Characterization of the PRs

omitted

2024-04-05

Data overview

```
glimpse(prs_characterization)
```

```
## Rows: 2,035
## Columns: 10
## $ pr_number           <int> 1302, 1304, 1313, 1322, 1329, 1332, 1343, 1346, 13~
## $ repo                <chr> "accumulo", "accumulo", "accumulo", "accumulo", "a~
## $ created_at          <dttm> 2019-07-31 19:34:39, 2019-08-01 18:52:50, 2019-08~
## $ merged_at           <dttm> 2019-08-02 15:59:29, 2019-08-01 20:56:42, 2019-08~
## $ additions           <int> 235, 89, 555, 112, 404, 1, 20, 14, 3, 9, 19, 27, 3~
## $ deletions           <int> 348, 91, 521, 109, 255, 1, 26, 30, 2, 8, 17, 31, 1~
## $ changed_files_count <int> 15, 5, 20, 5, 16, 1, 1, 1, 1, 1, 2, 1, 7, 7, 1, 11~
## $ commits_count       <int> 4, 2, 3, 1, 4, 1, 1, 2, 1, 1, 1, 3, 4, 5, 1, 3, 1,~
## $ code_churn          <int> 583, 180, 1076, 221, 659, 2, 46, 44, 5, 17, 36, 58~
## $ duration            <dbl> 1.85057870, 0.08601852, 6.99309028, 1.35975694, 9.~
```

PRs count

```
prs_characterization %>% count()
```

```
##      n
## 1 2035
```

PRs count by repo

```
prs_characterization %>%
    group_by(repo) %>%
    count()
```

```
## # A tibble: 12 x 2
## # Groups:   repo [12]
##    repo                      n
##    <chr>                 <int>
```

```
##  1 accumulo                994
##  2 cayenne                  55
##  3 commons-collections      31
##  4 commons-io               76
##  5 commons-lang            128
##  6 helix                   278
##  7 httpcomponents-client   126
##  8 maven-surefire           25
##  9 opennlp                  96
## 10 struts                  145
## 11 wicket                   47
## 12 zookeeper                34
```

Commits count

```
sum(prs_characterization$commits_count)
```

```
## [1] 4370
```

Commits count by repo

```
prs_characterization %>%
    group_by(repo) %>%
    summarise(sum(commits_count))
```

```
## # A tibble: 12 x 2
##    repo                 `sum(commits_count)`
##    <chr>                              <int>
##  1 accumulo                            2040
##  2 cayenne                              150
##  3 commons-collections                   61
##  4 commons-io                           192
##  5 commons-lang                         214
##  6 helix                                812
##  7 httpcomponents-client                211
##  8 maven-surefire                        37
##  9 opennlp                              118
## 10 struts                               341
## 11 wicket                                96
## 12 zookeeper                             98
```

General data statistics by repo

```
prs_characterization %>%
    group_by(repo) %>%
    summarise(
        min_date = min(merged_at),
```

```
        max_date = max(merged_at),
        age_in_days = difftime(min(merged_at), max(merged_at), units = "days")
    ) %>%
    mutate(
    age_in_days = abs(as.numeric(age_in_days)),
    age_in_years = abs(as.numeric(age_in_days))/365.25
    )
```

```
## # A tibble: 12 x 5
##    repo         min_date            max_date            age_in_days age_in_years
##    <chr>        <dttm>              <dttm>                    <dbl>        <dbl>
##  1 accumulo     2019-08-01 20:56:42 2024-02-29 23:40:37       1673.         4.58
##  2 cayenne      2021-12-07 10:13:35 2024-02-09 08:12:54        794.         2.17
##  3 commons-col~ 2019-01-20 03:23:46 2024-01-20 15:02:23       1826.         5.00
##  4 commons-io   2016-12-02 02:08:46 2024-02-09 22:53:32       2626.         7.19
##  5 commons-lang 2016-11-22 21:25:59 2024-02-11 14:18:57       2637.         7.22
##  6 helix        2020-10-06 21:58:38 2024-02-27 18:14:32       1239.         3.39
##  7 httpcompone~ 2018-11-14 19:18:37 2024-02-29 16:12:35       1933.         5.29
##  8 maven-suref~ 2021-11-18 22:43:22 2023-08-23 16:41:09        643.         1.76
##  9 opennlp      2017-01-25 14:53:25 2024-01-13 13:53:26       2544.         6.96
## 10 struts       2018-11-20 11:45:50 2024-02-16 07:29:55       1914.         5.24
## 11 wicket       2018-08-01 03:29:48 2024-02-26 11:33:05       2035.         5.57
## 12 zookeeper    2022-01-26 12:23:10 2024-02-12 18:40:34        747.         2.05
```

```
prs_characterization %>%
    group_by(repo) %>%
    summarise(
        mean_commits = mean(commits_count),
        median_commits = median(commits_count),
        mean_changed_files = mean(changed_files_count),
        median_changed_files = median(changed_files_count)
    )
```

```
## # A tibble: 12 x 5
##    repo      mean_commits median_commits mean_changed_files median_changed_files
##    <chr>            <dbl>          <dbl>              <dbl>                <dbl>
##  1 accumulo          2.05              1               9.43                    2
##  2 cayenne           2.73              1               8.13                    4
##  3 commons-~         1.97              1              13.0                     2
##  4 commons-~         2.53              1.5             3.18                    2
##  5 commons-~         1.67              1               3.19                    2
##  6 helix             2.92              2               5.23                    3
##  7 httpcomp~         1.67              1               7.18                    2
##  8 maven-su~         1.48              1              21.8                     3
##  9 opennlp           1.23              1              12.0                     3.5
## 10 struts            2.35              1              12.1                     4
## 11 wicket            2.04              1               5.15                    2
## 12 zookeeper         2.88              1.5             5.59                    2
```

```
prs_characterization %>%
    group_by(repo) %>%
    summarise(
```

```
        mean_code_churn = mean(code_churn),
        median_code_churn = median(code_churn),
        mean_duration = mean(duration)
    )
```

```
## # A tibble: 12 x 4
##    repo                   mean_code_churn median_code_churn mean_duration
##    <chr>                            <dbl>             <dbl>         <dbl>
##  1 accumulo                         323.               40            3.18
##  2 cayenne                          389.               75            5.67
##  3 commons-collections              134.               32           41.3
##  4 commons-io                        74.3              28           12.5
##  5 commons-lang                      88.8              12           15.4
##  6 helix                            199.               64            7.23
##  7 httpcomponents-client            237.               22            1.36
##  8 maven-surefire                   439.               22           20.9
##  9 opennlp                          753.               86            2.26
## 10 struts                           995.              109            6.21
## 11 wicket                           136.               58            5.00
## 12 zookeeper                        232.               47.5         73.9
```

NCLOC and classes count by repo

- All values are related to the last commit of the last PR that was successfully executed by SonarQube.

