data on tags over time

Roger Wilson

adapted from datacamp.com

How can we tell what programming languages and technologies are used by the most people? How about what languages are growing and which are shrinking, so that we can tell which are most worth investing time in?

We will be looking at data from Stack Overflow, specifically Explorer to examine the relative popularity of languages like R, Python, Java and Javascript have changed over time.

By looking at the number of question tags associated with each language, we will analyze overall popularity trends.

Our dataset has one observation for each tag in each year. The dataset includes both the number of questions asked in that tag in that year, and the total number of questions asked in that year.

In [47]:
```
# load libraries
library(readr)
library(dplyr)
# Load dataset
by_tag_year <- read_csv('by_tag_year.csv')

head(by_tag_year)
dim(by_tag_year)
```

```
-- Column specification --------------------------------------------------------
cols(
  year = col_double(),
  tag = col_character(),
  number = col_double(),
  year_total = col_double()
)
```

year	tag	number	year_total
2008	.htaccess	54	58390
2008	.net	5910	58390
2008	.net-2.0	289	58390
2008	.net-3.5	319	58390
2008	.net-4.0	6	58390
2008	.net-assembly	3	58390

1. 40518
2. 4

adding a percentage column

In [17]:
```
# add fraction column
```

```
by_tag_year_fraction <- by_tag_year %>%
    mutate(fraction = number / year_total)

head(by_tag_year_fraction)
```

year	tag	number	year_total	fraction
2008	.htaccess	54	58390	9.248159e-04
2008	.net	5910	58390	1.012160e-01
2008	.net-2.0	289	58390	4.949478e-03
2008	.net-3.5	319	58390	5.463264e-03
2008	.net-4.0	6	58390	1.027573e-04
2008	.net-assembly	3	58390	5.137866e-05

popularity of R

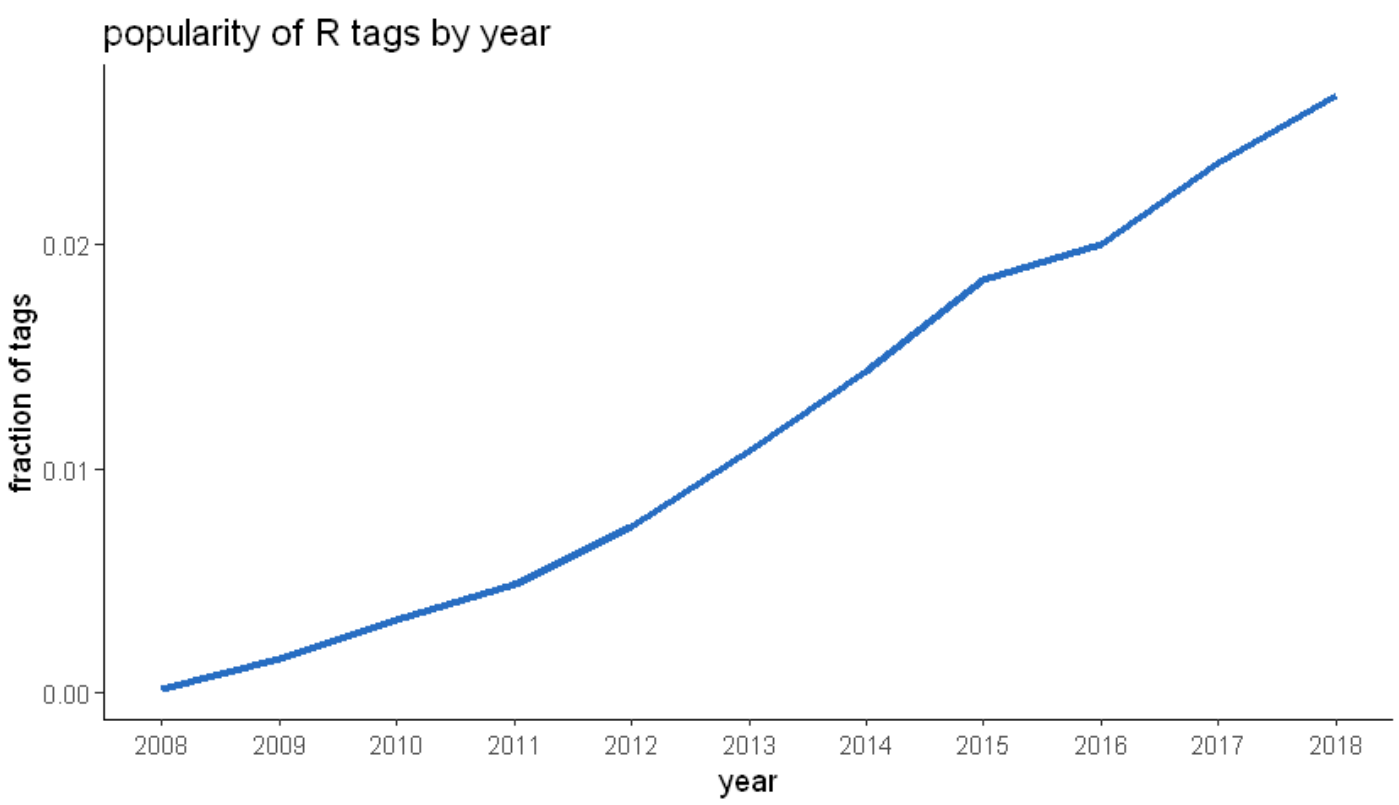
Let's first measure the popularity of R over time, because it's what we are using :D

