Lab 8

Pranav Bhagat

11/16/2019

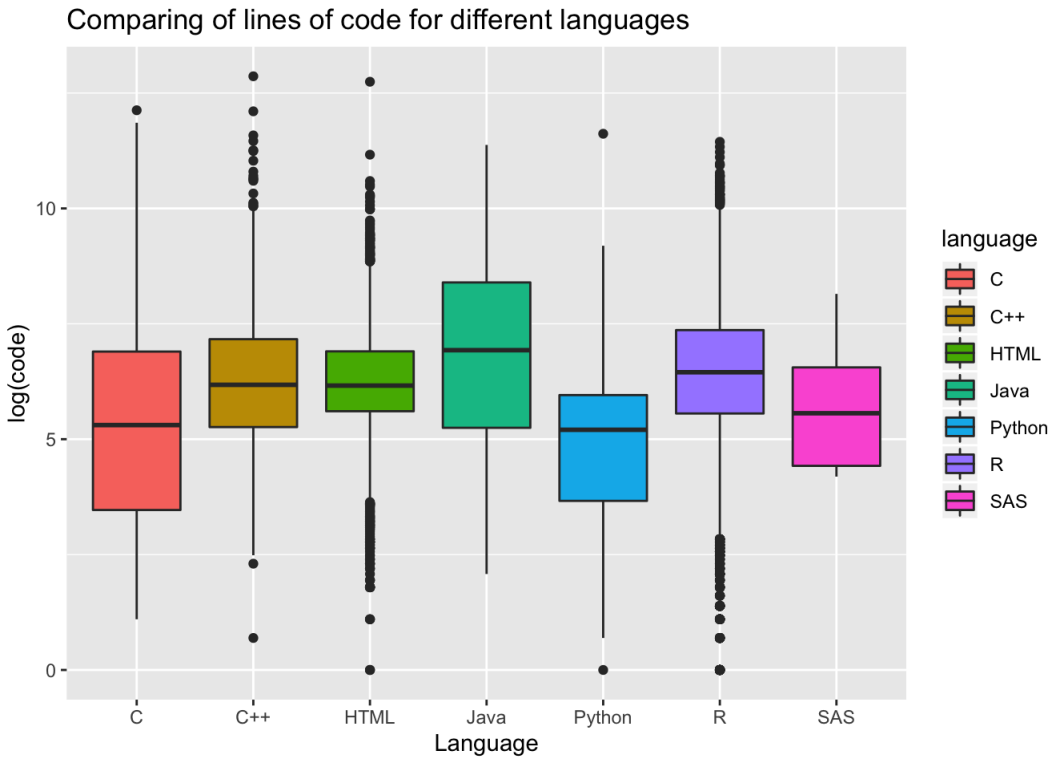
```
cran_code <- readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-11-12/loc_cran_packages.csv")
```

```
## Parsed with column specification:
## cols(
##   file = col_double(),
##   language = col_character(),
##   blank = col_double(),
##   comment = col_double(),
##   code = col_double(),
##   pkg_name = col_character(),
##   version = col_character()
## )
```

```
nofilter <- filter(cran_code, language == "C" | language == "Java" | language == "Javascript" | language == "HTML" | language == "Python" | language == "C++" | language == "R" | language == "SAS")
```

```
ggplot(nofilter, aes(x = as.factor(language), y = log(code), fill = language))+
  geom_boxplot()+
  ggtitle("Comparing of lines of code for different languages")+
  xlab("Language")
```

```
## Warning: Removed 3 rows containing non-finite values (stat_boxplot).
```



For the selected languages, Java has the highest median useage of code whereas C has the lowest. In addition, C also has the greatest IQR and the greatest variability, while HTML has the lowest. Also, while both R and HTML contain a lot of outliers, HTML has more extreme values.