УДК 69:657
DOI 10.33764/2687-041X-2024-2-16-19

О. И. Елфимова[1], Л. А. Максименко[1]*

Применение диаграммы BOXPLOT для анализа данных

[1]Сибирский государственный университет геосистем и технологий, г. Новосибирск, Российская Федерация
* e-mail: elfimovaoksana08@gmail.ru

Аннотация. В статье рассматривается диаграмма «ящик с усами» (BOX PLOT), которая используется для визуализации распределения данных и позволяет увидеть основные характеристики выборки, такие как медиана, квартили, выбросы. Она позволяет сравнивать распределения разных выборок и выявлять закономерности в данных. Целью работы явилось рассмотрение графического подхода к интерпретации данных с помощью диаграммы **BOX PLOT** (коробчатой диаграммы). Для решения поставленной цели были решены следующие задачи: рассмотрено понятие генеральной совокупности и выборки; меры центральной тенденции; проанализированы вариации в статистических данных; рассмотрено построение диаграммы на практических примерах.

Ключевые слова: визуализация, генеральная совокупность, выборка, меры центральной тенденции, статистические показатели

O. I. Elfimova[1], L. A. Maksimenko[1]

Using a BOXPLOT chart for data analysis

[1]Siberian State University of Geosystems and Technologies, Novosibirsk, Russian Federation
e-mail: elfimovaoksana08@gmail.ru

Abstract. The article discusses the «box with a mustache» diagram (BOX PLOT), which is used to visualize the distribution of data and allows you to see the main characteristics of the sample, such as median, quartiles, outliers. It allows you to compare the distributions of different samples and identify patterns in the data. The aim of the work was to consider a graphical approach to data interpretation using a BOXPLOT diagram (box diagram). To achieve this goal, the following tasks were solved: the concept of the general population and sampling was considered; measures of the central trend; the variations in statistical data are analyzed; the construction of the diagram on practical examples is considered.

Keywords: visualization, general population, sampling, measures of the central trend, statistical indicators

Введение

Диаграмма «ящик с усами» может быть применима при исследованиях в различных областях деятельности [1]. Например, при визуализации распределения данных, диаграмма «ящик с усами» помогает представить различные характеристики кадастровых данных, такие как площадь участков, стоимость недвижимости и другие параметры. Это позволяет исследователям быстро оценить основные статистические показатели и выявить выбросы или аномалии в данных. Диаграмма «ящик с усами» позволит сравнивать распределения кадастровых данных для разных групп или категорий. Например, можно сравнить распределение стоимости недвижимости в разных районах или типах земельных

участков, что поможет выявить различия и понять особенности каждой группы. Диаграмма «ящик с усами» позволяет идентифицировать выбросы или аномальные значения в кадастровых данных, что может быть полезно для обнаружения ошибок в данных или выявления потенциальных проблем, таких как недооценка или переоценка недвижимости. Таким образом, диаграмма «ящик с усами» является полезным инструментом для исследований кадастра, позволяющим визуализировать и анализировать различные характеристики данных.

Методы и материалы

BOX PLOT коробчатая диаграмма (или «ящик с усами») для анализа данных, на которой отражены медиана (значение серединного элемента для набора данных), нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. Мера центральной тенденции служит для описания множества значений единственным числом. Основные меры центральной тенденции: арифметическое среднее – сумма всех наблюденных значений, делённая на их количество; медиана – значение, которое делит упорядоченные по возрастанию (убыванию) наблюдения пополам рис.1; мода – наиболее часто встречающееся значение. Вариабельность значений признака показывают: дисперсия, стандартное отклонение, размах, квартильный размах.