In [19]:
```
# filter for r tags
r_over_time <- by_tag_year_fraction %>%
    filter(tag == 'r')

head(r_over_time)
```

year	tag	number	year_total	fraction
2008	r	8	58390	0.0001370098
2009	r	524	343868	0.0015238405
2010	r	2270	694391	0.0032690516
2011	r	5845	1200551	0.0048685978
2012	r	12221	1645404	0.0074273552
2013	r	22329	2060473	0.0108368321

visualizing change over time

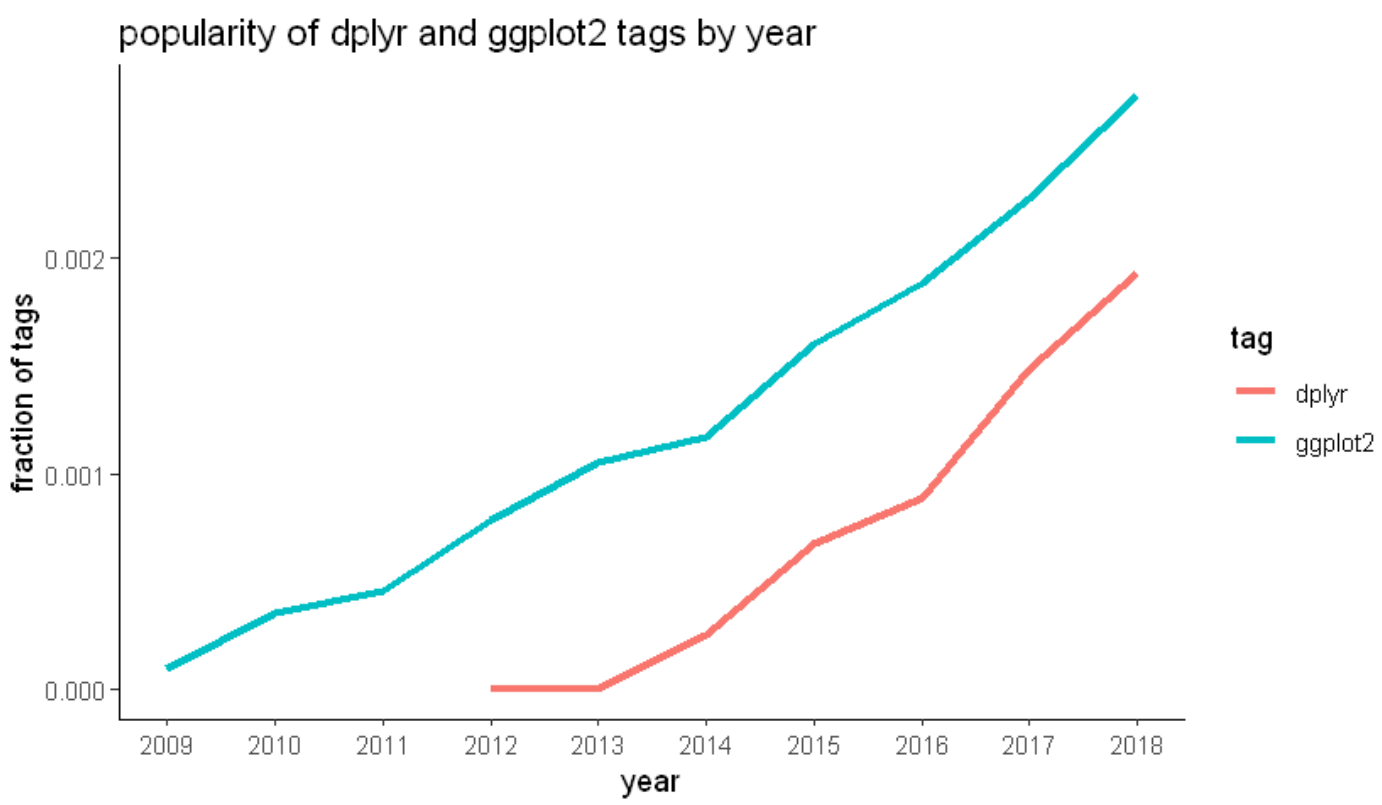
In [92]:
```
library('ggplot2')

# plotting . . .
options(repr.plot.width = 7, repr.plot.height = 4)
ggplot(r_over_time) +
    geom_line(aes(x = year, y = fraction), size = 1.25, color = '#276DC2') +
    scale_x_continuous(breaks = 0:2100) +
    labs(title = 'popularity of R tags by year') +
    ylab('fraction of tags') +
    theme_classic()
```



