

НАУЧНАЯ СТАТЬЯ
УДК 004.6

АНАЛИЗ ДАННЫХ С ПОМОЩЬЮ WEB SCRAPING PYTHON

Анастасия Александровна ВОРОБЬЕВА

ФГБОУ ВО «Тамбовский государственный университет им. Г.Р. Державина»,
г. Тамбов, Российская Федерация

Аннотация. Описаны принципы работы с Web Scraping Python. Рассмотрены алгоритмы анализа данных с помощью программных средств, описаны основные инструменты для извлечения данных и работы с ними. Также разобран пример анализа рынка недвижимости с помощью данной технологии. Продемонстрированы различные графики зависимости между параметрами и на основе этого сделаны выводы.

Ключевые слова: анализ больших данных, Web Scraping Python, рынок недвижимости

Для цитирования: Воробьева А.А. Анализ данных с помощью Web Scraping Python // Державинский форум. 2022. Т. 6, № 3. С. 533-544

ORIGINAL ARTICLE

DATA ANALYSIS USING WEB SCRAPING PYTHON

Anastasiya A. VOROBIEVA

Derzhavin Tambov State University, Tambov, Russian Federation

Abstract. The principles of working with Web Scraping Python are described. Algorithms of data analysis using software tools are considered, the main tools for data extraction and work with them are described. An example of real estate market analysis using this technology is also analyzed. Various graphs of the dependence between the parameters are demonstrated and conclusions are drawn based on this.

Keywords: big data analysis, Web Scraping Python, real estate market

For citation: Vorobyeva A.A. Analiz dannykh s pomoshch'yu Web Scraping Python [Data Analysis using Web Scraping Python]. *Derzhavinskiy forum – Derzhavin Forum*, 2022, vol. 6, no. 3, pp. 533-544. (In Russian, Abstr. in Engl.)

ВВЕДЕНИЕ

Web Scraping – это процесс извлечения необходимых данных с разных веб-страниц для обработки, сравнения и различного анализа. С появлением новых технологий Web Scraping проводить стало более быстро и удобно. Нет необходимости собирать данные вручную и использовать дополнительные средства для дальнейшей обработки. Такой инструмент, как Web Scraping Python, поможет решить такую задачу намного эффективнее. Можно собирать различные типы данных, такие как текст, изображения, URL-адреса и др.

Web Scraping имеет большой спектр применений. Программисты, специалисты по маркетингу, Data Scientist регулярно используют Web Scraping для сбора общедоступных данных. Если данные, которые вы собираете, не требуют учетной записи для доступа и не блокируются сервером, а также имеют статус «общедоступны», то такие данные можно собирать для дальнейшей работы с ними. Полученные данные можно использовать для различного анализа [1].

1. ИНСТРУМЕНТЫ ДЛЯ АНАЛИЗА ДАННЫХ

Requests – это HTTP-библиотека, написанная на Python (под лицензией Apache2). С помощью модуля Requests можно управлять HTTP-запросами при помощи языка Python. Библиотека имеет широкий инструментарий библиотеки и рассчитана на все случаи взаимодействия с веб-приложениями. Что же такое запросы HTTP? Запрос – это сообщение, которое посылает клиент серверу. Всякий раз, когда пользователь открывает веб-страницу, браузер направляет множество запросов на сервер нужной веб-страницы. Сервер отвечает на них, он пересылает все необходимые данные для вывода страницы, после чего браузер ее отображает для пользователя.

Клиентом может выступать не только браузер, но и скрипт Python, использующий библиотеку Requests. В целом механизм работы выглядит так: клиент отправляет данные на URL, а сервер с этим URL считывает данные и отправляет клиенту ответ¹.

В составе запроса клиент отправляет данные по методу запроса. Наиболее используемыми методами запроса являются GET, POST и PUT. Запросы GET предназначены только для чтения данных без их изменения, а запросы POST и PUT обычно предназначаются для измене-

¹ Работа с PyCharm. URL: <https://www.jetbrains.com/ru-ru/pycharm/> (дата обращения: 10.06.2022).

ния данных на сервере. Так как мы не изменяем данные на странице, то метод запроса будет GET. Функция `requests.get()` позволит получить ответ от сервера.

`BeautifulSoup4(bs4)` предоставляет собой несколько простых функций в стиле Python для обработки навигации, поиска, изменения дерева анализа и других связанных функций. Такой набор инструментов предоставляет пользователям данные, которые следует захватить путем анализа информации. Благодаря своей простоте и легкости в использовании, полное приложение может быть написано без большого количества кода².

