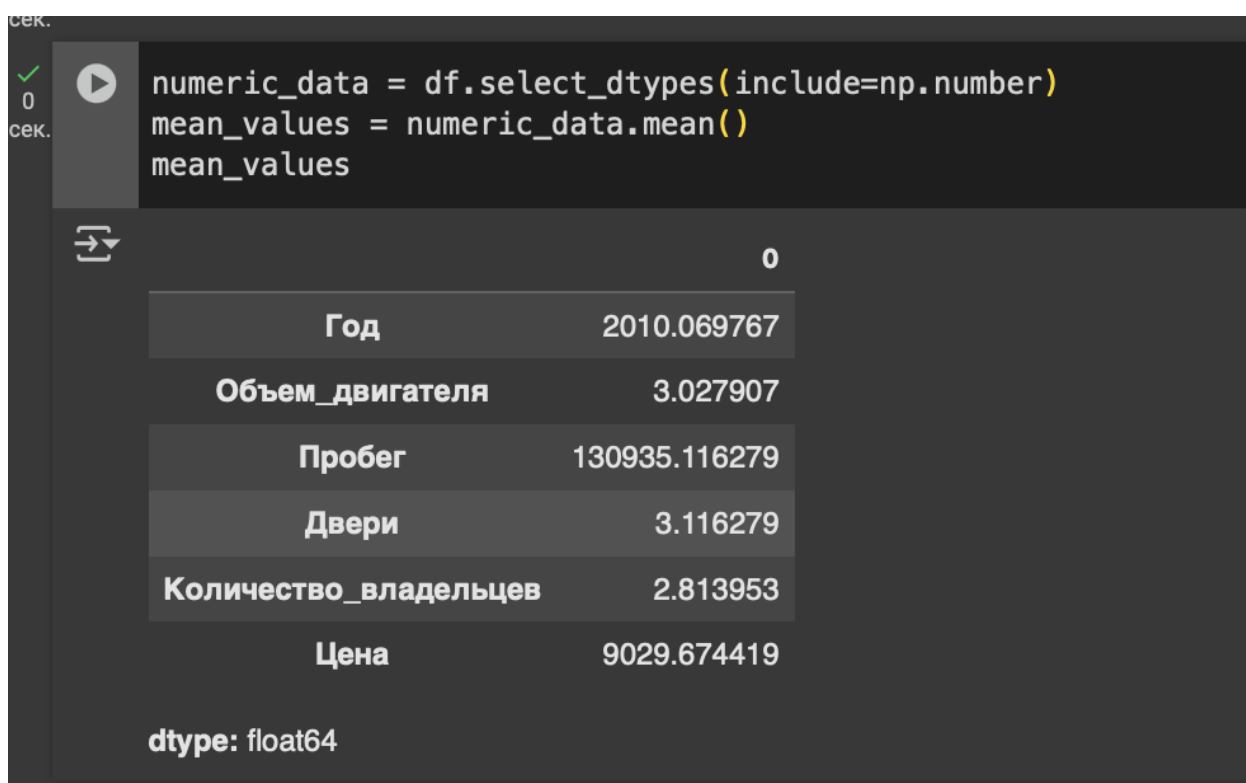


```
[1] import pandas as pd
import numpy as np

[4] df = pd.read_excel('data.xlsx')
```

Рисунок 1 - Импорт библиотек

На рисунке 2 реализуется процесс вычисления математического ожидания для числовых значений. Предварительно мы отбираем все числовые столбцы с помощью специализированной функции `.number()` путем обращения к библиотеке `numpy`. Далее мы вычисляем математическое ожидание с помощью функции `.mean()`. Результатом станет столбец с параметром и соответствующее значение.



```
numeric_data = df.select_dtypes(include=np.number)
mean_values = numeric_data.mean()
mean_values
```

	0
Год	2010.069767
Объем_двигателя	3.027907
Пробег	130935.116279
Двери	3.116279
Количество_владельцев	2.813953
Цена	9029.674419