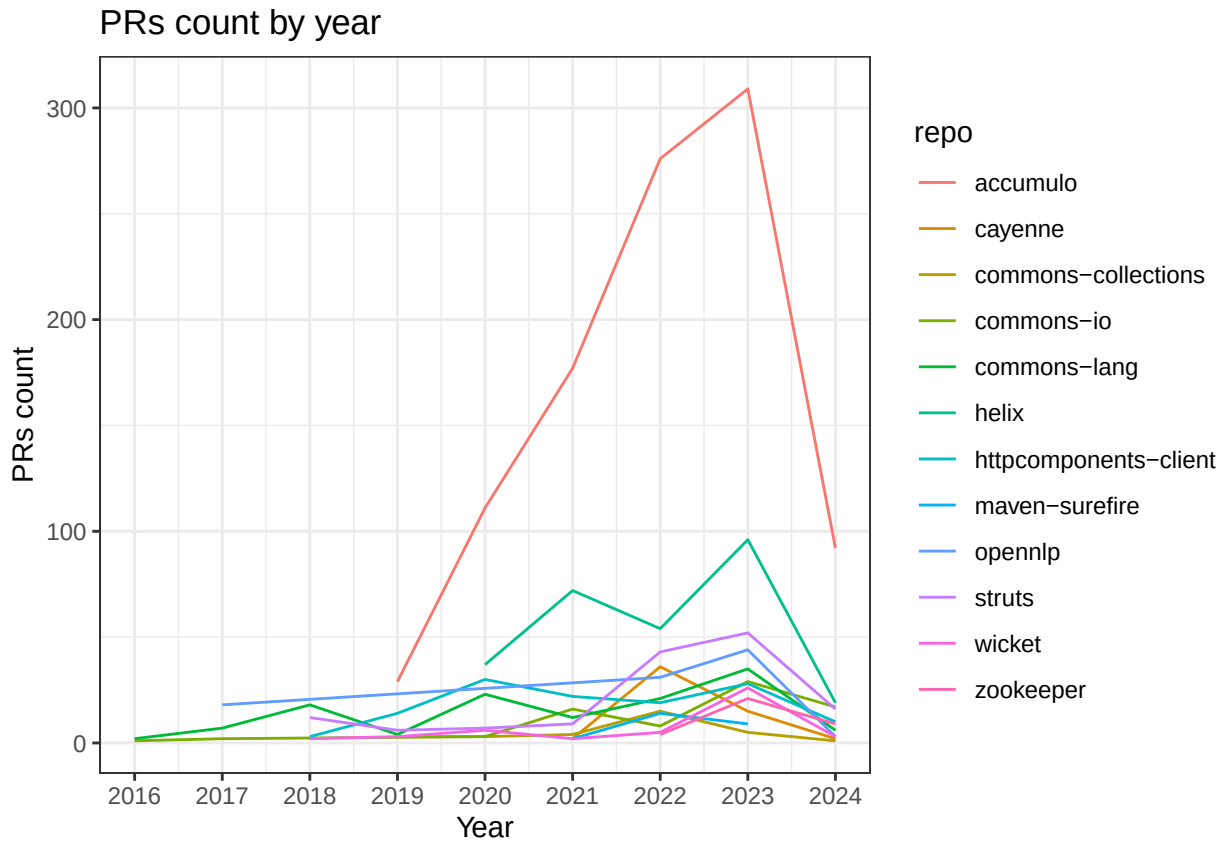
```
data.frame(repo = c("accumulo", "cayenne", "commons-collections","commons-io", "commons-lang", "helix",
```

```
##                    repo  NCLOC classes
## 1              accumulo 440441    5164
## 2               cayenne 318428    4716
## 3    commons-collections  67690     839
## 4           commons-io  30501     288
## 5          commons-lang  95929     918
## 6                helix 189487    2106
## 7  httpcomponents-client  76964     879
## 8         maven-surefire 110481    3036
## 9               opennlp 155900    2478
## 10               struts 234331    3419
## 11               wicket 251505    5254
## 12            zookeeper 131712    1542
```

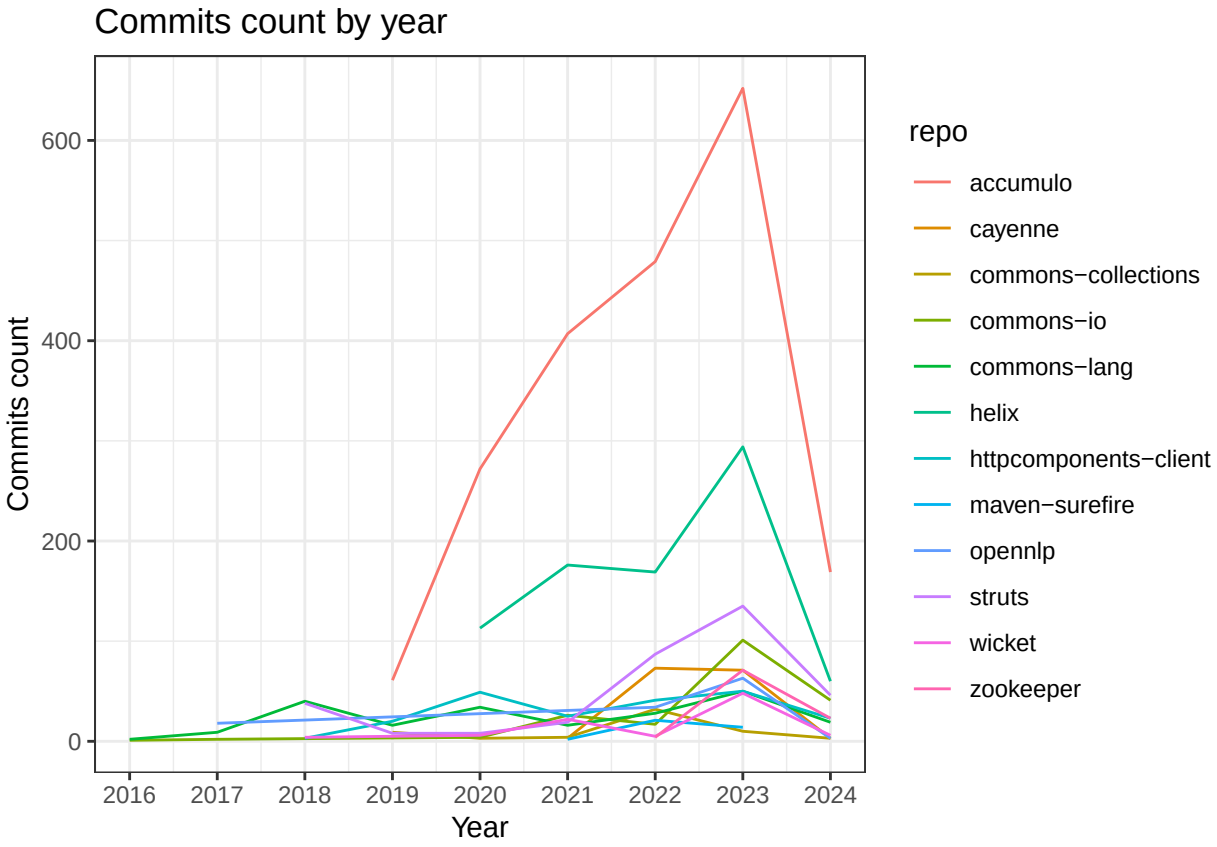
PRs and commits count through the years

```
prs_characterization %>%
    mutate(year = lubridate::year(merged_at)) %>%
    group_by(year, repo) %>%
    summarise(PRs_quantity = n(), .groups = 'drop') %>%
    ggplot(aes(x = year, y = PRs_quantity, color=repo)) +
    geom_line() +
```

```
    labs(x = "Year", y = "PRs count") +
    ggtitle("PRs count by year") +
    scale_x_continuous(breaks = unique(lubridate::year(prs_characterization$merged_at)))
```




```
prs_characterization %>%
    mutate(year = lubridate::year(merged_at)) %>%
    group_by(year, repo) %>%
    summarise(commits_quantity = sum(commits_count), .groups = 'drop') %>%
    ggplot(aes(x = year, y = commits_quantity, color=repo)) +
    geom_line() +
    labs(x = "Year", y = "Commits count") +
    ggtitle("Commits count by year") +
    scale_x_continuous(breaks = unique(lubridate::year(prs_characterization$merged_at)))
```