`BeautifulSoup4` представляет один интерфейс для разных парсеров. Разные парсеры, анализируя один и тот же документ, создадут различные деревья HTML. Стоит заметить, что `BeautifulSoup` автоматически преобразует входные документы в кодировку Unicode, а выходные документы – в кодировку utf-8. То есть не нужно учитывать метод кодирования, если в документе это не указано, в настоящее время `BeautifulSoup` не может автоматически распознавать метод кодирования.

Наиболее популярные – это HTML и XML парсеры. Чтобы разобрать HTML-документ, необходимо передать его в конструктор класса `BeautifulSoup()`. Можно передать строку или открытый дескриптор файла.

Для глубоко вложенных HTML-документов навигация может быстро стать утомительной. К счастью, `BeautifulSoup` поставляется с функцией поиска, поэтому нам не нужно перемещаться, чтобы получить HTML-элементы. Функции `find()` и `find_all()` помогут по тегу отыскать нужный фрагмент среди всего полученного кода. Можем искать теги определенного класса, предоставляя аргумент `class_`. `BeautifulSoup` использует `class_`, потому что `class` является зарезервированным ключевым словом в Python [2].

2. ПОСТАНОВКА ЗАДАЧИ

Будем рассматривать сайт недвижимости «Этажи», предварительно выбрав город Тамбов. Сайт располагается по URL-адресу <https://tambov.etagi.com/realty/>, он содержит 20 страниц с описанием квартир, на каждой отдельной вкладке находится 30 объектов. Каждая из квартир содержит определенные параметры (рис. 1).

² Работа с библиотекой `Requests`. URL: <https://pythonist.ru/ispolzovanie-biblioteki-requests-v-python/> (дата обращения: 17.06.2022).

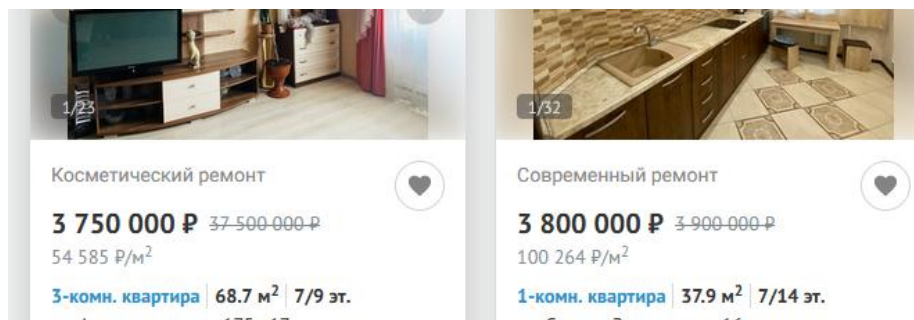


Рис. 1. Информация о квартирах
Fig. 1. Information about apartments

После того, как URL-адрес определен, необходимо выяснить, под какими HTML-тегами или атрибутами будут расположены нужные данные. Это можно сделать с помощью панели инструментов (рис. 2).

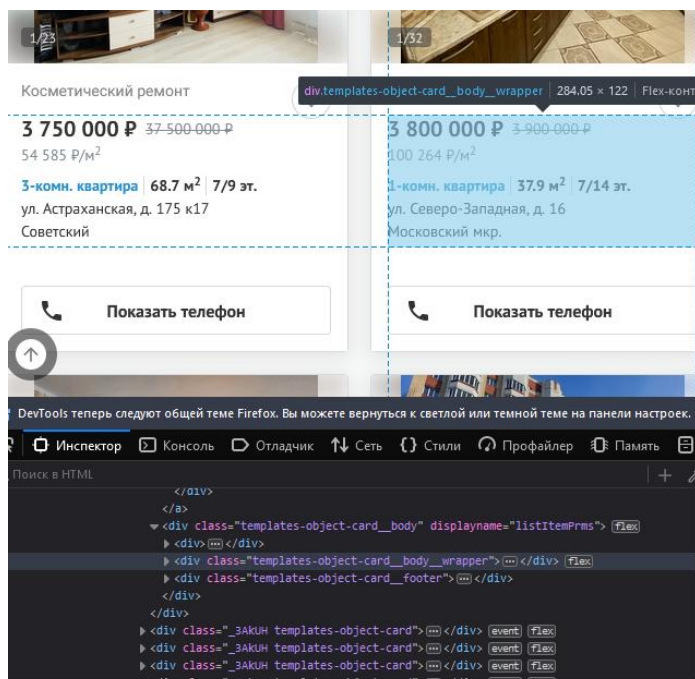


Рис. 2. Работа с тегами
Fig. 2. Working with tags

С помощью средств **Web Scraping Python** необходимо будет извлечь такие данные, как: общее число квартир, количество комнат, стоимость квартиры, общая площадь. Полученные данные нужно будет разбить на кластеры по количеству комнат, по этажам. Затем определить минимальную и максимальную цену, аналогично с площадью. После отобразить в виде гистограмм и графиков. Выяснить зависимость между разными параметрами.