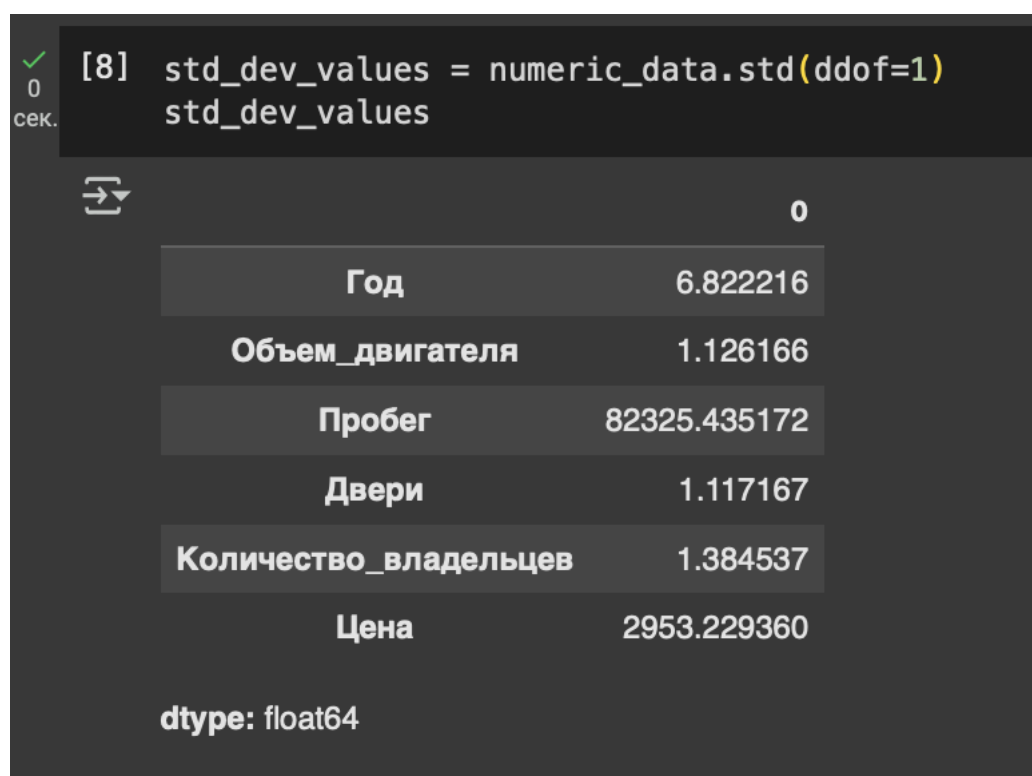
dtype: float64

Рисунок 2 – Процесс получения математического ожидания

Математическое ожидание года выпуска автомобилей — около 2010 года, что указывает на их относительную новизну. Средний объем двигателя — 3.03 литра, что характерно для автомобилей среднего и большого класса. Средний пробег — около 130 935

км, что говорит об активном использовании. В среднем у автомобилей 3 двери, а количество владельцев — примерно 2.81, что может свидетельствовать о нескольких перепродажах. Средняя цена — около 9029.67 единиц, что указывает на доступность. В целом, данные показывают преобладание автомобилей среднего возраста с умеренным пробегом и средней ценой.

На рисунке 3 представлен процесс вычисления стандартного отклонения для нескольких столбцов из DataFrame. Код вычисляет стандартные отклонения для столбцов 'Год', 'Объем_двигателя', 'Пробег', 'Двери', 'Количество_владельцев' и 'Цена'. Для каждого столбца используется функция `.std()` с параметром `ddof=1`, которая возвращает стандартное отклонение значений в столбце. При получении результата используется датасет с только числовыми значениями `numeric_data`. Результатом станет столбец с параметром и соответствующее значение.



The screenshot shows a Jupyter Notebook cell with the following code and output:

```
[8] std_dev_values = numeric_data.std(ddof=1)
std_dev_values
```

The output is a Series with the following values:

	0
Год	6.822216
Объем_двигателя	1.126166
Пробег	82325.435172
Двери	1.117167
Количество_владельцев	1.384537
Цена	2953.229360

The dtype is float64.

