7 языков программирования, которые следует использовать в науке о данных

22.02.2024

В условиях постоянного развития науки о данных вам необходимо владеть самыми современными технологиями в этой области. В этой статье мы рассмотрим лучшие языки программирования, используемые в науке о данных. За последнее десятилетие данные приобрели огромную ценность. У каждой крупной компании есть ценные данные, которые с помощью хорошего специалиста по исследованию данных могут принести пользу ее бизнесу. В других случаях — выявить стратегии, которые, возможно, работают не так хорошо. Отрасль развивается, и спрос на специалистов по исследованию данных растет. Если вы хотите стать специалистом по изучению данных, вам следует начать с изучения лучших языков программирования в этой области. Давайте рассмотрим самые распространенные языки в Data Science и почему их следует использовать.

Python

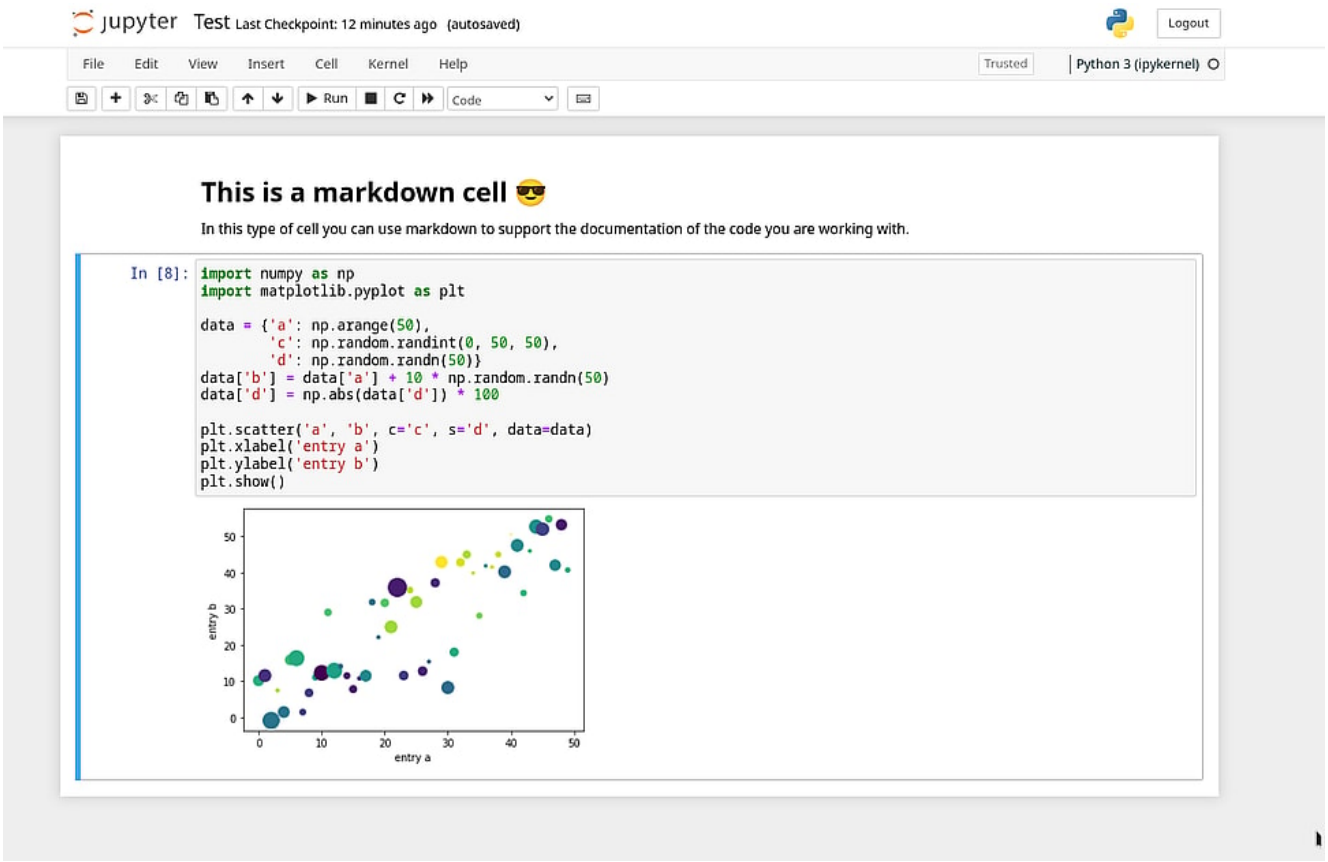
В настоящее время Python является самым используемым языком программирования. Это подтверждают несколько индексов языков программирования, таких как PYPL и TIOBE.