3. ОПИСАНИЕ АЛГОРИТМА

На рис. 3 представлена блок-схема, которая описывает действия алгоритма.

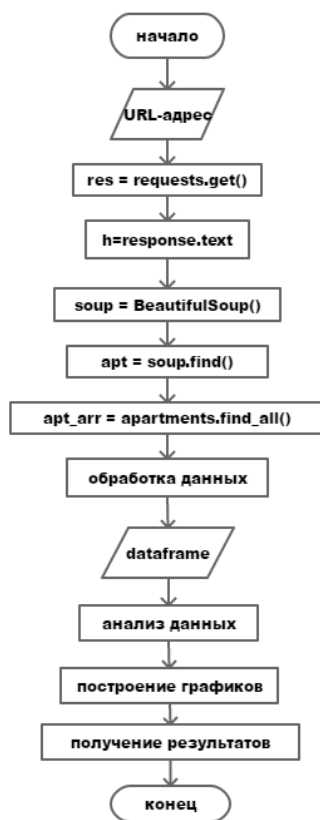
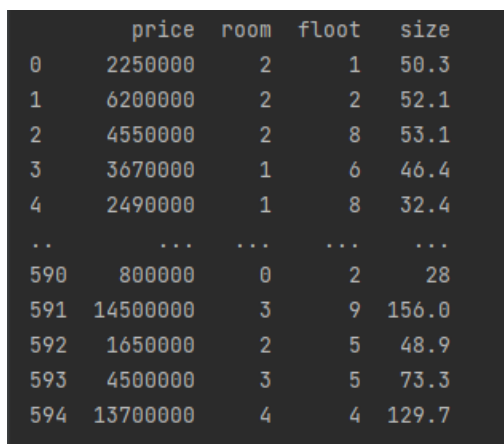


Рис. 3. Блок-схема
Fig. 3. Block diagram

На входе программы поступает URL-адрес сайта недвижимости. С помощью `requests.get()` получаем ответ от сервера. В случае успешного соединения программа продолжит работу. Метод `text` позволяет получить код `html`, после чего используем функцию `BeautifulSoup()`, которая обработает исходный код, все данные будут в переменной `soup`. С помощью функции `soup.find()` находим необходимую информацию, передав в функцию название нужного тега. Так как на сайте может находиться больше одного объекта, необходимо найти все такие объекты с помощью функции `find.all()`. После того как мы извлекли данные, необходимо выбрать нужные, а именно, цену, количество комнат, этаж и площадь квартиры. С готовым датафреймом и будут производиться дальнейшие действия.

4. АНАЛИЗ ДАННЫХ РЫНКА НЕДВИЖИМОСТИ

В самом начале мы работаем с датафреймом данных, который состоит из нескольких столбцов. Скриншот с данными представлен на рис. 4.



	price	room	floot	size
0	2250000	2	1	50.3
1	6200000	2	2	52.1
2	4550000	2	8	53.1
3	3670000	1	6	46.4
4	2490000	1	8	32.4
..	...	...	...	...
590	800000	0	2	28
591	14500000	3	9	156.0
592	1650000	2	5	48.9
593	4500000	3	5	73.3
594	13700000	4	4	129.7

Рис. 4. Структура датафрейма
Fig. 4. Dataframe Structure

4.1. АНАЛИЗ ЦЕН НА КВАРТИРЫ

Построение `Boxplots` для визуализации данных о стоимости квартир. Диаграммы размаха («ящик с усами») (`Box and Whisker Plot`, или `Box`

Plot) – это удобный способ представления различных групп числовых данных через квартили. Нахождение распространения данных – это неотъемлемая часть анализа. Программная реализация помогает построить Boxplots для стоимости всех 595 квартир [3]. Диаграмма представлена на рис. 5.

