ПРОГРАММНЫЕ СРЕДСТВА СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ

Гаращенко Александр Алексеевич
Магистрант,
ФГБОУ ВО Иркутский национальный исследовательский технический университет

Развитие сферы анализа данных, именуемой в зарубежных источниках как data science, оказывает большое влияние на развитие современного общества. В частности, благодаря data science, стало возможным существенно развить технологии искусственного интеллекта и внедрить их во многие сферы повседневной жизни. Методы data science помогают в принятии оптимальных стратегических решений на коммерческих предприятиях, изучении сложных биологических структур, получении новых знаний, и это лишь некоторые из областей применения.

Для успешного осуществления анализа данных без больших временных затрат специалисту требуется мощное и удобное в использований программное средство. В данной статье будут рассмотрены наиболее часто используемые средства статистического анализа и визуализации данных, применяемых в data science. Речь пойдет о языках программирвоания Python и R.

R. Данный язык программирования появился в 1993 году. Он изначально разрабатывался для осуществления статистической обработки данных. На настоящий момент у данного проекта сформировалось крупное сообщество по всему миру, регулярно выходят новые версии языка.

Данный язык выглядит достаточно сложным для освоения с основ, во многом это связано с необходимостью разбираться в большом количестве методов и функций статистического анализа, данный язык для иных целей практически используется. При освоении R можно столкнуться с тем, что учебные материалы на русском языке достаточно немногочислены, основная часть пособий составляется на английском языке. Однако, после приобретения практических навыков по использованию данного языка анализ данных и их визуализация становятся существенно проще.

В базовом наборе R содержит основные функции по нахождению коэффициентов корреляции, дисперсии и другие. При необходимости исследователь может расширить библиотеку за счет подгружаемых пакетов. С помощью различных пакетов данную систему можно успешно применять в областях экономики, биологии, медицины и многих других. В ней доступно большое количество методов визуализации данных с получением качественных и подробных графиков.

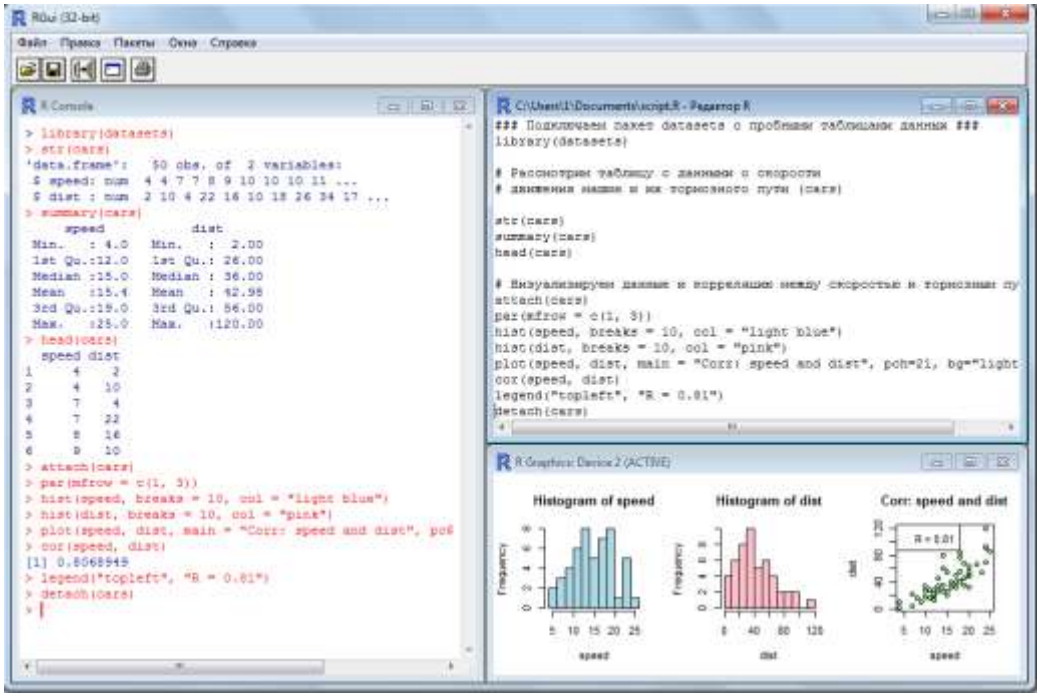

Рисунок 1. Окно среды программирования R

За счет своей узкой направленности R достаточно трудно интегрировать с другими языками и системами, тем самым он становится специфическим инструментом, который прекрасно справляется со своими задачами.

Python. Данный язык программирования увидел свет в 1991 году. В последнее время популярность его растет, во многом это обуславливается его простотой освоения и универсальностью. В частности, он может использоваться и для статистического анализа данных.

Количество доступных пакетов статистического анализа существенно меньше, чем у R. Однако существующие на данный момент пакеты включают в себя достаточный набор необходимых функций для вычислений, манипуляций с данными, машинного обучения и визуализации. Несмотря на это, многие функции, аналогичные в R, остаются не реализованными. Количество переведенной на русский язык учебной и справочной литературы существенно больше, чем для R. Универсальность использования Python позволяет легко включить статистическую обработку в любой проект.

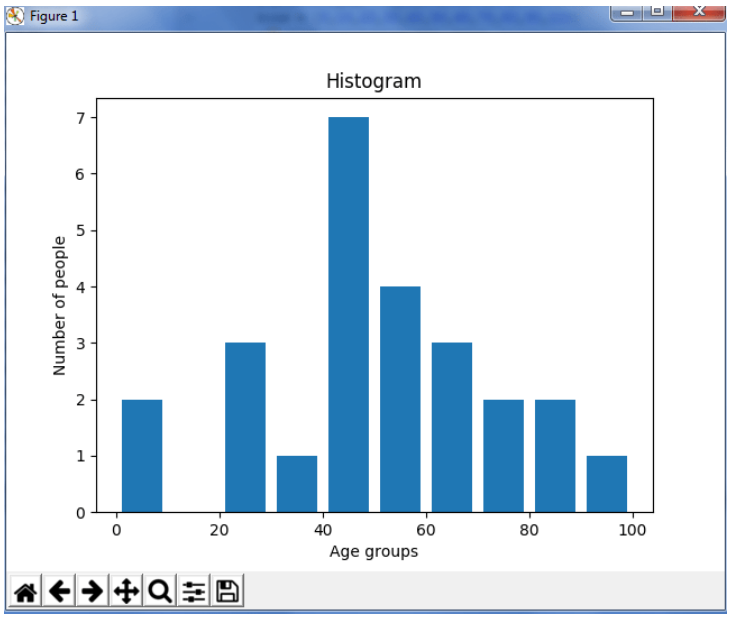

Рисунок 2. Пример использования столбчатой диаграммы в matplotlib (Python)

Резюмируя, можно сказать, что оба инструмента могут с равной успешностью использоваться в data science в зависимости от поставленных целей. По своим достоинствам и недостаткам они хорошо уравновешивают друг друга.

Список литературы

1. Ihaka R. R: Past and future history //Computing Science and Statistics. – 1998. – Т. 392396.

2. Tiobe: популярность Python в 2018 году значительно выросла [Электронный ресурс]. – Режим доступа: https://www.osp.ru/news/2019/0121/13036457 (дата обращения: 02.06.19).

3. Van Rossum G. Python 0. – 1991.