comparing dplyr and ggplot2

In [91]:

```
# a vector of selected tags
selected_tags <- c('dplyr', 'ggplot2')

# filtering for these tags
selected_tags_over_time <- by_tag_year_fraction %>%
    filter(tag %in% selected_tags)

# plotting . . .
ggplot(selected_tags_over_time) +
    geom_line(aes(x = year, y = fraction, color = tag), size = 1.25) +
    scale_x_continuous(breaks = 0:2100) +
    labs(title = 'popularity of dplyr and ggplot2 tags by year') +
    ylab('fraction of tags') +
    theme_classic()
```

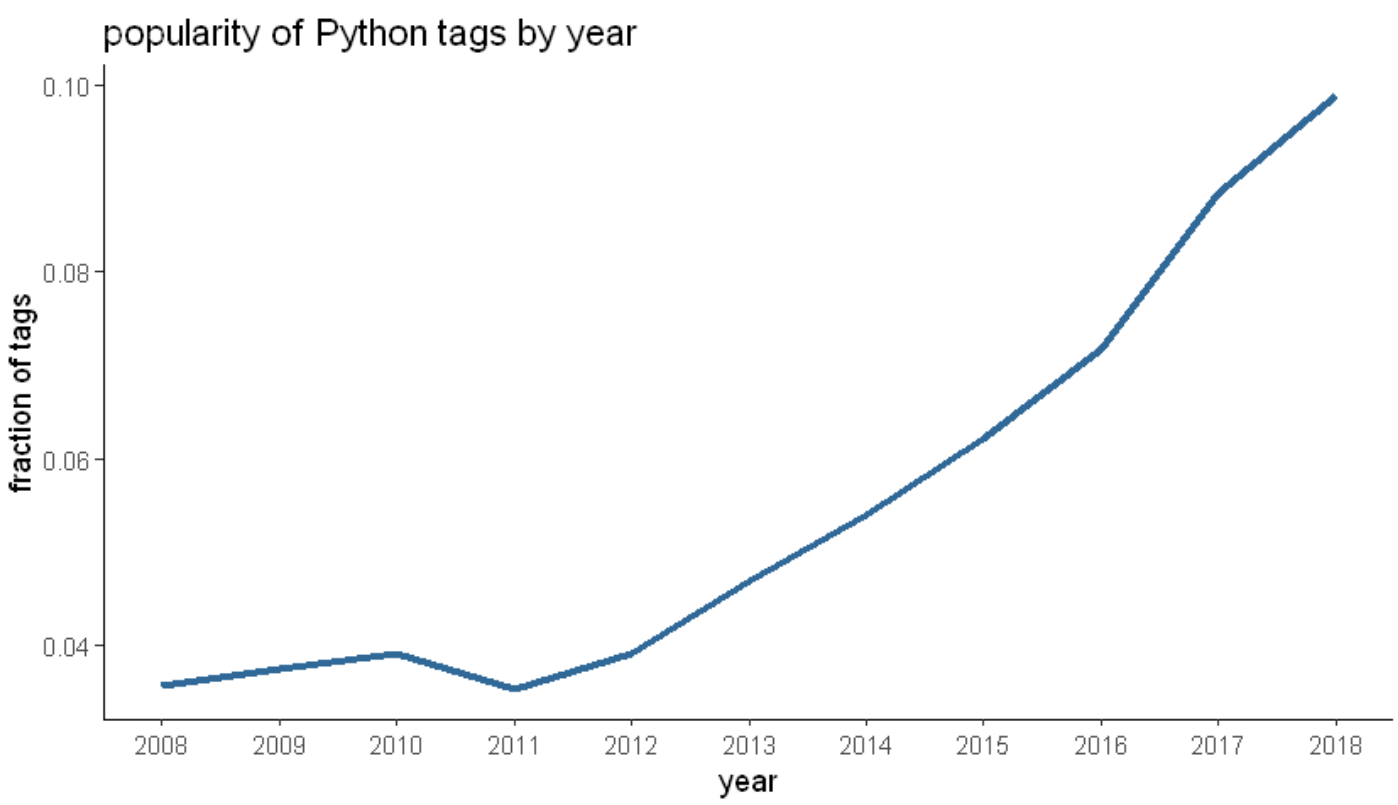


popularity of Python

In [44]:
```
# filter for python tags
python_over_time <- by_tag_year_fraction %>%
    filter(tag == 'python')
head(python_over_time)
```

year	tag	number	year_total	fraction
2008	python	2080	58390	0.03562254
2009	python	12906	343868	0.03753184
2010	python	27098	694391	0.03902412
2011	python	42313	1200551	0.03524465
2012	python	64456	1645404	0.03917336
2013	python	96803	2060473	0.04698096