Distribution of features

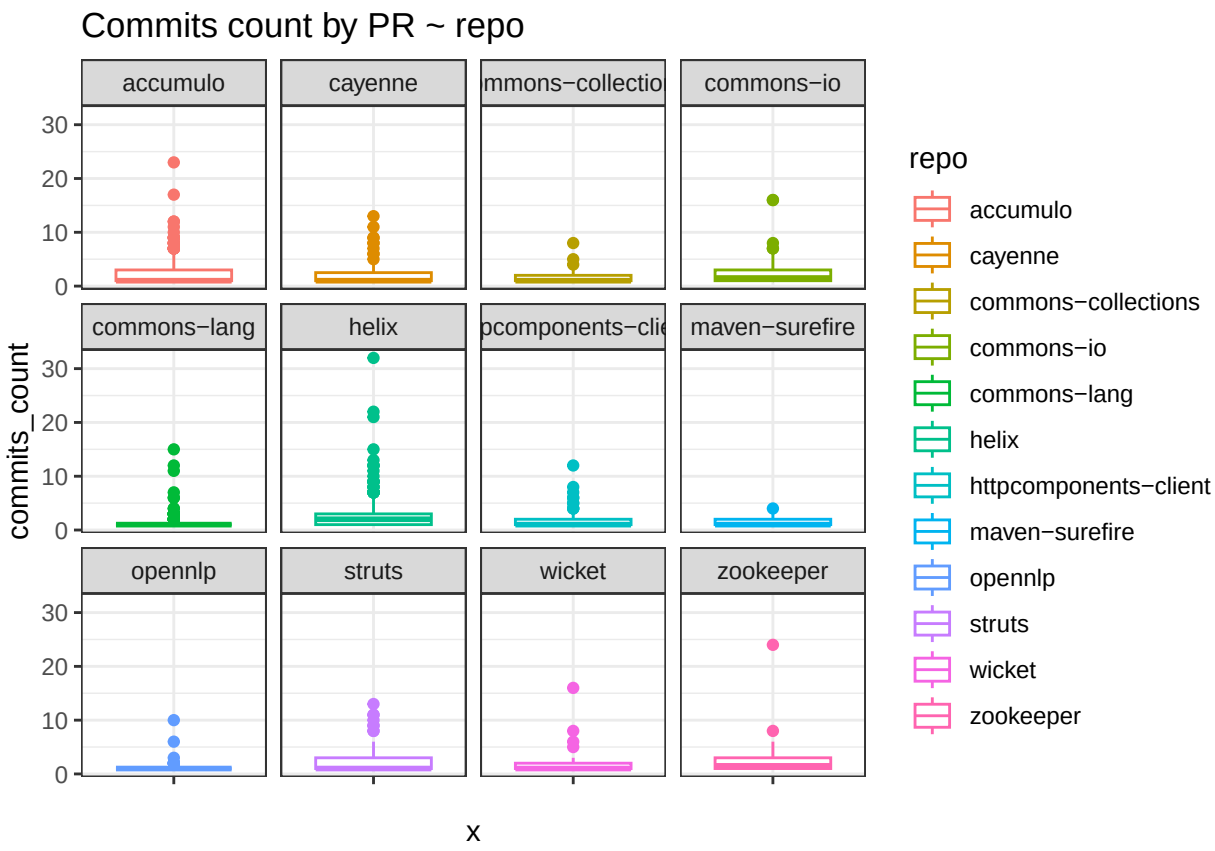
Commits count by PR

```
tapply(prs_characterization$commits_count, prs_characterization$repo, summary)
```

```
## $accumulo
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   2.052   3.000  23.000
##
## $cayenne
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   2.727   2.500  13.000
##
## $`commons-collections`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.968   2.000   8.000
##
## $`commons-io`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.500   2.526   3.000  16.000
##
```

```
## $`commons-lang`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.672   1.000  15.000
##
## $helix
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.921   3.000  32.000
##
## $`httpcomponents-client`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.675   2.000  12.000
##
## $`maven-surefire`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    1.00    1.48    2.00    4.00
##
## $opennlp
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.229   1.000  10.000
##
## $struts
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   2.352   3.000  13.000
##
## $wicket
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   2.043   2.000  16.000
##
## $zookeeper
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.500   2.882   3.000  24.000
```

```
prs_characterization %>%
    ggplot(aes(x = "", y = commits_count, color=repo)) +
      geom_boxplot() +
      facet_wrap(~repo) +
      labs(title = "Commits count by PR ~ repo")
```



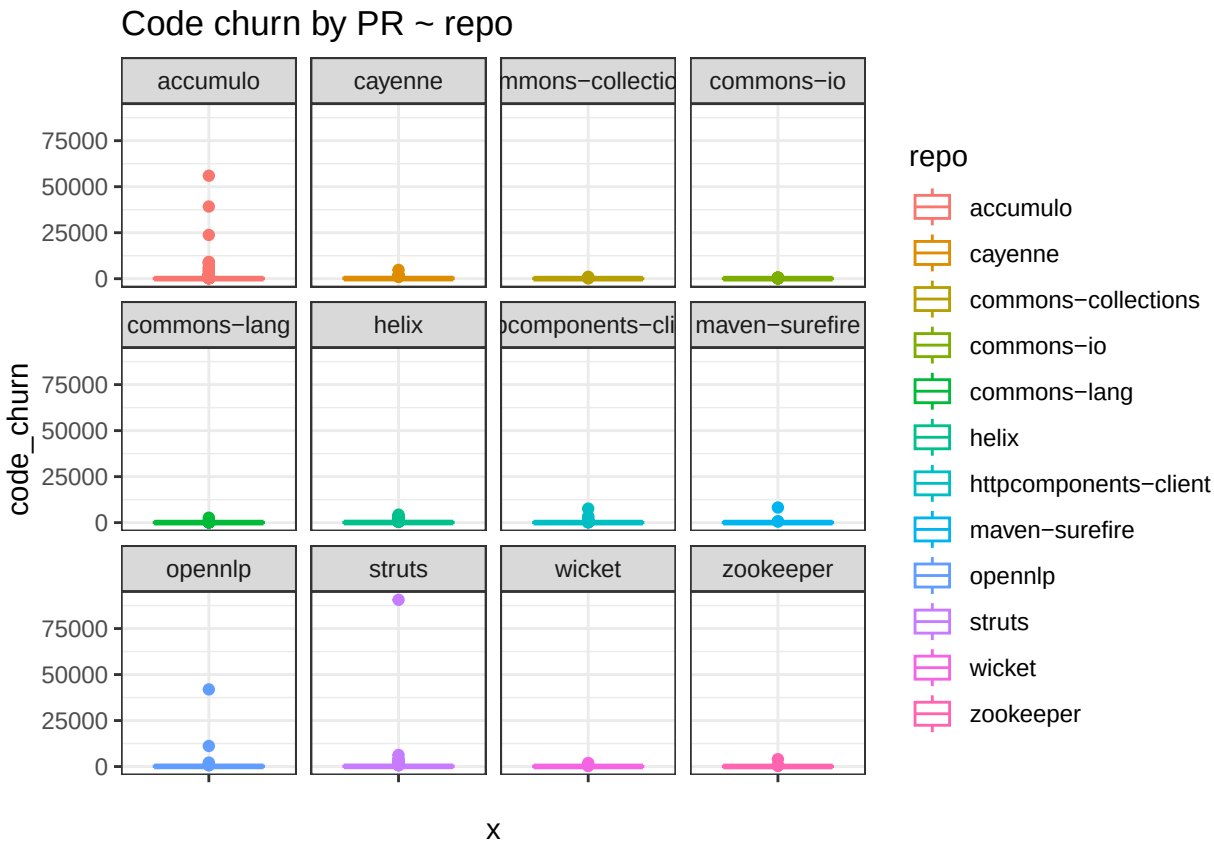
Code churn by PR

```
tapply(prs_characterization$code_churn, prs_characterization$repo, summary)
```

```
## $accumulo
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1      11      40     323     146   55972
##
## $cayenne
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    28.0    75.0   389.2   411.5  4793.0
##
## $`commons-collections`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0     5.5    32.0   133.6    73.5  1035.0
##
## $`commons-io`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   28.00   74.32   60.50  631.00
##
## $`commons-lang`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    4.00   12.00   88.83   46.00 2696.00
```

```
##
## $helix
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    18.0    64.0   198.6   188.2  4343.0
##
## $`httpcomponents-client`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     5.0    22.0   237.5    96.5  7633.0
##
## $`maven-surefire`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0     8.0    22.0   438.9   168.0  8229.0
##
## $opennlp
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     2.00    17.75    86.00   753.38   270.00 41938.00
##
## $struts
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    36.0   109.0   995.1   299.0 90610.0
##
## $wicket
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0    19.5    58.0   136.4   122.5  1945.0
##
## $zookeeper
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0     4.5    47.5   231.9   130.2  4005.0
```

```
prs_characterization %>%
    ggplot(aes(x = "", y = code_churn, color=repo)) +
      geom_boxplot() +
      facet_wrap(~repo) +
      labs(title = "Code churn by PR ~ repo")
```



