ИСПОЛЬЗОВАНИЕ ВИЗУАЛИЗАЦИИ ДЛЯ ЭФФЕКТИВНОГО АНАЛИЗА ДАННЫХ

Марухин Е.М.[1], Марухина О.В.[2]
[1] МБОУ Лицей при ТПУ, 031, e-mail: fordgod9@gmail.com
[2] НИ ТПУ, ИШИТР, доцент ОИТ, e-mail: Marukhina@tpu.ru

Введение

Существует множество способов визуального представления данных. Однако существует всего несколько способов, которыми можно изобразить данные таким образом, чтобы можно было увидеть что-то визуально и наблюдать новые закономерности. Визуализация данных не так проста, как кажется; это искусство, требующее большой практики и опыта (точно так же, как рисовать картину – нельзя стать мастером живописи с первого дня, для этого требуется много практики).

Целью нашей работы являлось исследование важности применения методов визуализации в решении задач анализа данных.

Описание и решение задачи

Эффективная визуализация помогает анализировать и понимать данные [1]. Одним из классических примеров является «квартет Энскомба» (Фрэнсис Джон Энскомб, (13.05.1918 г. –17.10.2001 г.), английский математик-статистик), состоящая из четырех наборов данных, у которых простые статистические свойства идентичны, но их графики существенно отличаются. Каждый набор состоит из 11 пар чисел. Квартет был составлен в 1973 году для иллюстрации важности применения графиков для статистического анализа и влияния выбросов значений на свойства всего набора данных [2].

Проиллюстрируем свойства этого примера (программа реализована на языке Python с использованием библиотеки seaborn) и покажем, какую роль играет визуализация в решении задач анализа данных. Суть заключается в следующем: есть четыре набора данных (сгруппированных по столбцу набора данных) со значениями *x* и *y*. Необходимо изучить и сравнить эти группы с точки зрения связи между переменными, в том числе, с использованием методов визуализации.

```
import seaborn as sns
# Загружаем датасет для «квартета Энскомба»
anscombe = sns.load_dataset('anscombe')
# Выводим значения описательных статистик по каждой группе
anscombe.groupby('dataset').describe()
```

	x								y							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
dataset																
I	11.0	9.0	3.316625	4.0	6.5	9.0	11.5	14.0	11.0	7.500909	2.031568	4.26	6.315	7.58	8.57	10.84
II	11.0	9.0	3.316625	4.0	6.5	9.0	11.5	14.0	11.0	7.500909	2.031657	3.10	6.695	8.14	8.95	9.26
III	11.0	9.0	3.316625	4.0	6.5	9.0	11.5	14.0	11.0	7.500000	2.030424	5.39	6.250	7.11	7.98	12.74
IV	11.0	9.0	3.316625	8.0	8.0	8.0	8.0	19.0	11.0	7.500909	2.030579	5.25	6.170	7.04	8.19	12.50

Рис. 1. Описательная статистика для датасетов в задаче «Квартет Энскомба»

Результаты вычислений, представленные на рисунке 1, показывают, что такие характеристики как среднее значение (*mean*) и стандартное отклонение (*sd*) идентичны по *x* и *y* для каждого из четырех наборов данных. Тем не менее, судя по другим параметрам (минимум, максимум, значениям процентилей), можно утверждать точно, что наборы данных имеют различия. Таким образом, на основании произведенных вычислений нельзя сделать вывод о характере связи между переменных наборов данных. Для дальнейшего исследования связи между *x* и *y* найдем корреляцию между этими переменными.

```
# Вычисляем матрицы корреляций для всех датасетов
anscombe.groupby('dataset').corr()
```

		x	y
dataset			
I	x	1.000000	0.816421
	y	0.816421	1.000000
II	x	1.000000	0.816237
	y	0.816237	1.000000
III	x	1.000000	0.816287
	y	0.816287	1.000000
IV	x	1.000000	0.816521
	y	0.816521	1.000000