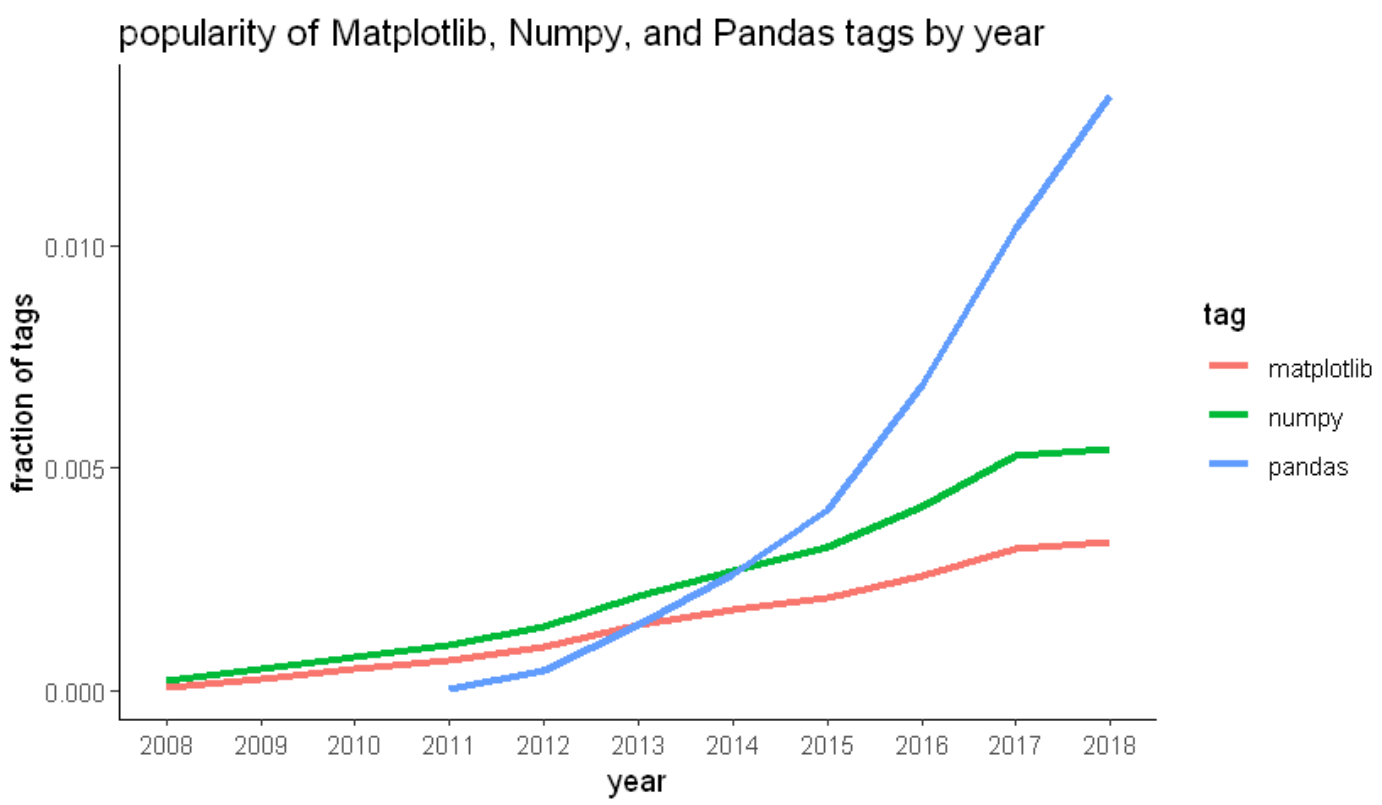
In [90]:
```
# plotting . . .
ggplot(python_over_time) +
    geom_line(aes(x = year, y = fraction), size = 1.25, color = '#306998') +
    scale_x_continuous(breaks = 0:2100) +
    labs(title = 'popularity of Python tags by year') +
    ylab('fraction of tags') +
    theme_classic()
```



adding in pandas, matplotlib and numpy

In [89]:

```
# a vector of selected tags
python_selected_tags <- c('pandas', 'numpy', 'matplotlib')

# filtering for these tags
python_selected_tags_over_time <- by_tag_year_fraction %>%
    filter(tag %in% python_selected_tags)

# plotting . . .
ggplot(python_selected_tags_over_time) +
    geom_line(aes(x = year, y = fraction, color = tag), size = 1.25) +
    scale_x_continuous(breaks = 0:2100) +
    labs(title = 'popularity of Matplotlib, Numpy, and Pandas tags by year') +
    ylab('fraction of tags') +
    theme_classic()
```



an overall comparison

In [39]:

```
# finding the total number of questions for each tag
sorted_tags <- by_tag_year %>%
# .... YOUR CODE FOR TASK 6 ....
    group_by(tag) %>%
    summarize(tag_total = sum(number)) %>%
    arrange(desc(tag_total))

head(sorted_tags)
```