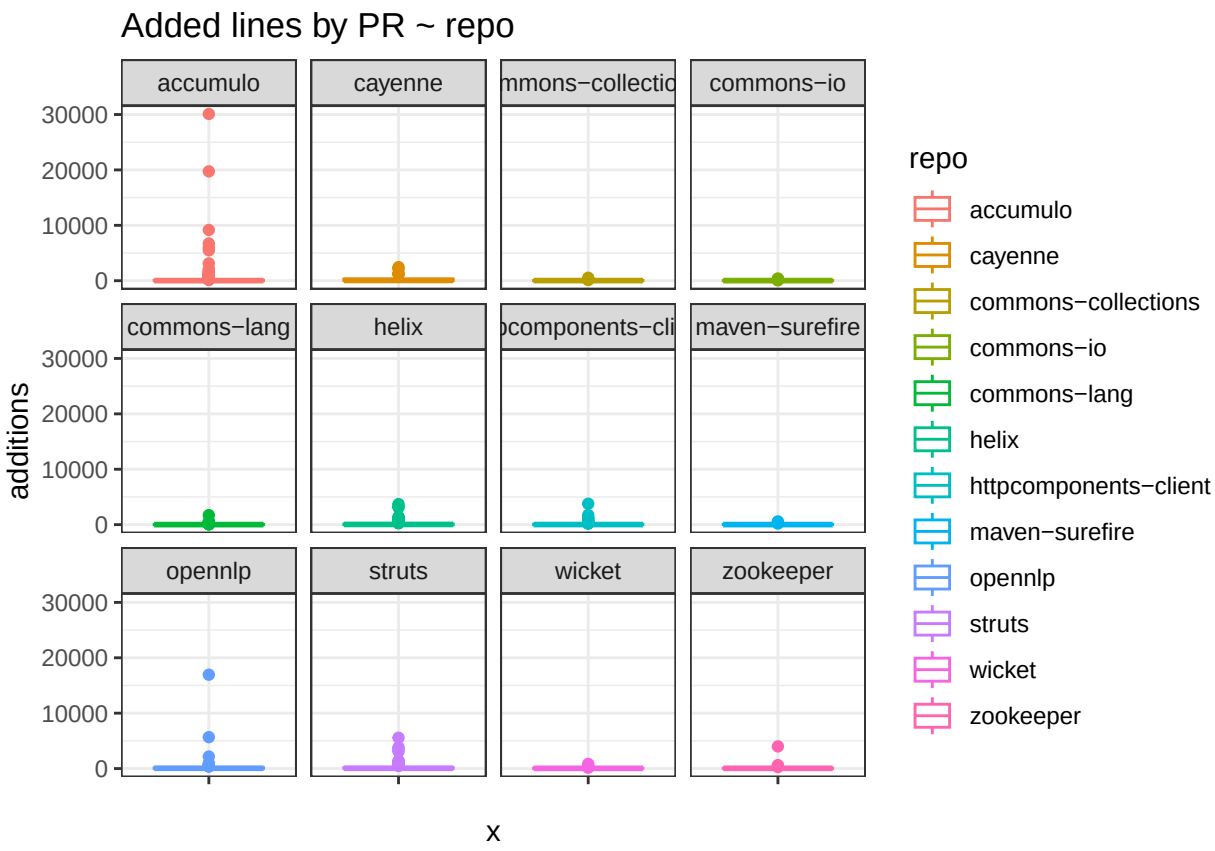
Added lines

```
tapply(prs_characterization$additions, prs_characterization$repo, summary)
```

```
## $accumulo
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.00     5.00    21.00   167.02    83.75 30103.00
##
## $cayenne
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    19.0    54.0   278.5   307.5  2447.0
##
## $`commons-collections`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     2.0    23.0    54.1    40.0   536.0
##
## $`commons-io`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    4.00   19.00   41.62   43.00  399.00
##
## $`commons-lang`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    2.00    6.00   52.49   27.00 1717.00
```

```
##
## $helix
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    11.0    36.0   152.6   119.0  3718.0
##
## $`httpcomponents-client`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    2.00   13.00  133.55   78.25 3783.00
##
## $`maven-surefire`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    3.00    9.00   70.44   47.00  579.00
##
## $opennlp
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.00     8.75    55.50   371.28   156.25 16944.00
##
## $struts
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    15.0    61.0   251.7   188.0  5565.0
##
## $wicket
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   10.50   38.00   85.28   76.00  848.00
##
## $zookeeper
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.00   42.50  197.15   94.75 4004.00
```

```
prs_characterization %>%
    ggplot(aes(x = "", y = additions, color=repo)) +
      geom_boxplot() +
      facet_wrap(~repo) +
      labs(title = "Added lines by PR ~ repo")
```