Worldwide, Mar 2022 compared to a year ago:

Rank	Change	Language	Share	Trend
1		Python	28.27 %	-2.0 %
2		Java	18.03 %	+0.8 %
3		JavaScript	8.86 %	+0.4 %
4		C#	7.51 %	+0.6 %
5		C/C++	7.32 %	+0.6 %
6		PHP	5.71 %	-0.4 %

Таблица наиболее используемых языков программирования PYPL.

Python — один из самых мощных и гибких языков, который также широко используется в науке о данных. Основная причина — простой и элегантный синтаксис, а также большая коллекция библиотек сторонних разработчиков. Инструмент, который можно встретить повсюду в области науки о данных, — это Jupyter. Блокноты Jupyter позволяют быстро увидеть результаты работы с кодом, построить график данных и создать документацию по коду с помощью блоков разметки. Это инструмент не только для Python, но наиболее распространенная комбинация — Python и Jupyter.

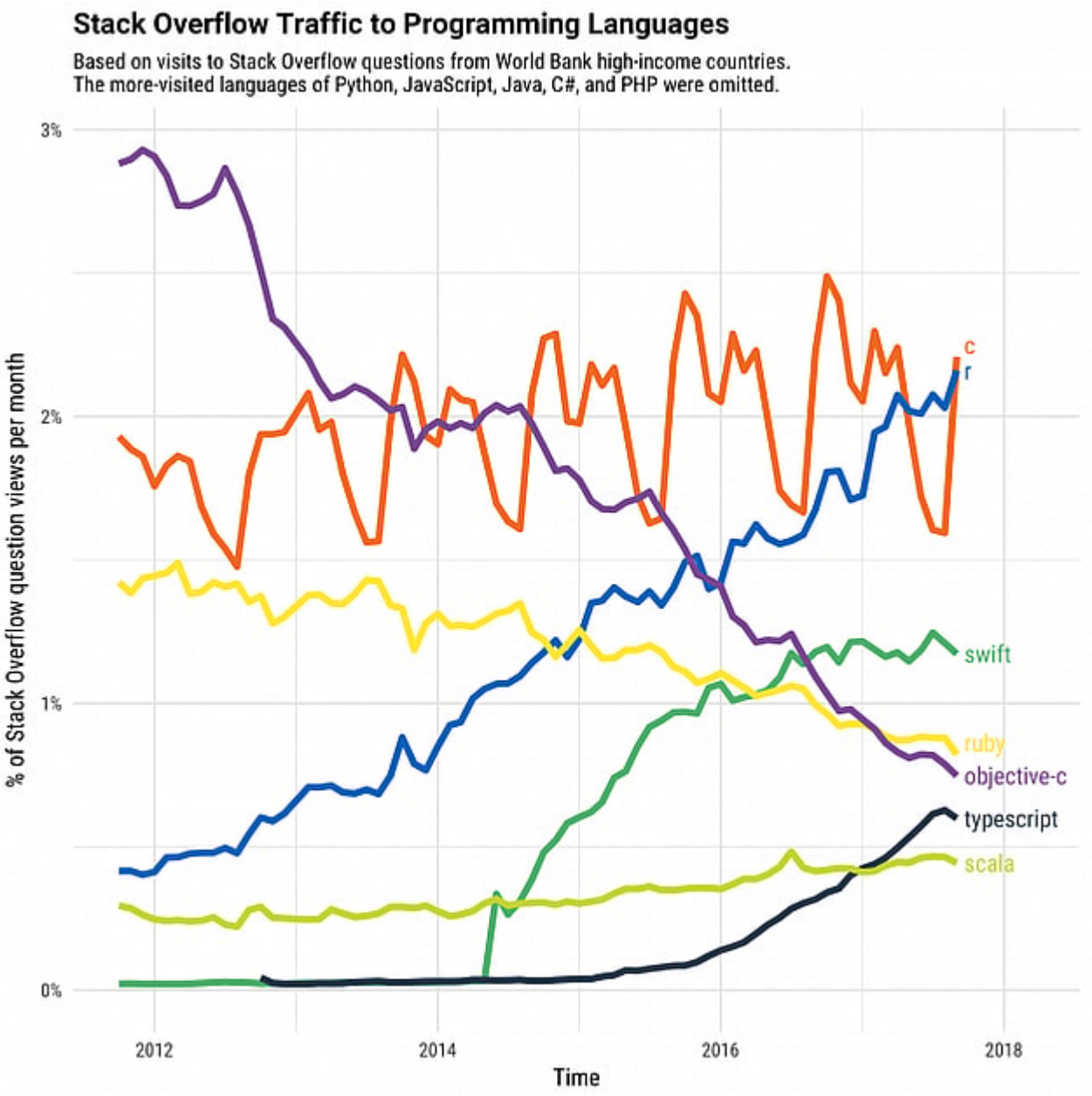


Блокнот Jupyter

Сообщество Python всегда дружелюбно к новичкам. У вас всегда есть форумы и сайты вроде Stack Overflow, чтобы разрешить ваши сомнения.

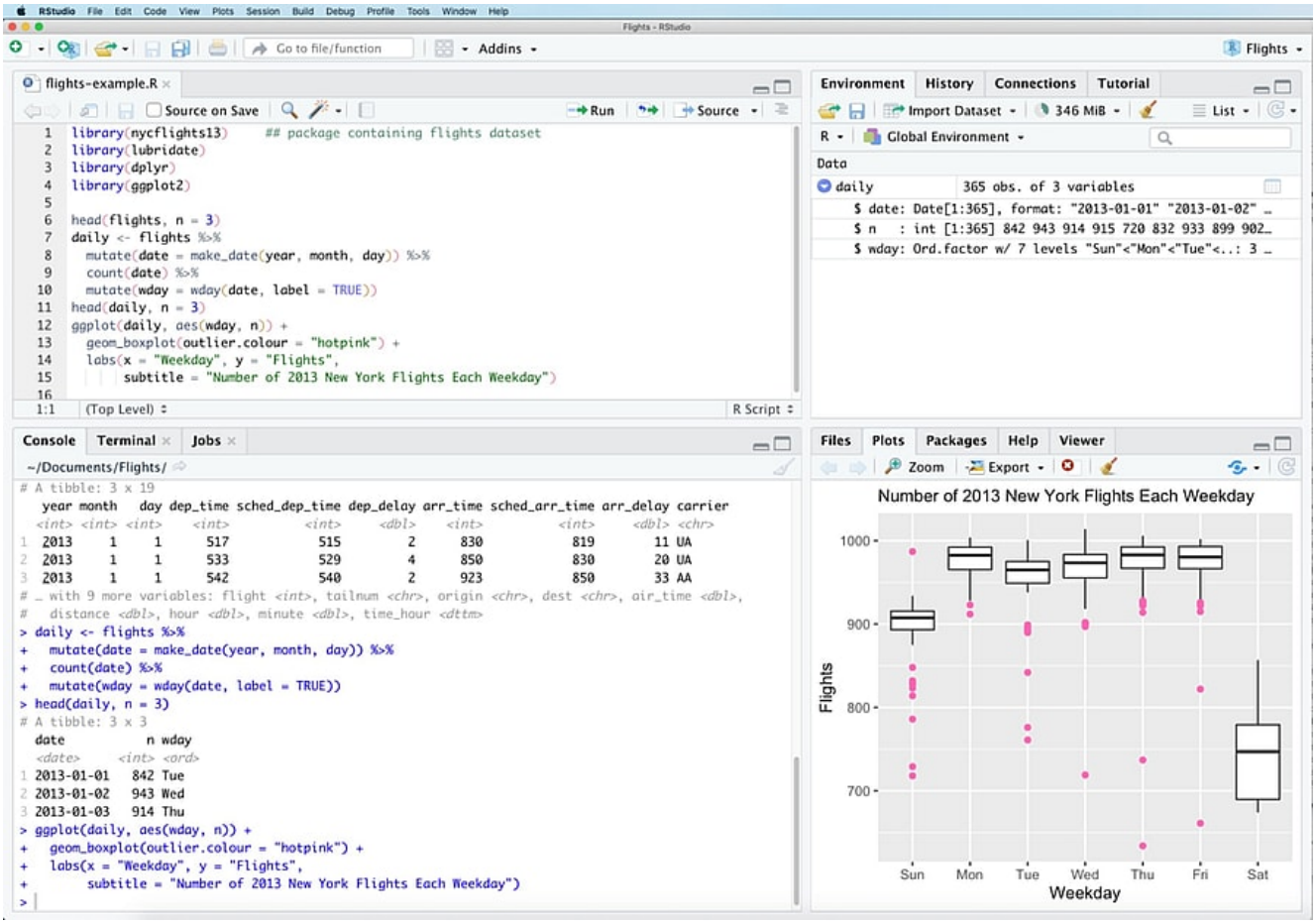
R

R — это язык программирования с открытым исходным кодом, впервые представленный в 1993 году и используемый для статистических вычислений, анализа данных и машинного обучения. Согласно анализу Stack Overflow, популярность R росла на протяжении последних нескольких лет.



