Lab 1 - Checkpoint 4

Joab de Araújo

6 de abril de 2018

```
suppressMessages(library("tidyverse"))
suppressMessages(library("here"))
library(tidyverse)
library(here)
library(knitr)
library(ggplot2)
theme_set(theme_bw())
```

```
projetos = read_csv(here::here("data/projetos.csv"))
```

```
## Parsed with column specification:
## cols(
##   gh_project_name = col_character(),
##   team = col_double(),
##   lang = col_character(),
##   sloc_end = col_integer(),
##   sloc_med = col_double(),
##   activity_period = col_integer(),
##   num_commits = col_integer(),
##   commits_per_month = col_double(),
##   tests_per_kloc = col_double(),
##   total_builds = col_integer(),
##   build_success_prop = col_double(),
##   builds_per_month = col_double(),
##   tests_added_per_build = col_double(),
##   tests_successful = col_double(),
##   test_density = col_double(),
##   test_size_avg = col_double()
## )
```

```
projetos = projetos %>%
    filter(lang != "javascript")
```

Nesse relatório serão apresentado os resultados obtidos, para responder os seguintes questionamentos:

1. O tamanho da equipe influencia no total de commits por mês?
2. No geral, em relação ao tempo de atividade do projeto, o número de commits por mês aumenta?

Para responder ambas as perguntas foram utilizados as variáveis:

- **commits_per_month** e **team**, para responder a primeira pergunta;
- **activity_period** e **commits_per_month**, para responder a segunda pergunta.

Antes de apresentar os resultados, a seguir será apresentado uma breve descrição de cada variável utilizada:

- **commits_per_month**: É a média mensal de commits;
- **team**: Tamanho do time(número de desenvolvedores) de cada projeto;
- **activity_period**: É o período de atividade do projeto

Essas variáveis foram obtidas do site TravisTorrent e toda a análise dos dados foi feito usando a linguagem R.

Respondendo a primeira pergunta

Afim de observar se o tamanho do time influencia no número de commits por mês, optou-se por usar o gráfico de Dispersão, e com uso de uma linha que representa a média, pode-se observar que em java não pode-se dizer que exista alguma relação entre o tamanho do time e o número de commits por mês.