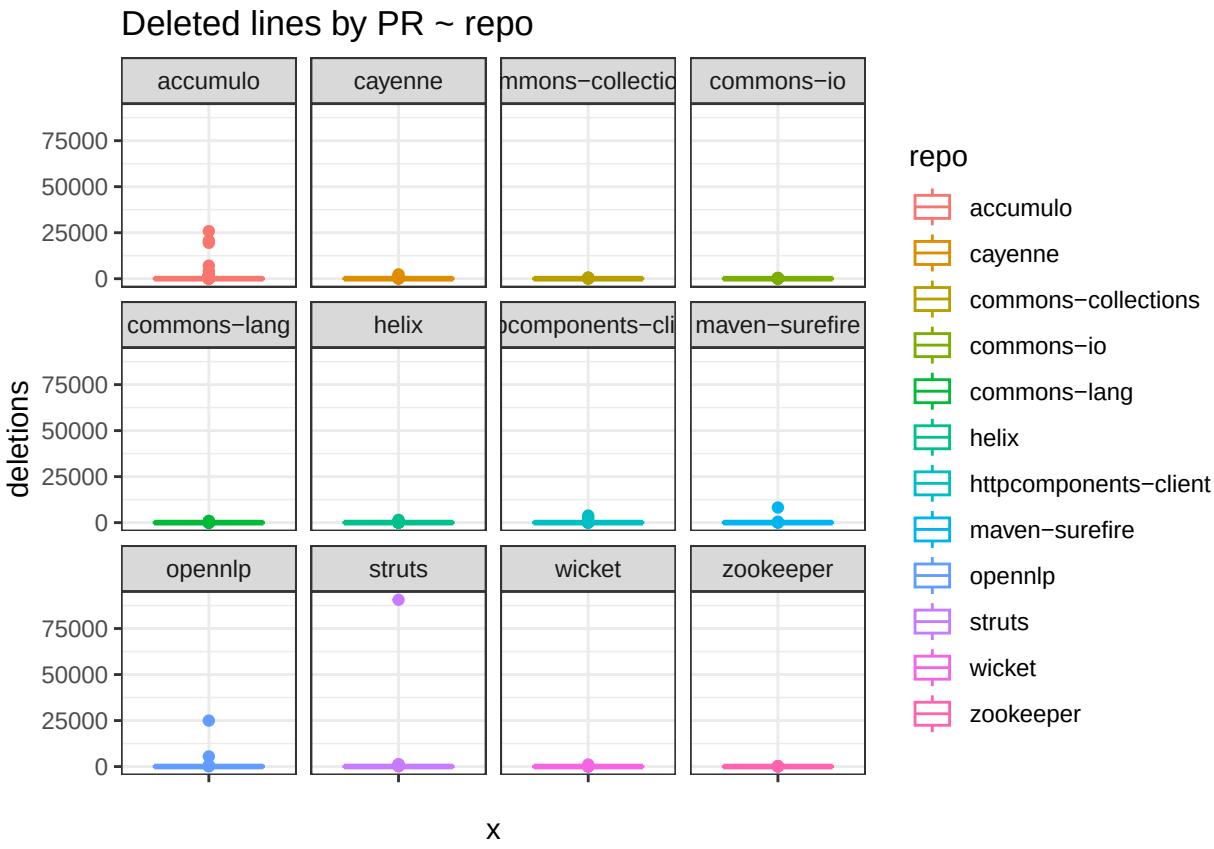
Deleted lines

```
tapply(prs_characterization$deletions, prs_characterization$repo, summary)
```

```
## $accumulo
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.00     3.00    12.00   156.00    49.75 25869.00
##
## $cayenne
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     2.0    14.0   110.7    43.5  2346.0
##
## $`commons-collections`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    5.00   79.52   34.00  557.00
##
## $`commons-io`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     1.0     6.5    32.7    19.5   423.0
##
## $`commons-lang`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    4.00   36.34   15.25  979.00
```

```
##
## $helix
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    2.00   11.00   45.97   35.00 1506.00
##
## $`httpcomponents-client`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    3.50  103.93   26.75 3850.00
##
## $`maven-surefire`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     3.0     9.0   368.4    20.0  8228.0
##
## $opennlp
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     2.0    16.5   382.1    99.5 24994.0
##
## $struts
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     6.0    28.0   743.4   112.0 90609.0
##
## $wicket
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    4.00   12.00   51.11   26.50 1097.00
##
## $zookeeper
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    4.00   34.76   36.50  307.00
```

```
prs_characterization %>%
    ggplot(aes(x = "", y = deletions, color=repo)) +
      geom_boxplot() +
      facet_wrap(~repo) +
      labs(title = "Deleted lines by PR ~ repo")
```



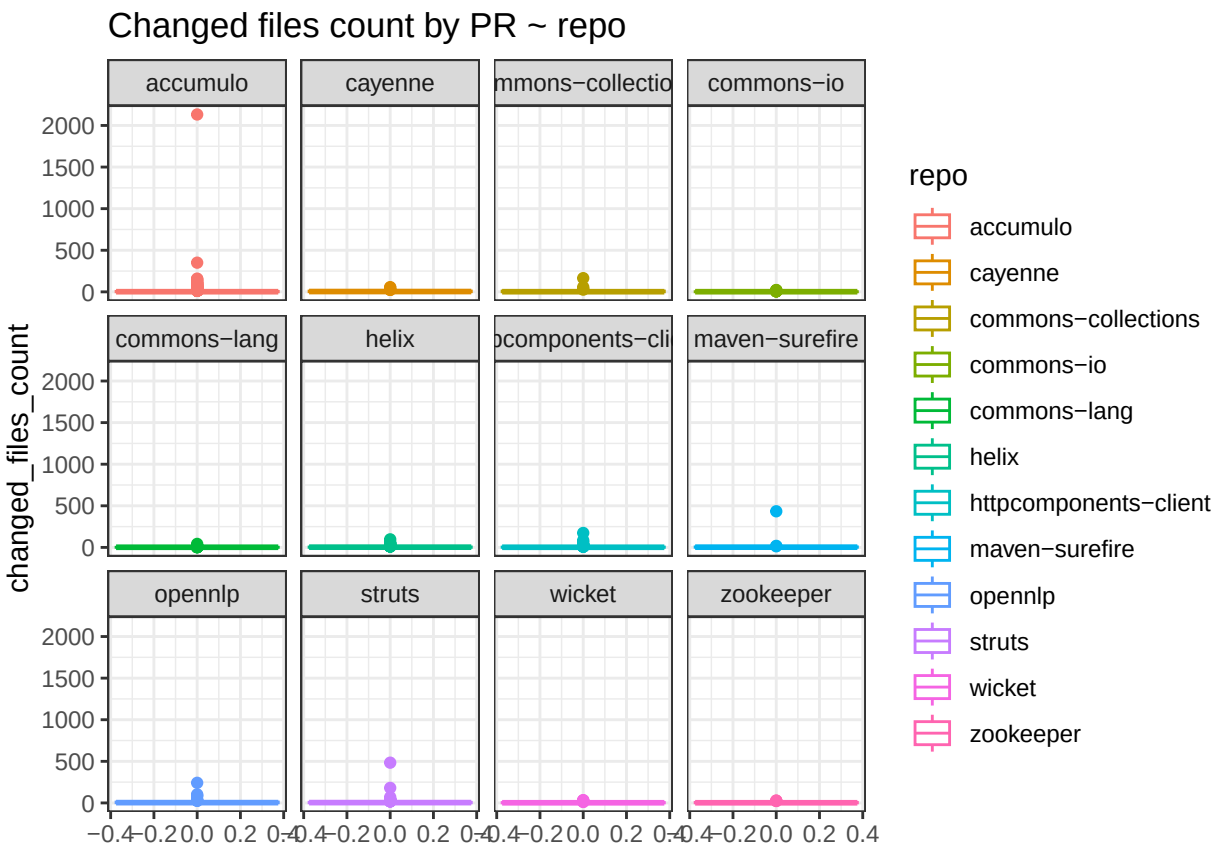
Changed files count

```
tapply(prs_characterization$changed_files_count, prs_characterization$repo, summary)
```

```
## $accumulo
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    2.00    9.43    6.00 2132.00
##
## $cayenne
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   4.000   8.127  10.000  57.000
##
## $`commons-collections`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    2.00   12.97   10.50  165.00
##
## $`commons-io`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   3.184   2.250  22.000
##
## $`commons-lang`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   3.188   2.000  42.000
```

```
##
## $helix
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   3.000   5.234   5.000  97.000
##
## $`httpcomponents-client`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   7.183   4.000 174.000
##
## $`maven-surefire`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    3.00   21.84    6.00  434.00
##
## $opennlp
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    3.50   11.98   11.00  242.00
##
## $struts
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    4.00   12.08    8.00  484.00
##
## $wicket
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   5.149   5.000  33.000
##
## $zookeeper
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   5.588   6.750  29.000
```

```r
prs_characterization %>%
    ggplot(aes(y = changed_files_count, color=repo)) +
      geom_boxplot() +
      facet_wrap(~repo) +
      labs(title = "Changed files count by PR ~ repo")
```



