УДК 004.67

ОБРАБОТКА И АНАЛИЗ СТАТИСТИЧЕСКИХ ДАННЫХ С ПОМОЩЬЮ ЯЗЫКОВ ПРОГРАММИРОВАНИЯ PYTHON И R

Д. Р. Аширова

студентка, Белорусский государственный университет, г. Минск, Беларусь, diashr3@gmail.com

Научный руководитель: **Е. В. Сошникова**

старший преподаватель, Белорусский государственный университет, г. Минск, Беларусь, soshnikova@bsu.by

В статье рассмотрен функционал языков программирования R и Python в целях обработки и анализа данных, а также определены их основные сравнительные преимущества и недостатки. В рамках работы перечислены основные библиотеки, которые используются для работы со статистическими данными.

Ключевые слова: язык программирования; анализ данных; библиотеки; эконометрическое моделирование.

PROCESSING AND ANALYSIS OF STATISTICAL DATA USING THE PYTHON AND R PROGRAMMING LANGUAGES

D. R. Ashirova

student, Belarusian State University, Minsk, Belarus, diashr3@gmail.com

Supervisor: **E. V. Soshnikova**

senior lecturer, Belarusian State University, Minsk, Belarus, soshnikova@bsu.by

The article examines the functionality of the R and Python programming languages for data processing and analysis, as well as identifies their main comparative advantages and disadvantages. The work lists the main libraries that are used to work with statistical data.

Keywords: programming language; data analysis; libraries; econometric modeling.

В современном мире анализ статистических данных требует использование гибких и трудоемких подходов, позволяющих работать с большими массивами информации и обеспечивать своевременную их актуализацию и обработку. Выбор того или иного инструмента для анализа данных зависит

от целей, которые преследуются в рамках работы, объема используемой информации, имеющейся квалификации и навыков исследователя.

В области Data Science широко распространено использование таких языков программирования, как Python и R. Язык Python характеризуется понятным синтаксисом и большим количеством библиотек. Язык R разрабатывался целенаправленно для работы со статистическими данными, и в нем представлен широкий инструментарий для визуализации данных [4].

Как правило, в научной и исследовательской практике принято, что использование предметно-ориентированных языков, таких как MATLAB или R, подходит для проверки и генерации новых идей, выполнения локальных проектов. Если речь идет о больших производственных системах, то они в основном реализуются на высокоуровневых языках (C#, Java, C++). Однако наблюдаются переходы в сторону реализации масштабных проектов в рамках одного языка программирования для того, чтобы избежать проблемы «двух языков» [2].

Язык программирования R реализуется на принципах открытого кода, который позволяет создавать свои библиотеки и редактировать уже существующие, полностью интерпретируя их под цели и задачи исследователя.

Одним из важных этапов при работе с данными является их качественный и агрегированный сбор. В настоящее время многие статистические сайты и онлайн-сервисы предоставляют API-ключи, которые позволяют автоматизировать интеграцию данных в Python и R. Помимо этого, такой способ сбора данных позволяет своевременно их актуализировать без мануального вмешательства [3].

Для эконометрической обработки, визуализации и иных этапов работы со статистической информацией в рамках языков программирования создано множество решений и инструментов – библиотек. Для Python широко используются следующие библиотеки: NumPy – оптимизирует вычислительные алгоритмы с многомерными массивами; pandas (работает с NumPy) – позволяет анализировать табличные данные (группировка, среднее, создание сводных таблиц); Matplotlib – для визуализации; statsmodels – библиотека с алгоритмами статистики и эконометрики (регрессионные модели, дисперсионный анализ, авторегрессионные модели и др.). Для языка программирования R также представлено большое количество библиотек: dplyr – аналог pandas; ggplot2 – библиотека для визуализации; shiny – позволяет создавать интерактивную визуализацию (дашборды); caret – для машинного обучения.

Использование этих инструментов позволяет заключить весь процесс работы с данными в одну замкнутую систему, которая начинается с самого первого этапа – загрузки данных, а заканчивается формированием

отчетов и дашбордов. R имеет более обширную базу библиотек, которые связаны с визуализацией и интерактивной графикой в целом, однако, в среде Python они также имеют место (например, библиотека seaborn). При изначальном выборе языка программирования необходимо определить конкретные задачи, которые требуется рассмотреть в ходе исследования, чтобы корректно выбрать эффективные технологии и инструменты.

Так, Python подходит для специалистов, которые впервые работают с языками программирования из-за своей простоты синтаксиса и своего широкого распространения [1]. Как правило, этот язык применяется при реализации больших коммерческих проектов, которые будут поддерживаться долгое время, а их структура будет дополняться и расти. В среднем, порог вхождения в Python значительно ниже, однако, если необходимо часто и узконаправленно работать с обработкой статистической информации с последующей визуализацией, то R удовлетворяет этим запросам полностью.

Производительность двух языков программирования зависит от набора используемых библиотек и типа реализуемых задач. По мере увеличения объема данных, R начинает уступать Python в быстродействии в связи с особенностями работы с памятью. Помимо этого, R подходит только для анализа данных и статистических исследований, в то время как возможности Python намного шире. Эти факторы могут быть решающими при выборе методов реализации целей исследователем.

Таким образом, языки программирования Python и R позволяют работать со статистическими данными в самых разных формах: сбор данных, их очистка, эконометрическое моделирование и даже их визуализация. Открытый исходный код и обилие библиотек позволяют практически персонализировать работу в рамках исследований и ведения проектов, что несомненно является одним из главных преимуществ перед программным обеспечением, которое работает с закрытым кодом.

Библиографические ссылки

1. *Антипко А. В.* Инструменты для анализа данных: сравнение Python, R и других популярных платформ [Электронный ресурс] // Молодой учёный. 2023. № 33(480). С. 14. URL: https://moluch.ru/archive/480 (дата обращения: 23.09.2024).

2. *Маккинни У.* Python и анализ данных: Первичная обработка данных с применением pandas, NumPy и Jupiter / пер. с англ. А. А. Слинкина. 3-е изд. М. : МК Пресс, 2023. 536 с.

3. 8 самых популярных языков программирования для работы с Big Data [Электронный ресурс]. URL: https://practicum.yandex.ru/blog/yazyki-programmirovaniya-dlya-big-data (дата обращения: 23.09.2024).

4. Python vs. R: что выбрать для Data Science начинающему специалисту? [Электронный ресурс]. URL: https://tproger.ru/articles/python-vs-r-for-data-science (дата обращения: 23.09.2024).