Растущая популярность R

Несмотря на то, что R широко используется исследователями, сегодня его применяют и крупные технологические компании, такие как Google, Facebook и Twitter, для анализа данных и статистики. О преимуществах этого языка можно говорить часами. R, как и Python, является интерпретируемым языком, поэтому вы можете выполнять свой код без использования компилятора. В то же время R является кроссплатформенным, поэтому вам не нужно беспокоиться о своей ОС. R – настолько популярный язык, что у вас есть множество редакторов и IDE на выбор. Но на протяжении многих лет RStudio была самой популярной IDE для разработки на R.



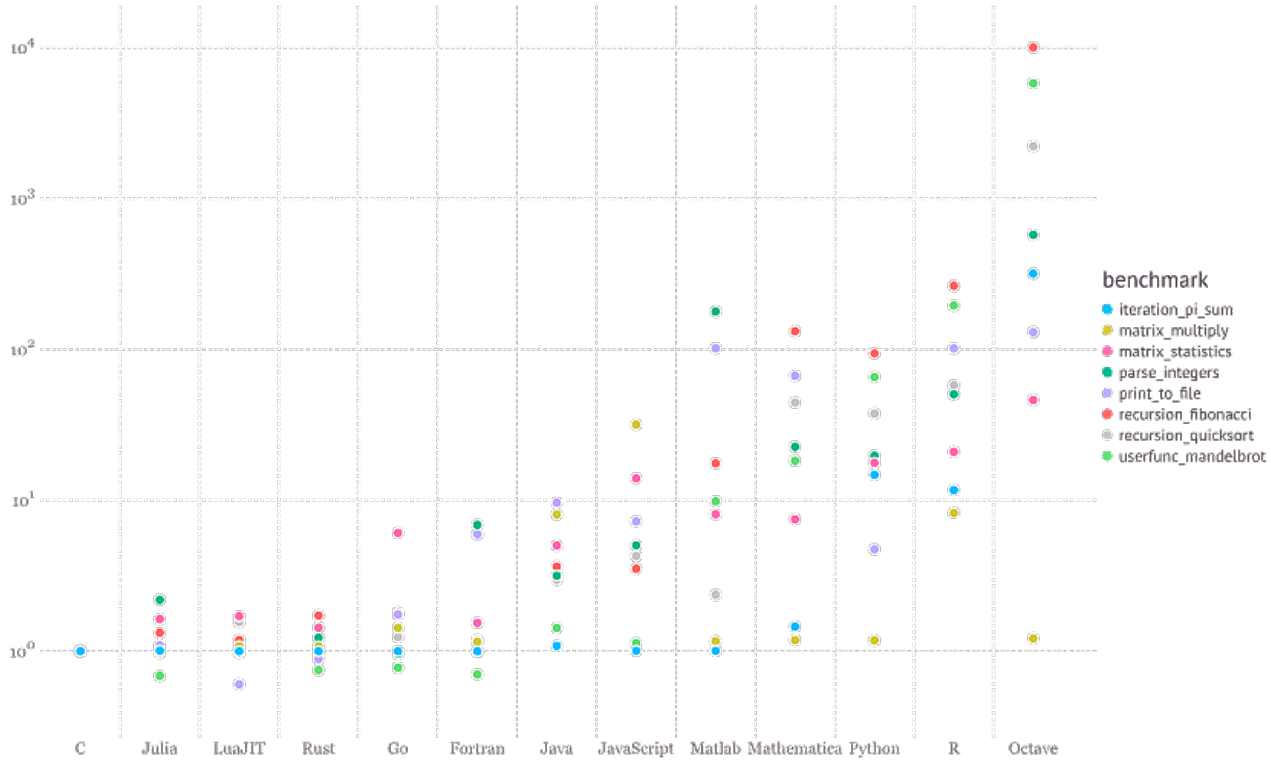
RStudio

Вы можете выйти за рамки обычного использования статистики. С помощью R вы получаете доступ к огромному количеству библиотек, которые позволяют создавать приложения любого типа. Например, с помощью пакета Shiny вы можете разрабатывать эстетичные веб-приложения, не выходя из среды R IDE. Если вы занимаетесь статистикой или исследованиями, использование R не должно вызывать сомнений.