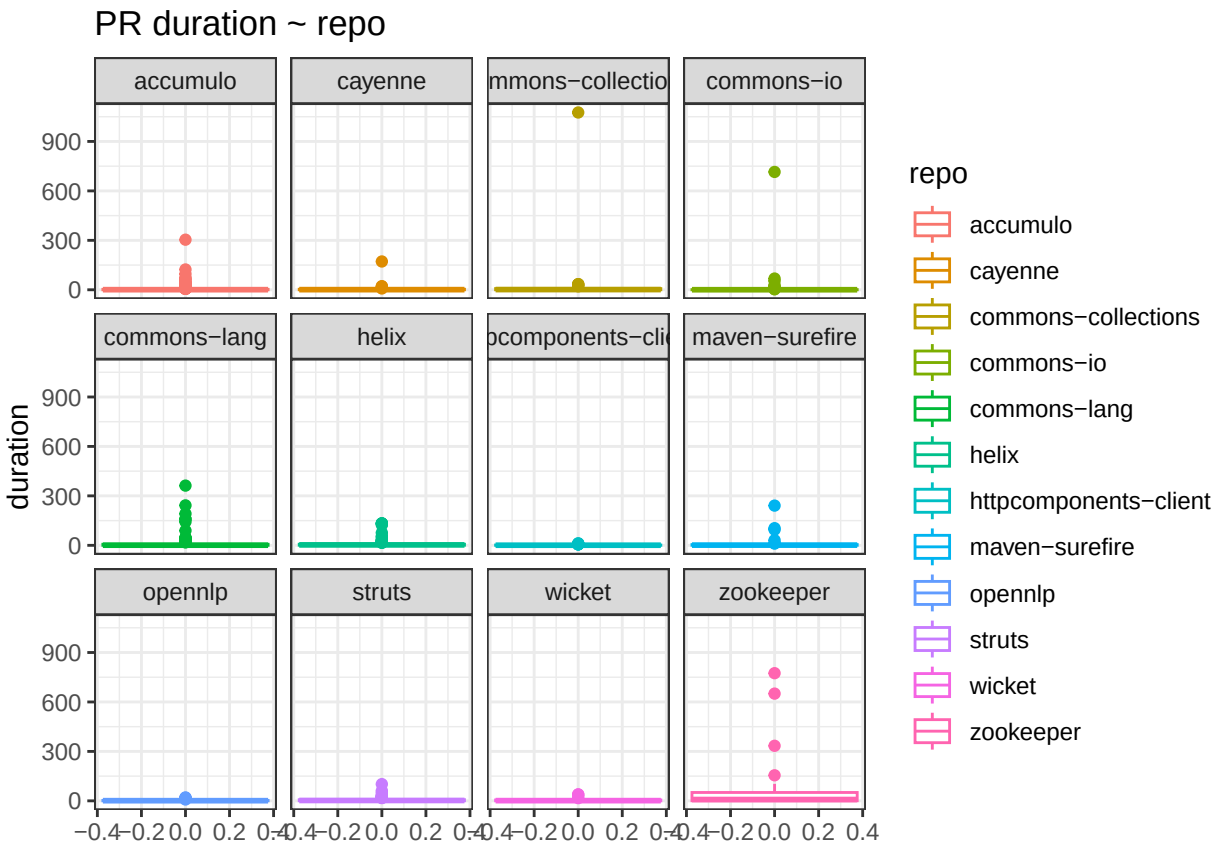
PR duration

```
tapply(prs_characterization$duration, prs_characterization$repo, summary)
```

```
## $accumulo
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00164   0.09984   0.77136   3.17734   2.80212 303.90714
##
## $cayenne
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00892   0.24649   0.93683   5.67438   3.36617 171.71755
##
## $`commons-collections`
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    0.0035    0.1510    1.4189   41.3452    7.2682 1075.6453
##
## $`commons-io`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0270   0.1300   0.3285  12.4828   1.1068 715.0823
##
## $`commons-lang`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0030   0.1234   0.7316  15.4330   6.4838 362.2128
```

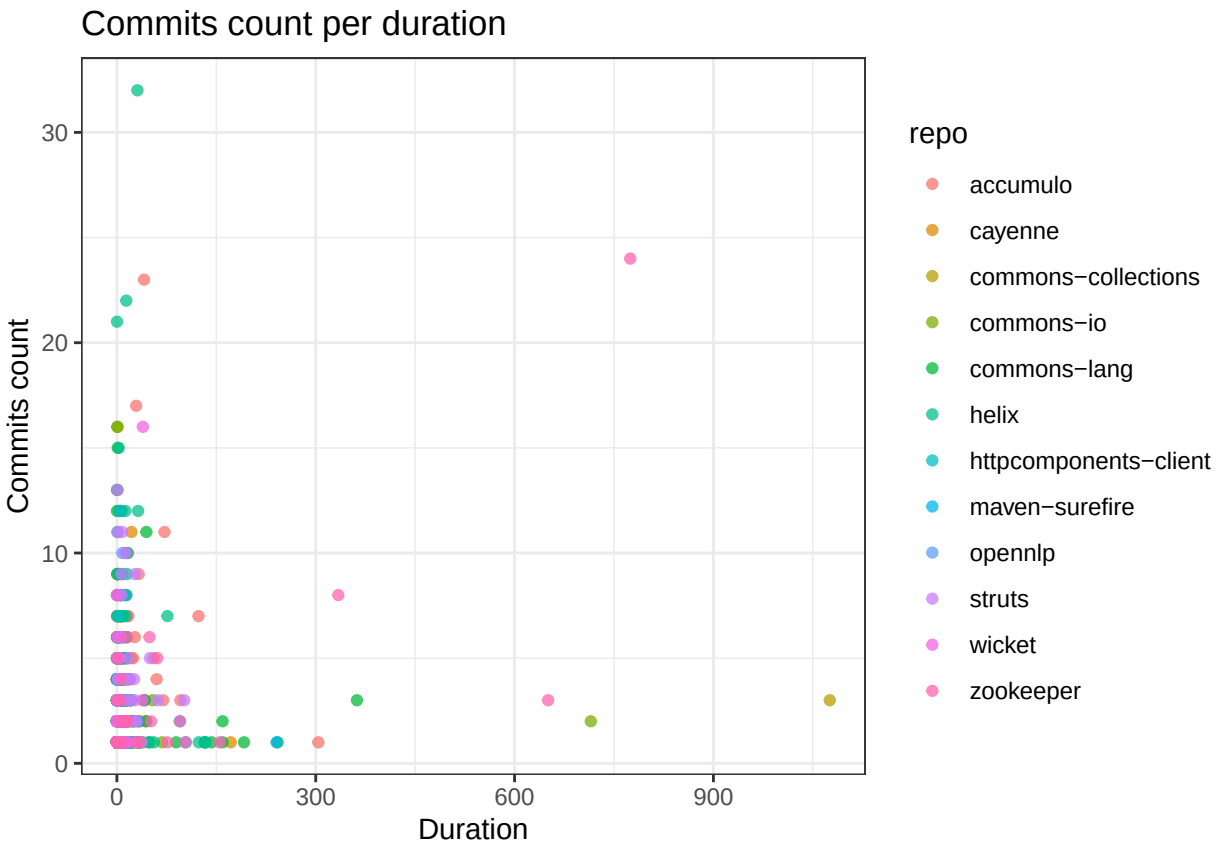
```
##
## $helix
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00017   0.79010   2.56567   7.22734   6.60545 132.76405
##
## $`httpcomponents-client`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00265  0.15797  0.58484  1.36375  1.90901 12.36778
##
## $`maven-surefire`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.1133   0.5058   1.2199  20.8595   4.4510 240.9541
##
## $opennlp
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00419  0.29868  1.04260  2.26427  2.61935 20.87879
##
## $struts
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.01024   0.61083   2.30698   6.21261   6.93494 101.57461
##
## $wicket
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00584  0.34491  1.37601  5.00295  5.38928 39.14480
##
## $zookeeper
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0442   0.8901   7.8947  73.9420  50.8732 774.5678
```

```
prs_characterization %>%
    ggplot(aes(y = duration, color=repo)) +
      geom_boxplot() +
      facet_wrap(~repo) +
      labs(title = "PR duration ~ repo")
```