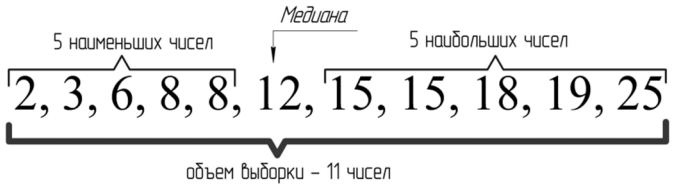

Рис.1. Пример определения медианы

Для визуального сравнения одного распределения с другим обычно строят несколько «ящиков с усами». Расстояния между различными частями ящика можно использовать для определения степени разброса (дисперсии) и асимметрии данных, а также для выявления выбросов, как показано на диаграмме, построенной по результатам сдачи экзамена по разным предметам для трех групп обучаемых.

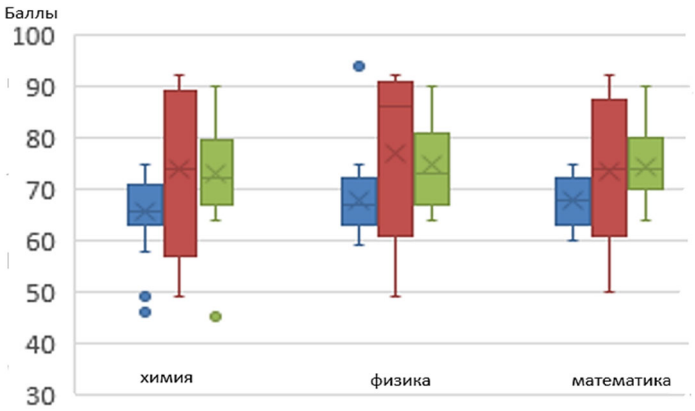

Рис. 2. Пример диаграммы BOX PLOT

Из выше представленного рисунка видны наименьшее и наибольшее значение набранных баллов, для каждой группы обучающихся. Размах в статистическом смысле предполагает разность наибольшего и наименьшего числа. Усы нам говорят о промежутке, в котором находятся все данные. Линия, пересекающая ящик является медианой. Точки на границе ящика — это медианы каждой половины. Таким образом, все данные разделены на четыре равные по численности группы, называются квартилями (1Q,2Q,3Q,4Q).

При выполнении работы были определены электронные источники для извлечения и анализа кадастровой информации:

– Агентство недвижимости ЦИАН [3];
– Real Capital Analytics [4];
– Росреестр. Статика и аналитика [5];
– Портал открытых данных РФ [6];
– Сайт федеральной службы государственной статистики [7];
– Набор открытых данных Министерства просвещения РФ [8];
– Открытые данные на сайте Министерства труда [9].

Результаты

По полученным данным агентства недвижимости ЦИАН на 2023 год были построены диаграммы BOX PLOT, что позволило провести исследования по стоимости жилой недвижимости в определенные промежутки времени в Советском районе в верхней и нижней зонах Академгородка г. Новосибирска (рис.3).

Для проведения исследования применили поисковые системы Google Chrome, Яндекс Браузер и программное обеспечение Excel, Word.