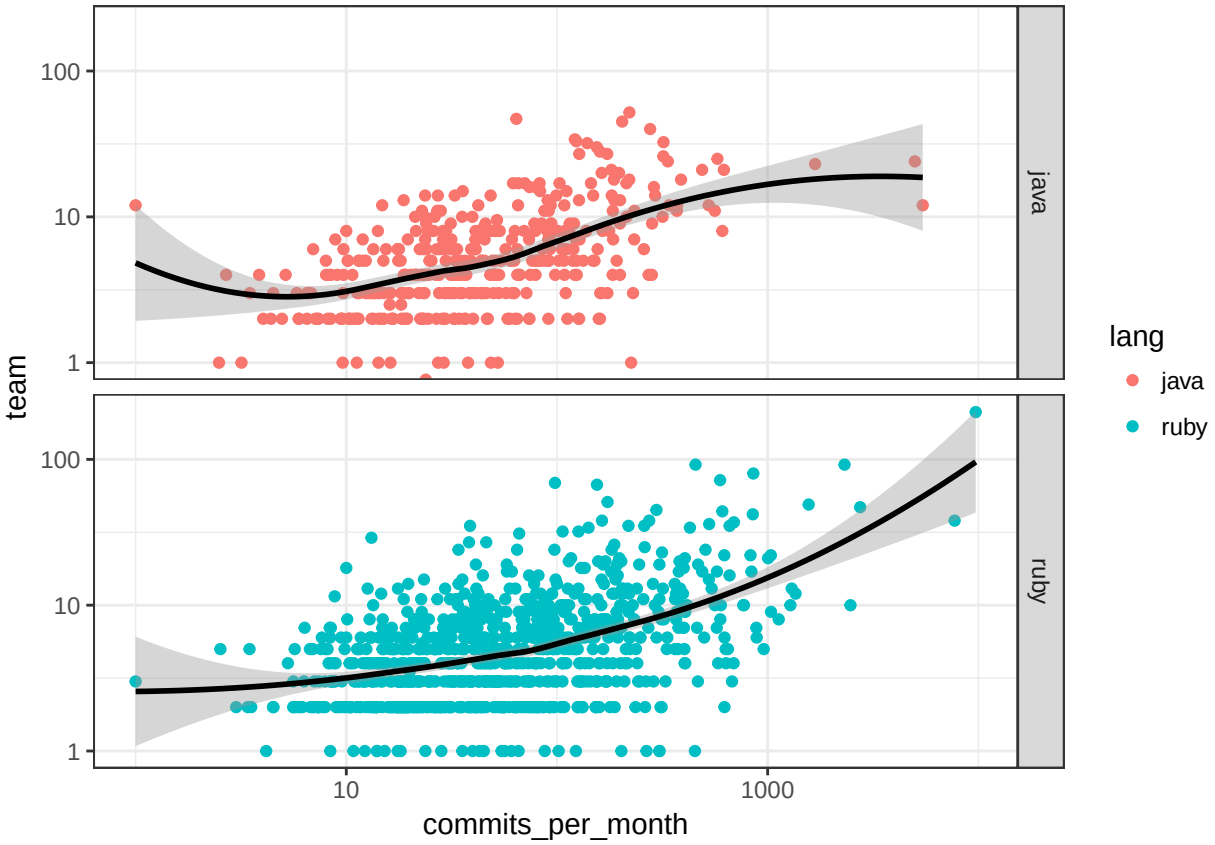
Por outro lado é possível observar uma leve correlação, já que a média sobe lentamente e se aproximando do fim há uma subida realtivamente rápida.

```
ggplot(projetos, aes(commits_per_month, team, colour=lang)) +
    geom_point() +
    geom_smooth(method = 'loess', colour="black") +
    facet_grid(lang ~ .)+
    scale_x_log10() +
    scale_y_log10()
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```



A seguir na tabela pode-se ver o resultado do cálculo de correlação de Pearson, e ver que há uma correlação considerável.

```
projetos %>%
  group_by(lang) %>%
 summarise(
     pearson = cor(team, commits_per_month, method="pearson")
 )
```