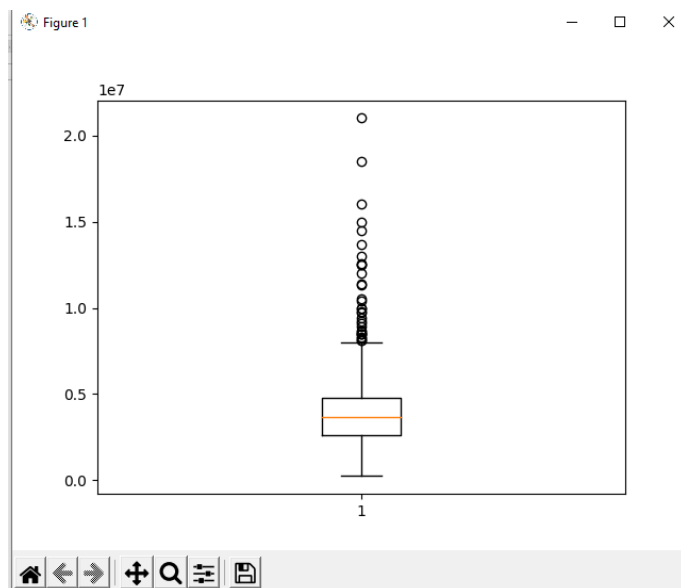


Рис. 5. Boxplots стоимости квартир
Fig. 5. Boxplots of the cost of apartments

Определим медиану с помощью функции `median()`. Значение, расположенное в середине, составляет 3683000 рублей. Первый квартиль, или 25-й процентиль – это 2650000 рублей. Третий квартиль, или 75-й процентиль – это 4800000 рублей. Их можно определить с помощью функции `quantile()`. Зная значения двух процентилей, найдем межквартильный диапазон. Он равен 2150000 рублей. Также не забудем про нижний диапазон – 575000 и верхний диапазон 8025000 рублей. Рассмотрев график более подробно, можно обнаружить, что есть выбросы, выбросы превосходят значение верхней границы, их порядка 30. Это значит, что среди рассмотренных квартир есть такие, которые превышают в цене остальные, что является некой аномалией.

4.2. ГИСТОГРАММА И КОЛИЧЕСТВО КОМНАТ

Гистограммы хороши для представления категориальных переменных и подходят для подсчета. Гистограммы отображают и сравнивают частоту, число или другие показатели. В данном исследовании гистограмма используется для наглядного представления количества комнат в квартирах. График представлен на рис. 6.

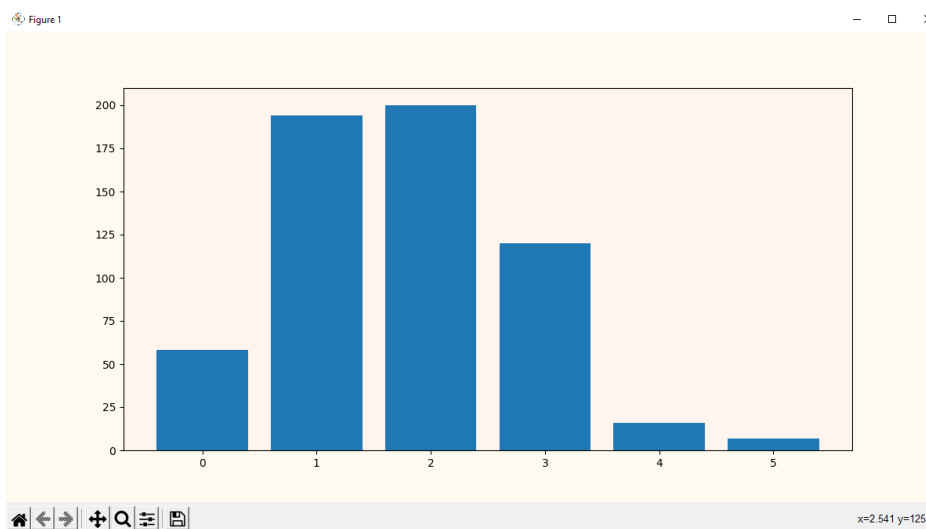


Рис. 6. Гистограмма по количеству комнат. По горизонтали – количество комнат, по вертикали – общее количество таких квартир [3]

Fig. 6. Histogram by number of rooms. Horizontally – the number of rooms, vertically – the total number of such apartments

4.3. ДИАГРАММА РАССЕЙВАНИЯ И ПЛОЩАДЬ КВАРТИР

Диаграмма рассеяния (точечная диаграмма, англ. scatterplot) – математическая диаграмма, изображающая значения двух переменных в виде точек на декартовой плоскости. На диаграмме рассеяния каждому наблюдению (или элементарной единице набора данных) соответствует точка, координаты которой равны значениям двух каких-то параметров этого наблюдения. Если предполагается, что один из параметров зависит от другого, то обычно значения независимого параметра откладываются

по горизонтальной оси, а значения зависимого – по вертикальной. Диаграммы рассеяния используются для демонстрации наличия или отсутствия корреляции между двумя переменными.