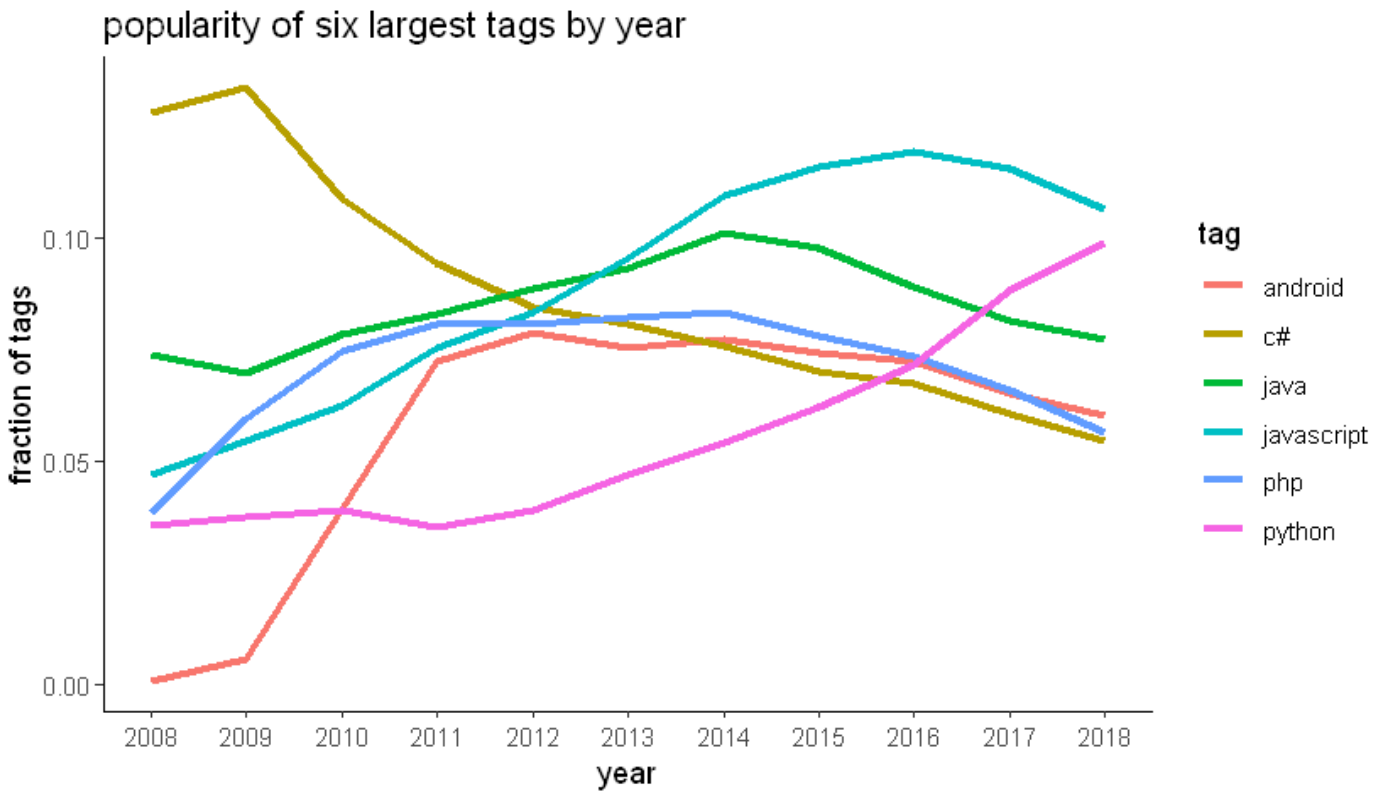
tag	tag_total
javascript	1632049
java	1425961
c#	1217450
php	1204291
android	1110261
python	970768

how have large programming languages changed over time?

In [88]:

```
# get the six largest tags
highest_tags <- head(sorted_tags$tag)

# filter for the six largest tags
by_tag_subset <- by_tag_year_fraction %>%
    filter(tag %in% highest_tags)

# plotting . . .
ggplot(by_tag_subset) +
```

```
    geom_line(aes(x = year, y = fraction, color = tag), size = 1.25) +
    scale_x_continuous(breaks = 0:2100) +
    labs(title = 'popularity of six largest tags by year') +
    ylab('fraction of tags') +
    theme_classic()
```



final thoughts

This project helped me put my dplyr skills to practice. I always rely on Stack Overflow for any intriguing question, so this project was also a great way to visualize the popularity of various languages through that community's lense. It was also interesing to compare the popularity of R and Python over time. While I have a better understanding of R, I can see why the use of Python has skyrocketed in recent years.

Stack Overflow tags are a great way track coding language popularity. This dataset only has records up to 2017, so if I ever were to build more upon this project, I would try to look into more recent years. I bet the demand for coding help has exponentially grown from even just 5 years ago (it's equally as crazy to think that 2017 was five years ago O.O)