Рисунок 3 – Процесс получения стандартного отклонения значений

На основе представленных данных, где использована функция для вычисления стандартного отклонения, можно сделать следующие выводы. Стандартное отклонение года выпуска автомобилей составляет примерно 6.82 года, что указывает на умеренный разброс в возрасте автомобилей. Стандартное отклонение объема двигателя — 1.13 литра, что говорит о некоторой вариативности в размерах двигателей. Пробег автомобилей имеет стандартное отклонение около 82 325 км, что свидетельствует о значительном разбросе в пробегах. Стандартное отклонение количества дверей — 1.12, что указывает на разнообразие типов кузовов. Количество владельцев имеет стандартное отклонение 1.38,

что может говорить о различной истории владения автомобилями. Стандартное отклонение цены — 2953.23 единицы, что указывает на умеренный разброс в ценах. В целом, данные показывают, что автомобили в наборе данных имеют значительную вариативность по пробегу и умеренную — по остальным параметрам.

На рисунке 4 реализуется процесс вычисления медианных значений для нескольких столбцов из DataFrame. Код вычисляет медианные значения для столбцов 'год', 'Объем_двигателя', 'Пробег', 'Двери' и 'Количество_владельцев'. Для каждого столбца используется функция `median()`, которая возвращает медиану всех значений в столбце. Медиана представляет собой значение, которое разделяет набор данных на две равные половины, что делает её полезной для анализа данных с выбросами или асимметричным распределением. Результаты вычислений сохраняются в переменные `year_median`, `hp_median`, `mileage_median`, `doors_median` и `owners_median`. В конце кода эти переменные выводятся, показывая медианные значения для каждого из столбцов.

```
[9] variance_values = numeric_data.var(ddof=1)
variance_values
```

Год	4.654264e+01
Объем_двигателя	1.268250e+00
Пробег	6.777477e+09
Двери	1.248062e+00
Количество_владельцев	1.916944e+00
Цена	8.721564e+06

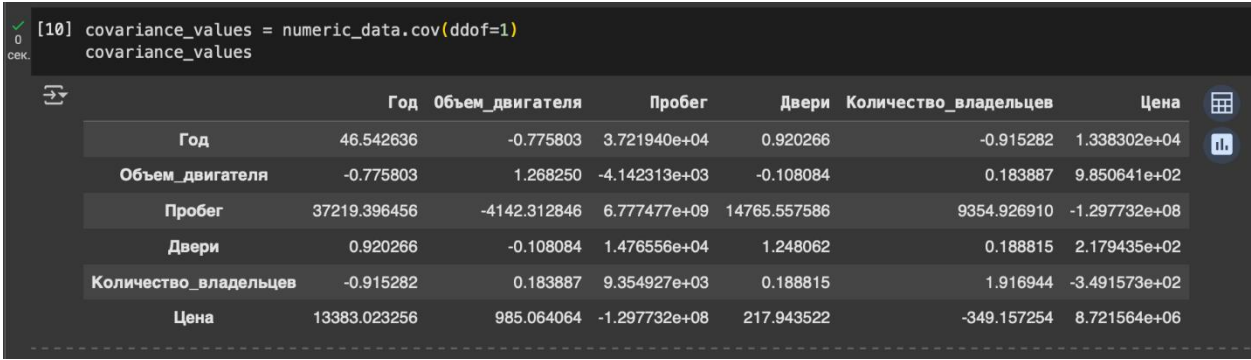
dtype: float64

Рисунок 4 – Процесс получения дисперсии значений

На основе представленных данных, где использована функция для вычисления дисперсии, можно сделать следующие выводы. Дисперсия года выпуска автомобилей составляет примерно 46.54, что указывает на умеренный разброс в возрасте автомобилей. Дисперсия объема двигателя — 1.27, что говорит о некоторой вариативности в размерах двигателей. Дисперсия пробега автомобилей составляет около 6.78 миллиардов км², что свидетельствует о значительном разбросе в пробегах. Дисперсия количества дверей — 1.25, что указывает на разнообразие типов кузовов. Дисперсия количества владельцев — 1.92,

что может говорить о различной истории владения автомобилями. Дисперсия цены составляет около 8.72 миллиона единиц, что указывает на умеренный разброс в ценах. В целом, данные показывают, что автомобили в наборе данных имеют значительную вариативность по пробегу и умеренную — по остальным параметрам.