Джулия

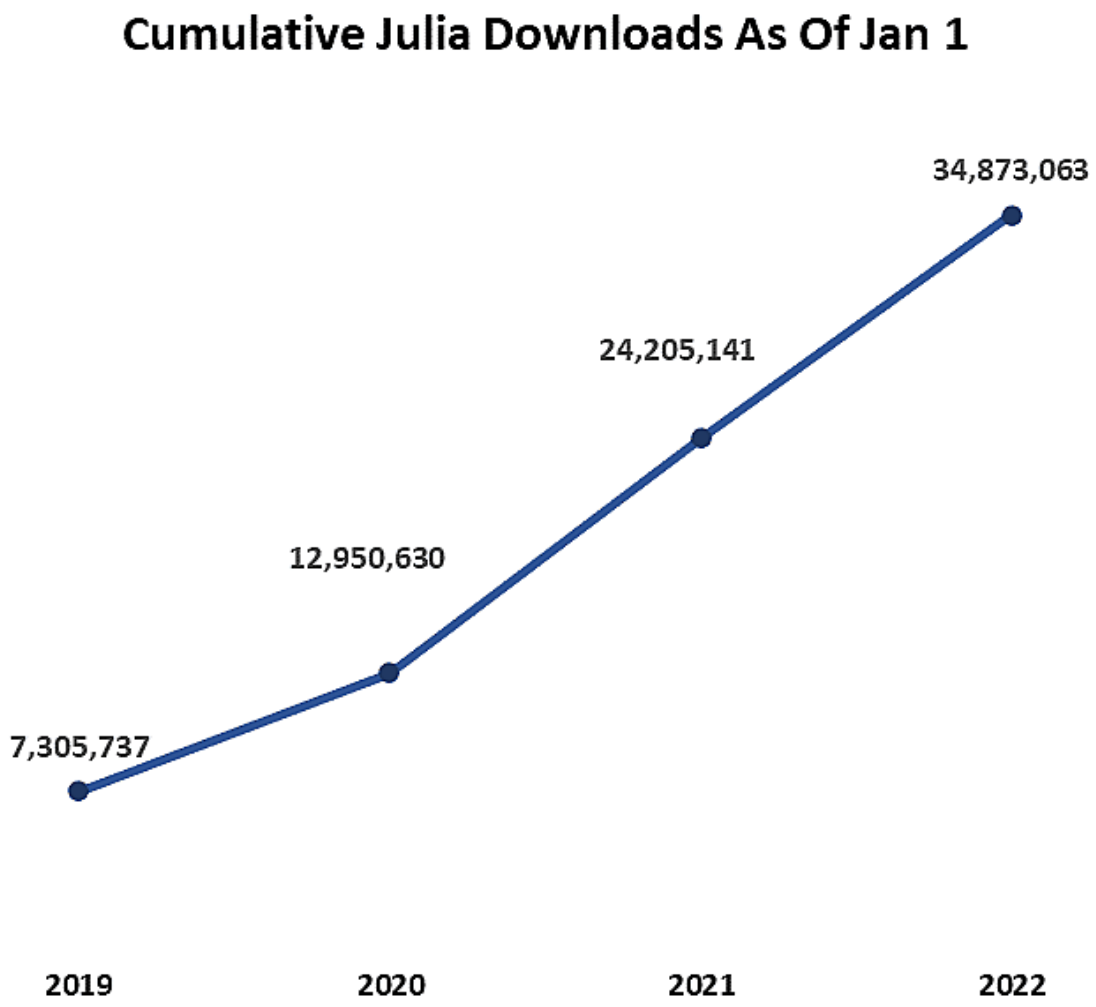
Julia берет лучшее из таких языков, как Python, Ruby, Lisp и R, сочетает это со скоростью языка C и включает привычную математическую нотацию, как в Matlab. Мы можем назвать Julia амбициозной попыткой создать язык, достаточно хороший для общего программирования и в то же время удивительный для конкретных дисциплин информатики, таких как машинное обучение, добыча данных, распределенные и параллельные вычисления. Одно из главных преимуществ Julia — скорость, сопоставимая с такими

языками, как C, Rust, Lua и Go. Это объясняется тем, что он компилируется по принципу Just-In-Time (JIT).



Эталоны Джулии

За последние несколько лет Julia значительно увеличила свою пользовательскую базу. Это видно по количеству загрузок по состоянию на 2022 год.