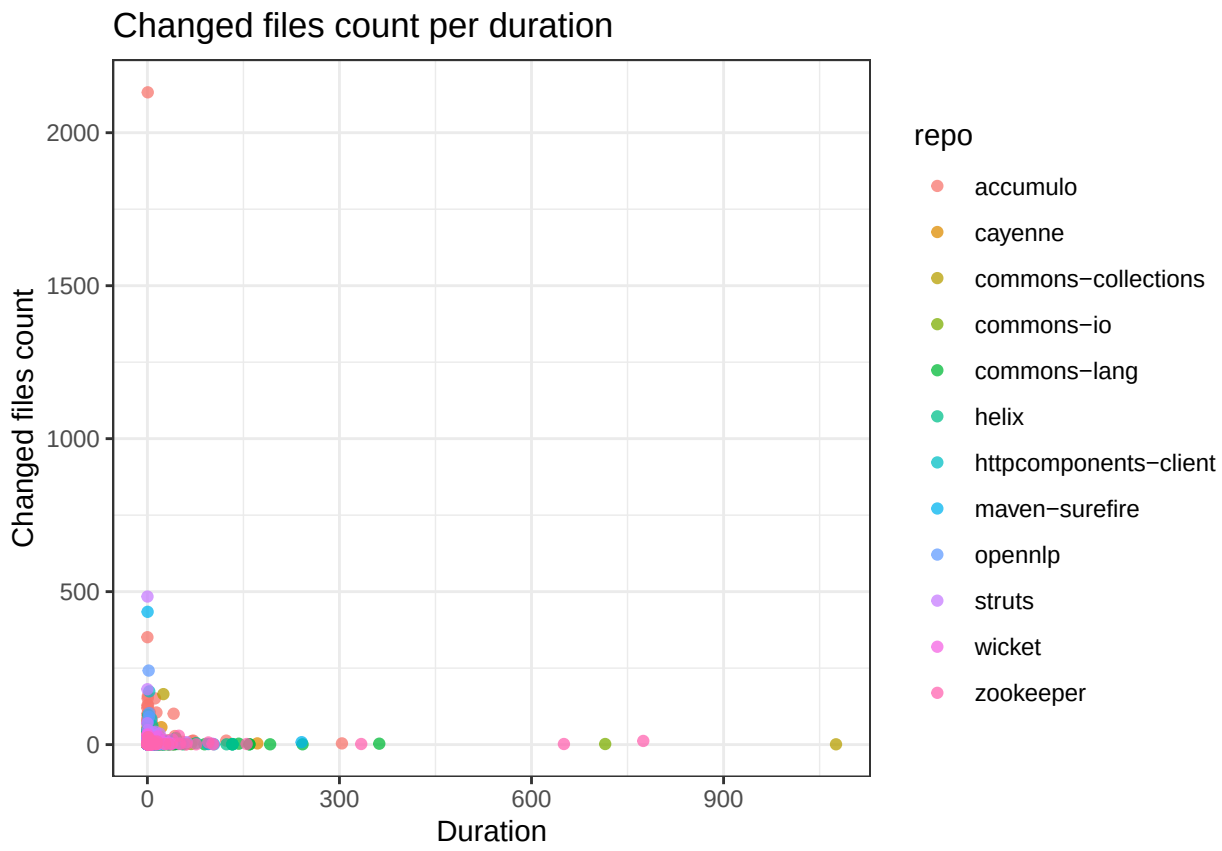


Scatters

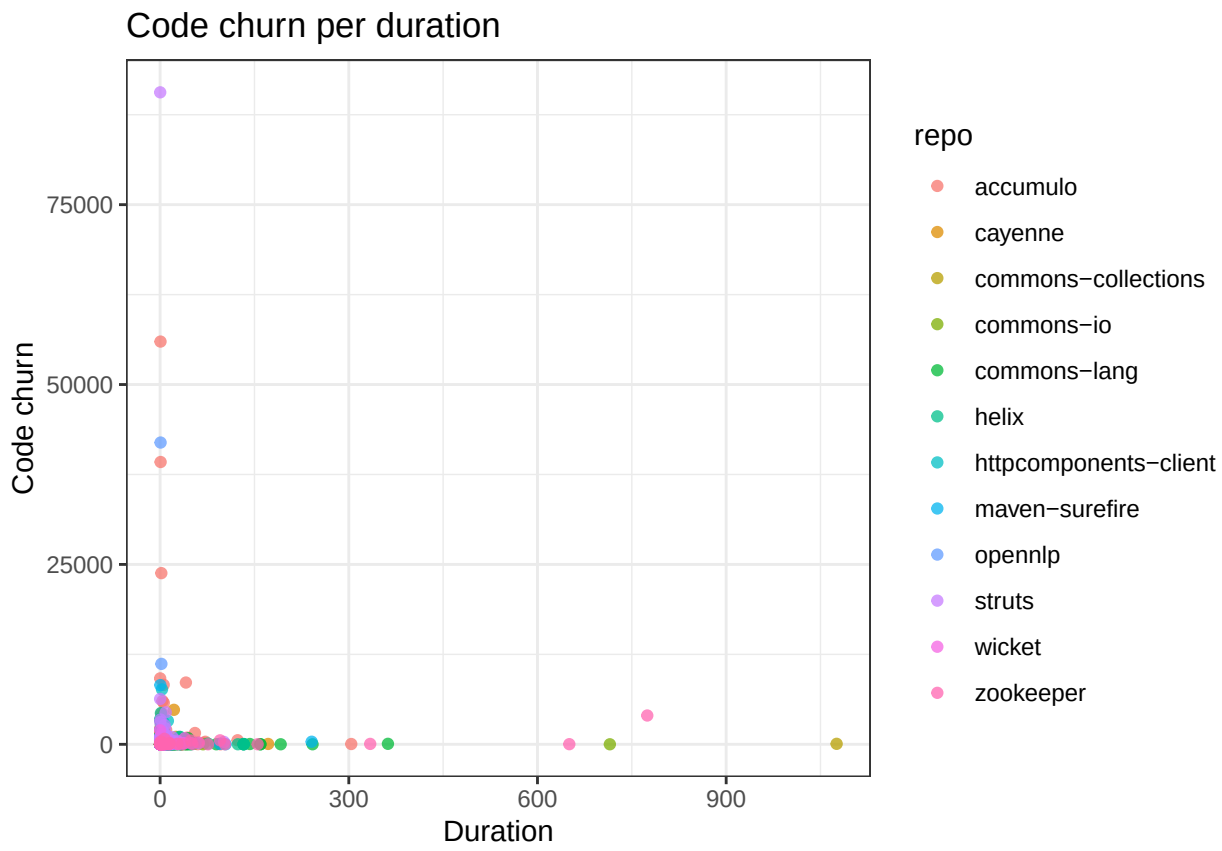
```
prs_characterization %>%
    ggplot(aes(x = duration, y = commits_count, color=repo)) +
    geom_point(alpha=0.75) +
    labs(x = "Duration", y = "Commits count") +
    ggtitle("Commits count per duration")
```