Для того чтобы отобразить, как изменяются площади относительно количества комнат, выбрана точечная диаграмма [3]. Она изображена на рис. 7. Здесь можно проследить верхний диапазон для каждого вида квартир.

Figure 1

— □ ×

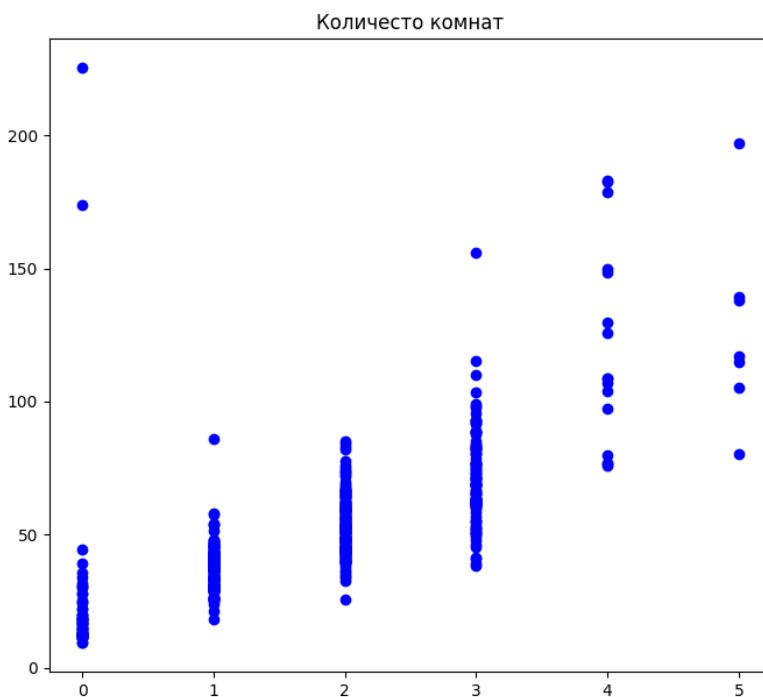


Рис. 7. Диаграмма распределения. По горизонтали – количество комнат, по вертикали – площадь помещения

Fig. 7. Distribution diagram. Horizontally – the number of rooms, vertically – the area of the room

4.4. ГРУППИРОВКА ЗНАЧЕНИЙ

Группировка – это метод, при котором вся исследуемая совокупность разделяется на группы по какому-то существенному признаку. Группировка представляет собой способ подразделения рассматриваемой совокупности данных на однородные по изучаемым признакам группы. Это делается с целью изучения структуры этой совокупности либо взаимосвязей между отдельными элементами этой совокупности. После этого можно проводить статические вычисления, например, находить стандартное отклонение, среднее, максимальное и минимальное значение, а также многое другое [3].

Здесь мы проведем группировку по количеству комнат и найдем вышеуказанные параметру, сделаем мы это с помощью функции `groupby()`. Результаты работы программы представлены на рис. 8.

	room	price			...	float			
		mean	min	max		mean	min	max	std
0	0	1.056466e+06	250000	6600000	...	3.258621	1	18	2.488996
1	1	3.399361e+06	649000	9400000	...	5.654639	1	19	3.788352
2	2	4.093600e+06	550000	13000000	...	4.235000	1	17	3.225180
3	3	5.225508e+06	650000	16000000	...	4.850000	1	16	3.443786
4	4	9.271000e+06	2999000	21000000	...	5.125000	1	14	3.500000
5	5	8.192857e+06	5130000	15000000	...	4.714286	1	10	3.592320

Рис. 8. Результаты группировки

Fig. 8. Grouping results

Группировка проводилась по всем столбцам, таким как: цена, количество комнат и размер. Сразу видно, что чем больше количество комнат, тем больше стоимость этой квартиры. В связи с этим растет минимальная и максимальная цена на квартиры разной площади.