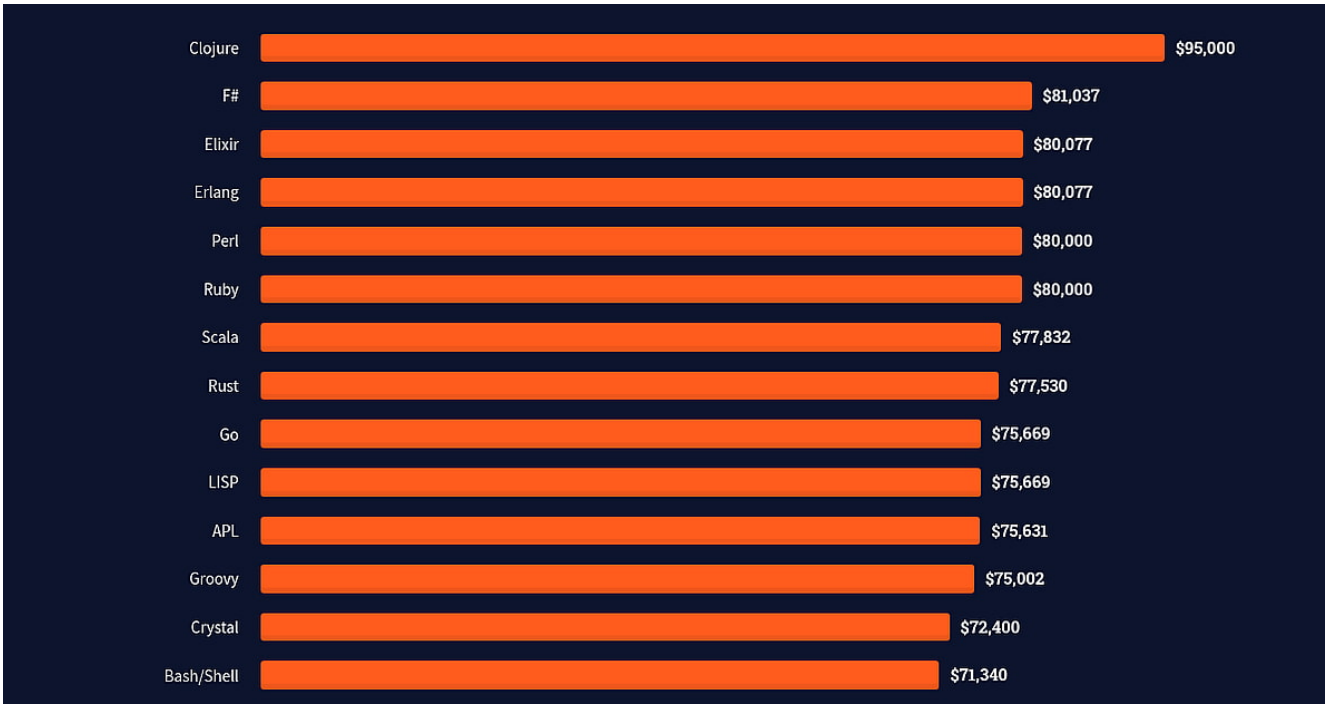
Джулия невероятно хороша в науке о данных, потому что:

- Этот язык легче изучать математикам. В нем используется синтаксис, похожий на математические формулы, которыми пользуются непрограммисты.
- Автоматическое управление памятью с ручным управлением сборщиком мусора.
- Оптимизирован для машинного обучения и статистики.
- Динамическая типизация, почти как в скриптовом языке.
- Множество библиотек Julia для взаимодействия с вашими данными(DataFrames.jl, JuliaGraphs и другие).

Сообщество Джулии настолько энергично, что они создали песню в честь этого языка. Если вам нужен язык с поддержкой науки о данных из коробки, простотой использования Python и скоростью C, Julia – ваш выбор.

Scala

Scala — это язык программирования высокого уровня, впервые представленный в 2004 году и работающий в JVM (Java Virtual Machine) или с JavaScript в вашем браузере. Он был создан для того, чтобы улучшить некоторые аспекты, которые программисты на Java считали утомительными и ограничивающими. Среди этих улучшений — включение функционального программирования в дополнение к уже знакомой объектно-ориентированной парадигме. Кроме того, Scala является более быстрым языком по сравнению с Python или даже самой Java. Многие специалисты по анализу данных включили Scala в свой инструментарий, потому что она неоценима, когда речь идет об анализе больших массивов данных. По данным исследования Stack Overflow 2021, Scala занимает 7-е место среди самых оплачиваемых языков в мире. Но с этой статистикой нужно быть осторожным, так как работа на Scala не так часто встречается в индустрии.