На рисунке 5 реализуется процесс вычисления ковариации для нескольких столбцов из датафрейма. Код вычисляет моду для столбцов 'Год', 'Объем_двигателя', 'Пробег', 'Двери' и 'Количество_владельцев'. Ковариация пар Год выпуска и Пробег составляет 3.7219. Ковариация пар Пробег и Цена составляет 6.7775. Ковариация пар Количество владельцев и Цена составляет -3.49. Ковариация значений Объем двигателя и Цена составляет 9.85



```
[10] covariance_values = numeric_data.cov(ddof=1)
      covariance_values
```

	Год	Объем_двигателя	Пробег	Двери	Количество_владельцев	Цена
Год	46.542636	-0.775803	3.721940e+04	0.920266	-0.915282	1.338302e+04
Объем_двигателя	-0.775803	1.268250	-4.142313e+03	-0.108084	0.183887	9.850641e+02
Пробег	37219.396456	-4142.312846	6.777477e+09	14765.557586	9354.926910	-1.297732e+08
Двери	0.920266	-0.108084	1.476556e+04	1.248062	0.188815	2.179435e+02
Количество_владельцев	-0.915282	0.183887	9.354927e+03	0.188815	1.916944	-3.491573e+02
Цена	13383.023256	985.064064	-1.297732e+08	217.943522	-349.157254	8.721564e+06

Рисунок 5 – Процесс получения ковариации значений

На основе представленных данных, где вычислена ковариация для нескольких столбцов датафрейма, можно сделать следующие выводы. Ковариация между годом выпуска и пробегом составляет 3.72e+04, что указывает на слабую положительную связь: более новые автомобили могут иметь меньший пробег. Ковариация между пробегом и ценой равна -1.30e+08, что свидетельствует о сильной отрицательной связи: автомобили с большим пробегом, как правило, дешевле. Ковариация между количеством владельцев и ценой составляет -349.16, что также указывает на отрицательную связь: автомобили с большим количеством владельцев могут быть дешевле. Ковариация между объемом двигателя и ценой равна 985.06, что говорит о слабой положительной связи: автомобили с большим объемом двигателя могут быть дороже. В целом, данные показывают, что пробег и количество владельцев имеют значительное влияние на цену, в то время как год выпуска и объем двигателя оказывают меньшее воздействие.

На основе анализа данных, включающего вычисление средних значений, стандартных отклонений, дисперсий и ковариаций, можно сделать следующие общие выводы. Автомобили в наборе данных в среднем выпущены около 2010 года, с умеренным пробегом около 130 935 км и средней ценой примерно 9029.67 единиц. Объем двигателя в среднем составляет 3.03 литра, а количество дверей — около 3. Стандартные отклонения и

дисперсии указывают на значительный разброс в пробеге и умеренный — в остальных параметрах.

Ковариационный анализ выявил, что пробег и количество владельцев имеют сильное отрицательное влияние на цену: автомобили с большим пробегом и большим количеством владельцев, как правило, дешевле. Объем двигателя и год выпуска показывают слабую положительную связь с ценой, что может указывать на то, что более новые автомобили и автомобили с большим объемом двигателя могут быть дороже.

В целом, данные показывают, что пробег и количество владельцев являются ключевыми факторами, влияющими на цену автомобилей, в то время как год выпуска и объем двигателя оказывают меньшее, но все же заметное влияние

Вывод:

Проведенная работа была направлена на анализ и исследование данных с использованием статистических методов. В процессе работы были применены различные подходы, включая расчет средних значений, стандартных отклонений, дисперсий и ковариаций, что позволило оценить характер распределения данных, выявить особенности и тенденции. С помощью библиотек Python, таких как pandas и numpy, удалось провести детальный анализ данных, включая расчет ключевых статистических показателей. В процессе работы я научился применять различные статистические методы для анализа данных, интерпретировать результаты расчетов и делать выводы на основе полученных данных. Это позволило мне лучше понять, как различные параметры, такие как пробег, количество владельцев, год выпуска и объем двигателя, влияют на цену автомобилей. Работа также помогла развить навыки обработки и анализа данных, что является важным аспектом в области data science и аналитики.