РЕЗУЛЬТАТЫ

Недвижимость в Тамбове имеет большое разнообразие, есть различные вариации комнат, размеры квартир, а также варьируется и стоимость. Минимальную цену из всех представленных квартир определили с помощью функции `min()`, она равна 250000 рублей. Максимальную це-

ну квартир определили с помощью функции `max()`, она равна 21000000 рублей. Стоит отметить, что самая дорогая квартира стоит дороже в 84 раза по сравнению с самой дешевой. Среднее же значение цены на все квартиры составляет 3986921 рублей.

С помощью гистограммы можно сделать несколько наблюдений: так, диапазон комнат в квартирах варьируется от 0 до 5, где 0 – это студии, малосемейки, комнаты в общежитии. Также среди лидеров объявлений выступают 1- и 2-комнатные квартиры, их преобладающее большинство, около 200 каждого вида. Квартир 4- и 5-комнатных сравнительно меньше, следовательно, продаются они гораздо реже по сравнению с остальными. Этот факт можно объяснить тем, что таких квартир в принципе меньше, так как такие квартиры практически не строят.

С помощью точечной диаграммы рассмотрели, как изменяются площади относительно количества комнат. Большинство однокомнатных квартир имеет площадь в диапазоне от 20 до 50 квадратных метров. Прослеживается тенденция: чем больше комнат в квартире, тем больше ее площадь.

Еще одно интересное наблюдение: в городе Тамбов не продаются квартиры выше 20-го этажа. А минимальный этаж, который возможен, – это первый. По данным выборки средний этаж для однокомнатных квартир – шестой. Это говорит о том, что застройка в городе в основном – это здания многоэтажные и повышенной этажности.

ВЫВОДЫ

Web Scraping – это технология получения веб-данных путем извлечения их со страниц веб-ресурсов. В данном исследовании данные были взяты с сайта недвижимости. Вся реализация была сделана на Python. В примере было найдено несколько параметров, а именно, средняя цена, изменение цены в зависимости от количества комнат, представлены графики, отображающие эти закономерности.

Сегодня анализ данных занимает ключевую роль на многих направлениях бизнеса. Это требуется, прежде всего, для того чтобы принимать решения, опираясь на объективные данные, а не на интуицию и личный опыт. Многие крупные компании используют data-driven подход. Он позволяет составлять стратегии бизнеса и рекламных кампаний, выстраивать современную систему отчетности и не только.

Список источников

1. *Mitchell R. Web Scraping with Python*. 2nd edition. Sebastopol: O'Reilly Media, Inc., 2018. 280 p.
2. *Федоров Д.Ю.* Программирование на языке высокого уровня Python. М.: Юрайт, 2019. 161 с.
3. *Абдрахманов М.И.* Python. Визуализация данных: Matplotlib. Seaborn. Mayavi. Devpractice Team, 2020. 412 с. URL: <https://devpractice.ru/book-python-data-viz/> (дата обращения: 17.06.2022).

References

1. Mitchell R. *Web Scraping with Python*. Sebastopol, O'Reilly Media, Inc. Publ., 2018, 280 p.
2. Fedorov D.Y. *Programmirovaniye na yazyke vysokogo urovnya Python* [Programming in the High-Level Python Language]. Moscow, Urait Publ., 2019, 161 p. (In Russian).
3. Abdrakhmanov M.I. *Python. Vizualizatsiya dannykh. Matplotlib. Seaborn. Mayavi* [Python. Data Visualization. Matplotlib. Seaborn. Mayavi]. Devpractice Team Publ., 2020, 412 p. Available at: <https://devpractice.ru/book-python-data-viz/> (accessed 17.06.2022).

Информация об авторе

Воробьева Анастасия Александровна, магистрант по направлению подготовки «Математика (профиль «Обработка больших данных и интеллектуальные системы поддержки принятия решений»», Тамбовский государственный университет им. Г.Р. Державина, Российская Федерация, 392000 г. Тамбов, ул. Интернациональная, 33, Tmiooy@gmail.com

Information about the author

Anastasiya A. Vorobyeva, Master's Degree Student in "Mathematics (Profile Big Data Processing and Intelligent Decision Support Systems)" Programme, Derzhavin Tambov State University, Internatsionalnaya St., 33, Tambov 392000, Russian Federation, Tmiooy@gmail.com

Статья поступила в редакцию/The article was submitted 19.07.2022
Одобрена после рецензирования/Approved after reviewing 13.09.2022
Принята к публикации/Accepted for publication 27.09.2022