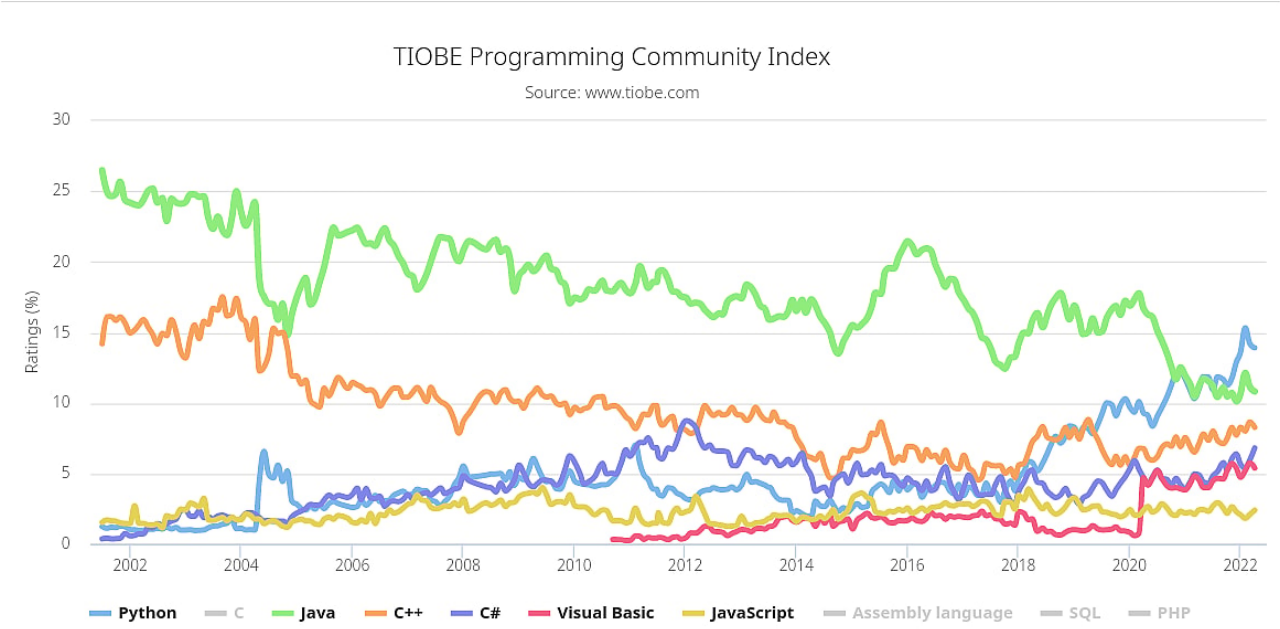


Поскольку Scala работает на JVM, у вас будет доступ к тонне существующих библиотек и некоторым пакетам только для Scala, используемым в области больших данных, математики, баз данных и компьютерных наук в целом. Если вы уже свободно владеете Java, Scala может стать подходящим языком для перехода к науке

о данных. Вот официальный тур , чтобы вы могли начать это приключение прямо сейчас.

Java

Java уже несколько десятилетий является одним из самых распространенных и любимых языков программирования. Это универсальный язык, который можно использовать практически во всех мыслимых ситуациях. Наука о данных — не исключение. Хотя Java в основном используется в мобильных и веб-приложениях, благодаря своей сильной пользовательской базе она применяется вместе с другими популярными фреймворками, такими как Hadoop или Spark, для анализа больших объемов данных. В заключение хотелось бы сказать, что не просто говорить о том, что Java лучше всего подходит для науки о данных, а понимать, что из-за большого количества разработчиков на Java и компаний, которые уже пишут на нем свои программы, удобнее все делать на одном языке.