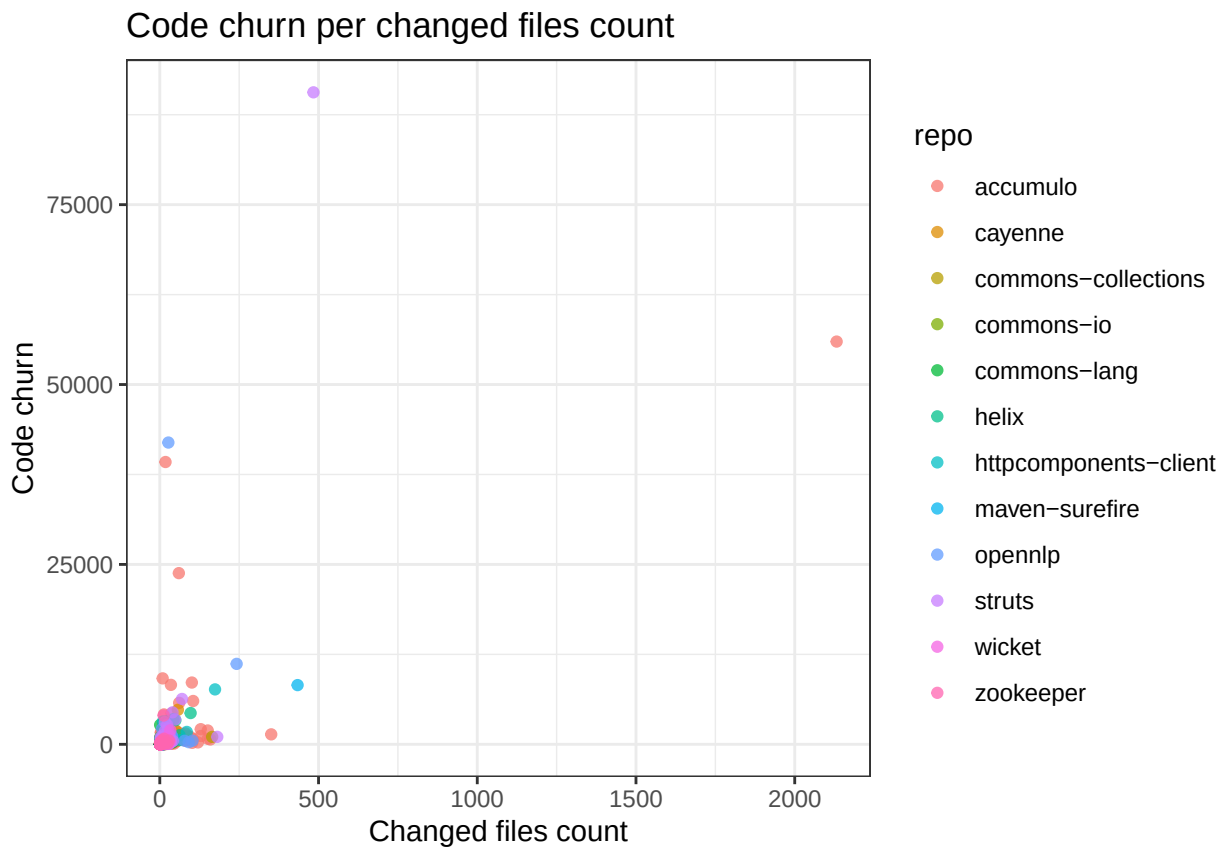

```
prs_characterization %>%
    ggplot(aes(x = duration, y = changed_files_count, color=repo)) +
    geom_point(alpha=0.75) +
    labs(x = "Duration", y = "Changed files count") +
    ggtitle("Changed files count per duration")
```




```
prs_characterization %>%
    ggplot(aes(x = duration, y = code_churn, color=repo)) +
    geom_point(alpha=0.75) +
    labs(x = "Duration", y = "Code churn") +
    ggtitle("Code churn per duration")
```

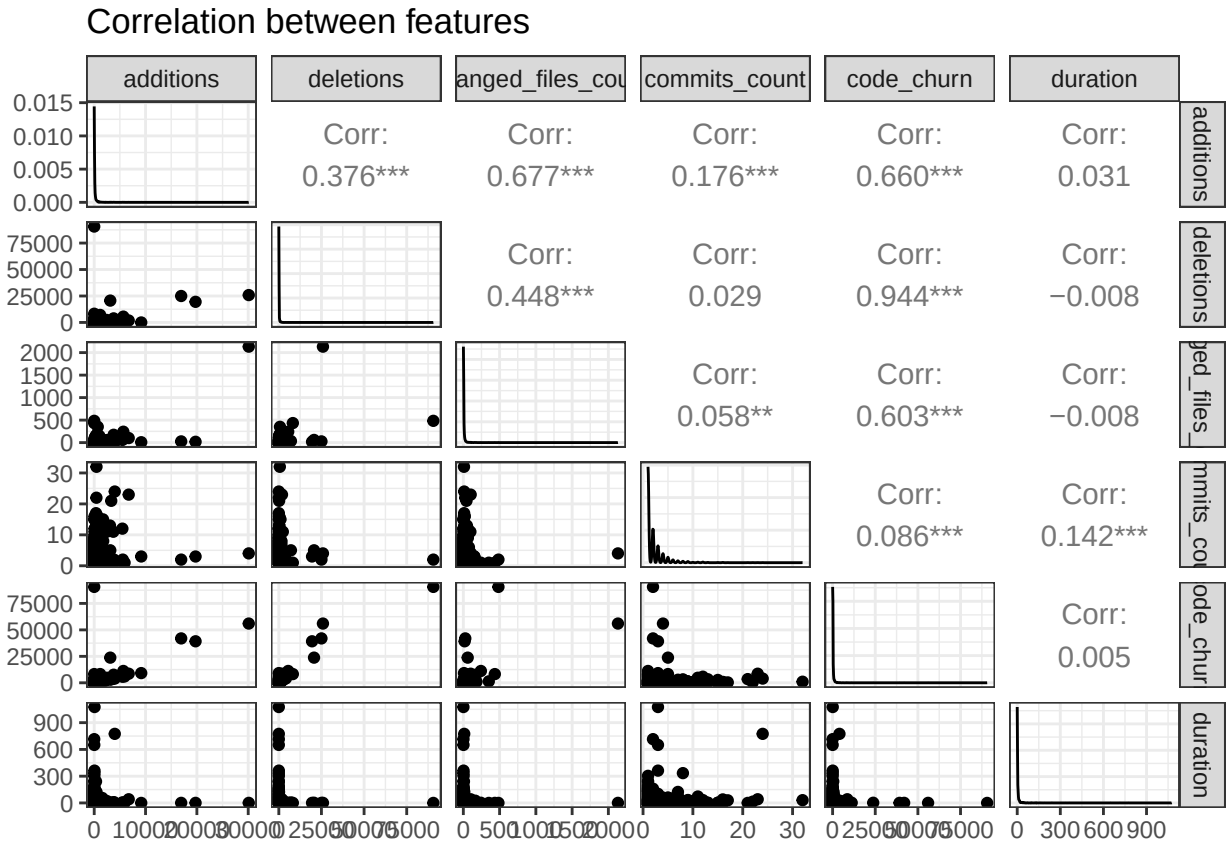



```
prs_characterization %>%
    ggplot(aes(x = changed_files_count, y = code_churn, color=repo)) +
    geom_point(alpha=0.75) +
    labs(x = "Changed files count", y = "Code churn") +
    ggtitle("Code churn per changed files count")
```



Correlations

```
prs_characterization %>%
    select(additions, deletions, changed_files_count, commits_count, code_churn, duration) %>%
    ggpairs(title="Correlation between features")
```


Correlation between features
additions
deletions
anged_files_cou
commits_count
code_churn
duration
Corr:
0.376***
Corr:
0.677***
Corr:
0.176***
Corr:
0.660***
Corr:
0.031
Corr:
0.448***
Corr:
0.029
Corr:
0.944***
Corr:
−0.008
Corr:
0.058**
Corr:
0.603***
Corr:
−0.008
Corr:
0.086***
Corr:
0.142***
Corr:
0.005
additions
deletions
ged_files_
mmits_co
ode_chur
duration