```
## # A tibble: 2 x 2
##   lang  pearson
##   <chr>   <dbl>
## 1 java    0.243
## 2 ruby    0.642
```

Ainda sobre a primeira pergunta

Uma outra pergunta que pode ser feita é:

Porque há uma relação maior para Ruby?

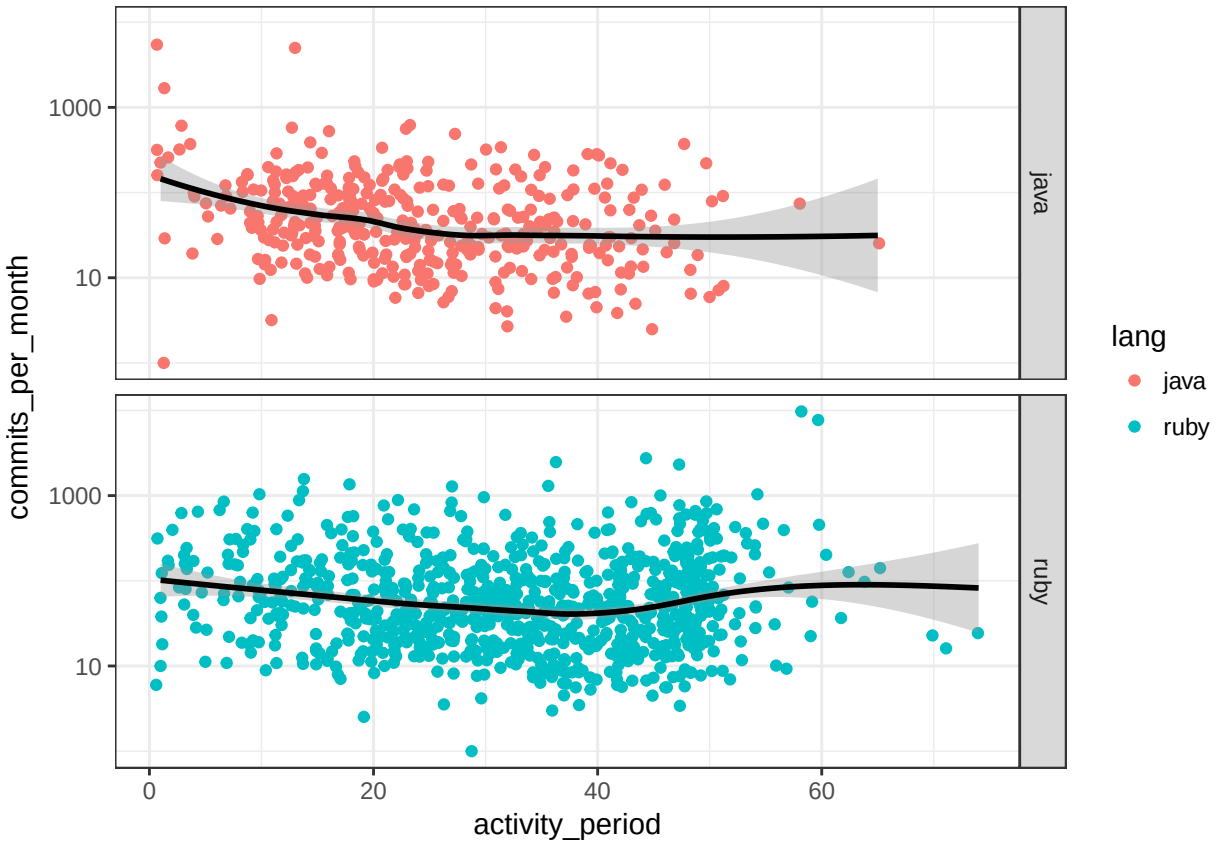
Não há como afirmar com certeza de 100%, mas pode-se supor que seja porque os maiores times são de projetos Ruby e ainda mais existem mais projetos Ruby, e no gráfico pode-se ver que na faixa de tamanho dos times vai de 0 a 10, ambas linguagens tem comportamentos semelhantes, e após isso o comportamento muda, enquanto ruby aumenta o tamanho dos times rapidamente, java mantém a sua média.

Com isso podemos supor que a correlação acontece para projetos com times maiores, enquanto em times menores essa correção aparentemente não existe.

Respondendo a segunda pergunta

Seguindo a mesma lógica que foi usado na pergunta anterior, agora em relação ao tempo de atividade do projeto, pode-se ver que pouco influencia o tempo de atividade do projeto, com o número de commits por mês.

```
ggplot(projetos, aes(activity_period, commits_per_month, colour=lang)) +
    geom_jitter() +
    facet_grid(lang ~ .) +
    geom_smooth(method = 'loess', colour="black") +
    scale_y_log10()
```



E na tavela seguinte com o cálculo de correlação de Pearson, pode-se ver claramente que não há praticamente nenhuma realação.

```
projetos %>%
  group_by(lang) %>%
 summarise(
     pearson = cor(activity_period, commits_per_month, method="pearson")

    )
```

```
## # A tibble: 2 x 2
##   lang  pearson
##   <chr>   <dbl>
## 1 java  -0.163
## 2 ruby   0.0686
```

Ainda sobre a segunda pergunta

Uma outra pergunta que pode ser feita é:

Porque não há relação?

Muitos projetos chegam ao ponto de não ter mais tanta necessidade de mudanças ou de tantas alterações, logo não necessita de tantos commits. Então pode-se dizer que provavelmente não exista relação, pois os projetos podem ter chegado a um nível de amadurecimento em que os commits se estabeleceram em uma média.