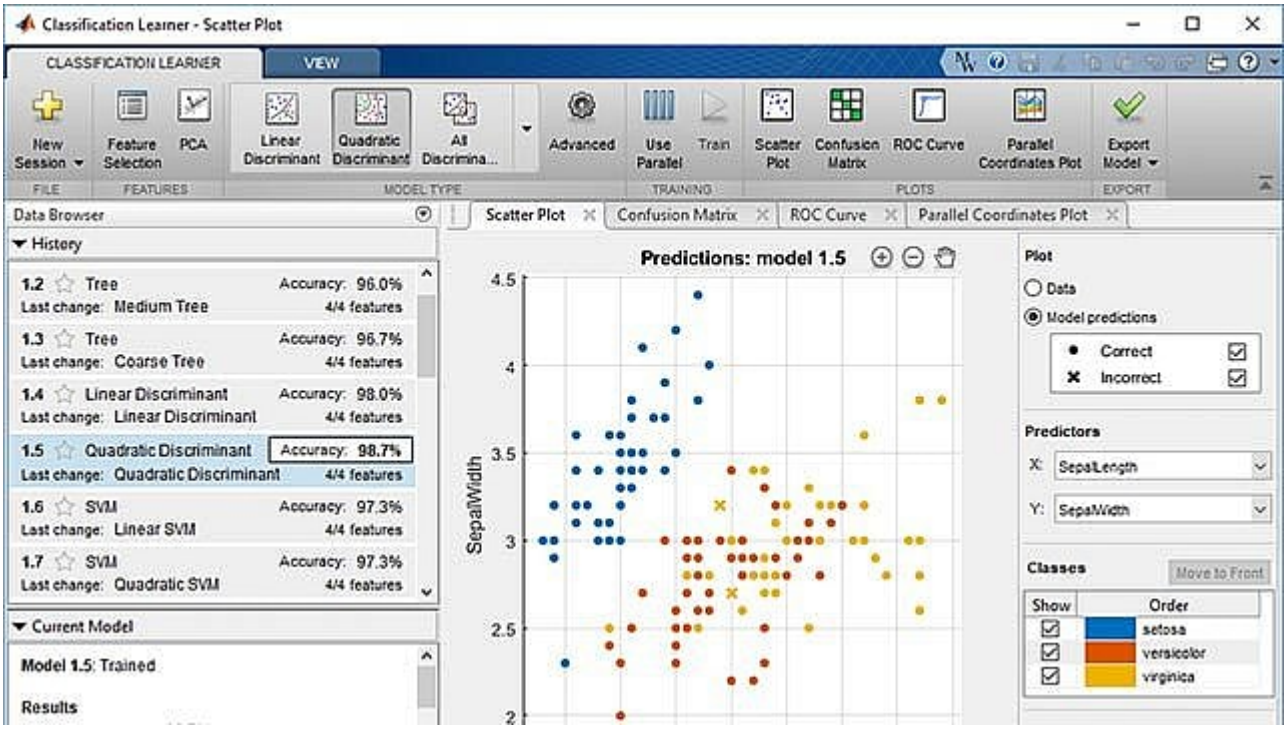


Использование Java на протяжении многих лет

При этом Java можно использовать в большинстве областей науки о данных, таких как управление базами данных, машинное обучение, Если вы знаете Java, гораздо проще освоить пару библиотек, чем изучать использование совершенно другого языка, например R или Julia.

MATLAB

MATLAB – это запатентованный язык программирования, используемый миллионами инженеров и ученых для математических и статистических вычислений.



Ученые, занимающиеся изучением данных, в основном используют этот язык для анализа данных и машинного обучения. Самое приятное – это то, что у вас есть все в одном рабочем пространстве. Он используется в основном в академических кругах, но все же это отличный выбор для создания глубокого фундамента концепций науки о данных. Единственным недостатком MATLAB является то, что это платное программное обеспечение, поэтому вы будете использовать этот язык в основном, если вы учитесь в университете или уже используете его на своей работе. Ознакомьтесь с официальным списком ресурсов MathWorks, чтобы начать обучение уже сегодня.

C++

Завершает этот список язык C++. Хотя он используется в основном для создания приложений и операционных систем, без него мы не увидели бы современного расцвета науки о данных.

Специалисты по исследованию данных предпочитают простые в использовании и отладке языки, такие как Python или R, потому что они не хотят тратить время на исправление странных ошибок на C/C++. Однако C++ играет важную роль в науке о данных, поскольку на нем написаны многие библиотеки, используемые в других языках. Создание модели машинного обучения требует вычислительных усилий, поэтому использование такого эффективного языка, как C++, имеет смысл. Если вы хотите участвовать в индустрии науки о данных, разрабатывая библиотеки для других языков, C++ может стать правильным выбором.

Заключение

В этом посте мы рассмотрели наиболее часто используемые языки программирования для науки о данных. Эта область развивается взрывными темпами, и сегодня самое время начать свою карьеру в качестве специалиста по изучению данных. Если вы только начинаете, я бы рекомендовал вам начать с Python или R. Как только вы получите некоторый опыт создания проектов в реальном мире, вы можете начать расширять свой набор инструментов, изучая другие языки, такие как Julia или Scala.