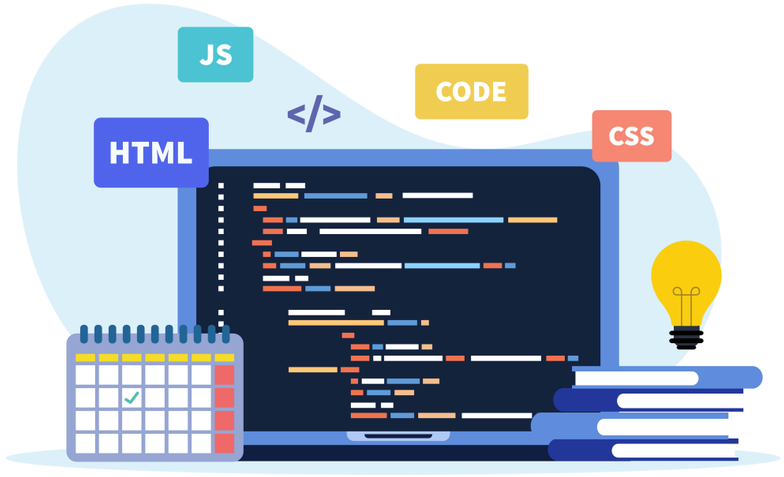


How can you determine which programming languages and technologies are most widely used? Which languages are gaining or losing popularity, helping you decide where to focus your efforts?

One excellent data source is Stack Overflow, a programming question-and-answer site with more than 16 million questions on programming topics. Each Stack Overflow question is tagged with a label identifying its topic or technology. By counting the number of questions related to each technology, you can estimate the popularity of different programming languages.

In this project, you will use data from the Stack Exchange Data Explorer to examine the relative popularity of R compared to other programming languages.

You'll work with a dataset containing one observation per tag per year, including the number of questions for that tag and the total number of questions that year.

`stack_overflow_data.csv`

Column	Description
`year`	The year the question was asked (2008-2020)
`tag`	A word or phrase that describes the topic of the question, such as the programming language
`num_questions`	The number of questions with a certain tag in that year
`year_total`	The total number of questions asked in that year

```
# Load necessary packages
library(readr)
library(dplyr)
library(ggplot2)
```

Hidden output

```r
# Load the dataset
data <- read_csv("stack_overflow_data.csv")
```

Hidden output

```r
# View the dataset
head(data)
glimpse(data)
```

index	year	tag	num_questions	year_total
1	2008	treeview	69	168541
2	2008	scheduled-tasks	30	168541
3	2008	specifications	21	168541
4	2008	rendering	35	168541
5	2008	http-post	6	168541
6	2008	static-assert	1	168541

Rows: 6 Expand

```
Rows: 420,066
Columns: 4
$ year          <dbl> 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 20…
$ tag           <chr> "treeview", "scheduled-tasks", "specifications", "render…
$ num_questions <dbl> 69, 30, 21, 35, 6, 1, 159, 10, 4, 20, 11, 5, 19, 2, 19, …
$ year_total    <dbl> 168541, 168541, 168541, 168541, 168541, 168541, 168541, …
```

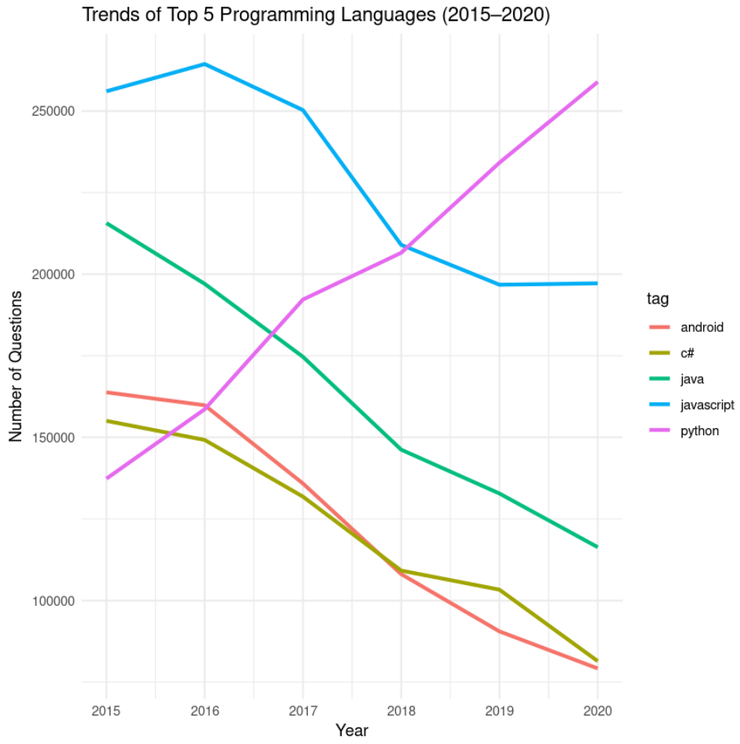
```
r_2020 <- data %>%
  mutate(percentage = (num_questions / year_total * 100)) %>%
  select(year, tag, num_questions, year_total, percentage) %>%
  filter(tag == "r", year == 2020)
r_2020

highest_tags <- data %>%
  filter(year %in% c(2015:2020)) %>%
  group_by(tag)%>%
  summarize(num_questions = sum(num_questions))%>%
  arrange(desc(num_questions))%>%
  slice_head(n=5)

# Filter original data to only top 5 tags
data_filtered <- data %>%
  filter(tag %in% highest_tags$tag, year %in% 2015:2020)

# Line plot over time
ggplot(data_filtered, aes(x = year, y = num_questions, color = tag)) +
  geom_line(linewidth = 1.2) +
  labs(title = "Trends of Top 5 Programming Languages (2015-2020)",
       x = "Year",
       y = "Number of Questions") +
  theme_minimal()
highest_tags
```

index	year	tag	num_questions	year_total	percentage
1	2020	r	52662	5452545	0.9658

Rows: 1 Expand

index	tag	num_questions
1	javascript	1373634
2	python	1187838
3	java	982747
4	android	737330
5	c#	730045

Rows: 5 Expand


Trends of Top 5 Programming Languages (2015–2020)
Number of Questions
250000
200000
150000
100000
2015
2016
2017
2018
2019
2020
Year
tag
android
c#
java
javascript
python