Рис. 2. Корреляционные матрицы по каждому из четырех наборов данных

Полученные значения коэффициентов корреляции свидетельствуют о прямой положительной зависимости между переменными. Но, к сожалению, больше никаких выводов на данном этапе сделать нельзя. Визуализируем данные с помощью диаграммы рассеяния.

```
# Визуализация данных – строим диаграммы рассеяния для всех датасетов
sns.scatterplot(data=anscombe, x='x', y='y', hue='dataset')
```

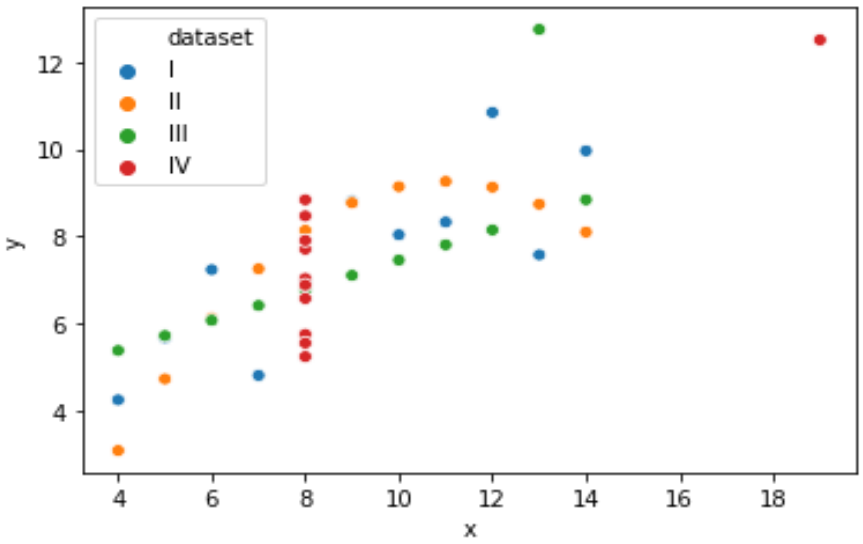

Рис. 3. Диаграммы рассеяния для исследуемых данных

Несмотря на «статистическую идентичность», видим, что это совсем разные наборы с точки зрения выбора модели, описывающей данные. Первый набор – линейная модель с шумом, на второй видна квадратичная зависимость, третий – линейная с выбросом, последний – константа с выбросом.

Наборы данных, которые идентичны по ряду статистических свойств, но создают разные графики, часто используются для иллюстрации важности графического представления при изучении данных [3].

Заключение

Хорошая визуализация помогает пользователям исследовать и осмысливать данные, обеспечивая ценность и глубокое понимание. Значимость визуализации была проиллюстрирована на примере, известном как квартет Энскомба. Это четыре набора данных, которые почти идентичны по описательным характеристикам, но имеют разное распределение и при графическом представлении дают совершенно разную картину.

Качественная визуализация данных имеет критическое значение для анализа данных и принятия решений на их основе. Визуализация позволяет быстро и легко замечать и интерпретировать связи и взаимоотношения, а также выявлять развивающиеся тенденции, которые не привлекли бы внимания в виде необработанных данных.

Список использованных источников

1. Желязны Д. Говори на языке диаграмм: Пособие по визуальным коммуникациям для руководителей / Пер. с англ. – М.: Институт комплексных стратегических исследований, 2004. – 220 с.
2. Anscombe’s quartet (Квартет Энскомба). [Электронный ресурс]. – URL: https://seaborn.pydata.org/ examples/ anscombes_quartet.html (дата обращения 23.01.2023).
3. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing (Идентичная статистика, разные графики: генерирование наборов данных с разнообразным внешним видом и идентичной статистикой). [Электронный ресурс]. – URL: https://www.autodesk.com/research/publications/same-stats-different-graphs (дата обращения 23.01.2023).