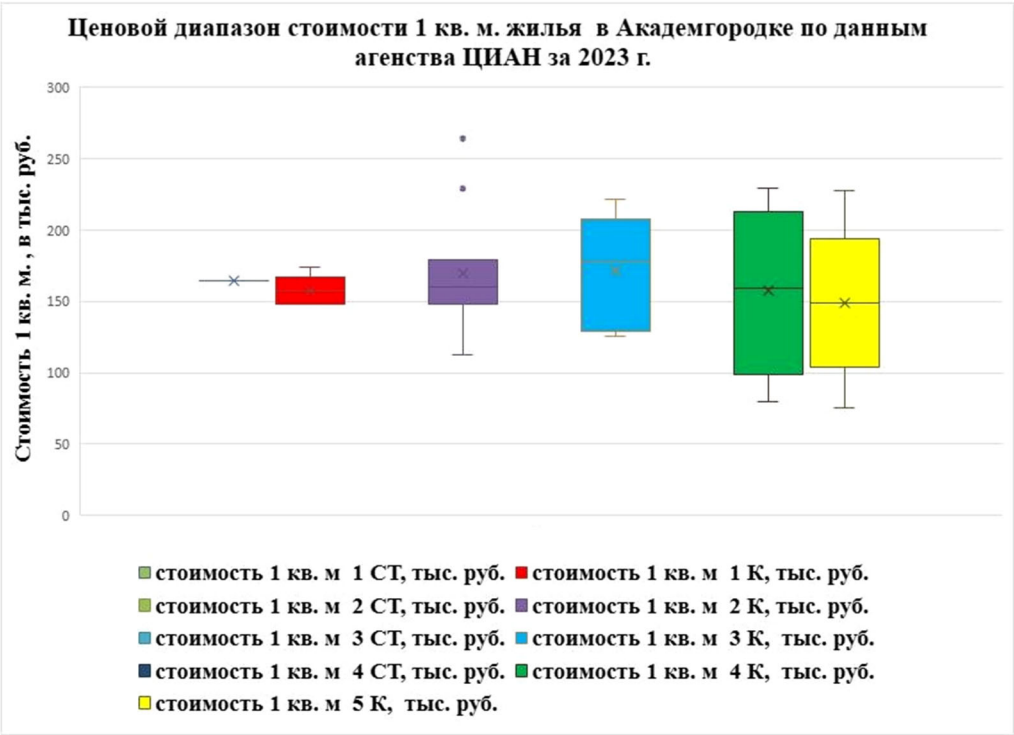

Рис. 3. Результат исследования по стоимости жилой недвижимости

Заключение

В результате проведения исследований было выявлено, что диаграмма BOX PLOT подходит для сравнения диапазона и распределения значений в группах числовых данных. Преимуществом является упорядочивание больших объемов данных и визуализация значений выбросов, представление сводных сведений о распределении данных, при этом диаграмма не подходит для тщательного анализа данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Каталог визуализации данных [Электронный ресурс]. – Режим доступа: https://datavizcatalogue.com/RU/.
2. Об информации, информационных технологиях и о защите информации [Электронный ресурс] : Федеральный закон от 27.07.2006 №149-ФЗ (ред. от 14.07.2022). – Доступ из справ.-правовой системы «Консультант Плюс».
3. Агентство недвижимости ЦИАН [Электронный ресурс]. – Режим доступа: https://www.cian.ru/analiz-rynka-nedvizhimosti-b2b/.
4. Real Capital Analytics [Электронный ресурс]. – Режим доступа: https://app.rcanalytics.com/.
5. Росреестр. Статика и аналитика [Электронный ресурс]. – Режим доступа: https://rosreestr.gov.ru/open-service/statistika-i-analitika/.
6. Портал открытых данных РФ [Электронный ресурс]. – Режим доступа: https://data.gov.ru/.
7. Сайт федеральной службы государственной статистики [Электронный ресурс]. – Режим доступа: https://rosstat.gov.ru/.
8. Набор открытых данных Министерства просвещения РФ [Электронный ресурс]. – Режим доступа: https://opendata.edu.gov.ru/opendata/.
9. Открытые данные на сайте Министерства труда [Электронный ресурс]. – Режим доступа: https://mintrud.gov.ru/opendata.
10. Максименко Л. А. Меры и типы признаков кадастровой информации = Measures and types of cadastral information features / Л. А. Максименко. - DOI 10.33764/2687-041X-2023-1-282-285. - Текст : непосредственный // Регулирование земельно-имущественных отношений в России: правовое и геопространственное обеспечение, оценка недвижимости, экология, технологические решения : сб. материалов 6 нац. науч.-практ. конф. с междунар. участием, посвящ. празднованию 90-летия НИИГАиК – СГГА – СГУГиТ, 23–25 нояб. 2022 г., Новосибирск: в 3 ч. – Новосибирск: СГУГиТ, 2023. – Ч. 1. – С. 282-285