Gender Research: RQ1 Analysis

(SETUP) Laod the datasets

To answer RQ1, we use two datasets.

- all_commiters_ds: contains information about all commiters of the projects
- tf_ds: contains information about all key developers of the projects

```
all_commiters_ds <- read.csv("../datasets/all_committers_repo.csv", head=T, sep=",")
tf_ds <- read.csv("../datasets/tf.csv", head=T, sep=";")
colnames(all_commiters_ds)
```

```
## [1] "email"    "language" "location" "name"     "pais"     "gender"  
## [7] "repo"
```

```
colnames(tf_ds)
```

```
##  [1] "country"          "created_at"       "email"           
##  [4] "forks"            "full_name"        "gender"          
##  [7] "gender2"          "git_url"          "id"              
## [10] "language"         "lines"            "location"        
## [13] "login"            "num_contributors" "rate_commits"    
## [16] "repository"       "size"             "user"            
## [19] "user_num_commits" "watchers"
```

```
nrow(all_commiters_ds)
```

```
## [1] 242621
```

```
nrow(tf_ds)
```

```
## [1] 2640
```

```
summary(tf_ds$gender)
```

```
##        female   male unisex 
##    142    130   2311     57
```

```
summary(tf_ds$gender2)
```

```
## female   male 
##     88   2552
```

```
tf_ds$num_contributors <- as.numeric(gsub(",","",as.character(tf_ds$num_contributors)))
tf_ds$lines <- as.numeric(gsub(",","",as.character(tf_ds$lines)))
tf_ds$size <- as.numeric(gsub(",","",as.character(tf_ds$size)))
tf_ds$watchers <- as.numeric(gsub(",","",as.character(tf_ds$watchers)))
```

Exploratory data analysis

- Number of distinct projects

```
##   count(distinct full_name)
## 1                      1184
```

- Number of projects with at least 5 key developers

```
##                        full_name totalTF
## 1              ansible/ansible      23
## 2            gitlabhq/gitlabhq      20
## 3        elastic/elasticsearch      17
## 4        kubernetes/kubernetes      17
## 5               python/cpython      15
## 6       apache/incubator-mxnet      13
## 7        facebook/react-native      13
## 8              pytorch/pytorch      13
## 9                  rails/rails      13
## 10            jedi4ever/veewee      12
## 11                 php/php-src      11
## 12              saltstack/salt      10
## 13                 spree/spree      10
## 14               FFmpeg/FFmpeg       9
## 15              Microsoft/CNTK       9
## 16           aspnet/AspNetCore       9
## 17               dotnet/corefx       9
## 18            emberjs/ember.js       9
## 19             github/linguist       9
## 20                   golang/go       9
## 21                   moby/moby       9
## 22             Bash-it/bash-it       8
## 23         WordPress/WordPress       8
## 24                apache/kafka       8
## 25 chriskempson/tomorrow-theme       8
## 26       cockroachdb/cockroach       8
## 27              facebook/folly       8
## 28        geekcomputers/Python       8
## 29         mesosphere/marathon       8
## 30           puppetlabs/puppet       8
```

```
## [1] 88
```

```
##   count(distinct full_name)
## 1                  7.432432
```

- Characteristics of the projects

```
projects_ds <- sqldf("select  full_name, lines, size, num_contributors, forks, watchers, count(distinct
                      from tf_ds
                      group by full_name, lines, size, num_contributors, forks, watchers")

summary(projects_ds[,c("lines", "num_contributors", "size", "forks", "watchers")])
```

```
##      lines          num_contributors      size              forks
##  Min.   :      0   Min.   :   0.0   Min.   :      9   Min.   :    8
##  1st Qu.:   5183   1st Qu.:  33.0   1st Qu.:   2648   1st Qu.:  531
##  Median :  23690   Median :  82.0   Median :  10852   Median : 1077
##  Mean   : 174414   Mean   : 199.9   Mean   :  78591   Mean   : 2337
##  3rd Qu.: 104933   3rd Qu.: 187.2   3rd Qu.:  45630   3rd Qu.: 2504
##  Max.   :9442645   Max.   :8413.0   Max.   :8299557   Max.   :64712
##                    NA's   :8
##     watchers
##  Min.   :  1097
##  1st Qu.:  5315
##  Median :  8003
##  Mean   : 12206
##  3rd Qu.: 13472
##  Max.   :300666
##
```

- It is better to remove smaller projects

```
tf_ds <- filter(tf_ds, lines >= 5183, tf_ds$num_contributors >= 33)

sqldf("select count(distinct full_name) from tf_ds")
```

```
##   count(distinct full_name)
## 1                       737
```

```
projects_ds <- sqldf("select  full_name, lines, size, num_contributors, forks, watchers, count(distinct
                      from tf_ds
                      group by full_name, lines, size, num_contributors, forks, watchers")

pds_summary <- as.data.frame(sapply(projects_ds[,c("lines", "num_contributors", "size", "forks", "watche

print(xtable(t(pds_summary)), type="latex")
```

```
## % latex table generated in R 3.6.1 by xtable 1.8-4 package
## % Tue Jan  7 08:27:09 2020
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrrrr}
##   \hline
##  & Min. & 1st Qu. & Median & Mean & 3rd Qu. & Max. \\
##   \hline
## lines & 5191.00 & 19523.00 & 57013.00 & 259367.63 & 195265.00 & 9442645.00 \\
##   num\_contributors & 33.00 & 80.00 & 145.00 & 292.77 & 297.00 & 8413.00 \\
##   size & 368.00 & 6625.00 & 20503.00 & 94498.98 & 78353.00 & 3950679.00 \\
##   forks & 54.00 & 774.00 & 1481.00 & 2949.94 & 3171.00 & 64712.00 \\
##   watchers & 1145.00 & 5882.00 & 9039.00 & 14284.96 & 16418.00 & 300666.00 \\
##    \hline
## \end{tabular}
## \end{table}
```

- Correlation: number of developers and number of TF developers
- Correlation: lines of code and number of TF developers

```
t3 <- sqldf("select full_name, num_contributors, lines, count(*) ignore
             from tf_ds
             group by full_name, num_contributors, lines
             order by 3 desc")

nrow(t3)
```

```
## [1] 737
```

```
head(t3)
```

```
##                full_name num_contributors   lines ignore
## 1      apple/turicreate               51 9442645      2
## 2        dotnet/coreclr              556 9290723      2
## 3           nodejs/node             2424 5421767      7
## 4         mongodb/mongo              365 4355800      7
## 5         dotnet/roslyn              349 4103895      3
## 6 kubernetes/kubernetes             2092 3566770     19
```

```
t4 <- merge(t2, t3)
cor.test(as.numeric(t4$num_contributors), t4$totalTF, method="spearman")
```

```
## Warning in cor.test.default(as.numeric(t4$num_contributors), t4$totalTF, :
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  as.numeric(t4$num_contributors) and t4$totalTF
## S = 61991, p-value = 7.052e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4151481
```

```
cor.test(as.numeric(t4$lines), t4$totalTF, method="spearman")
```

```
## Warning in cor.test.default(as.numeric(t4$lines), t4$totalTF, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  as.numeric(t4$lines) and t4$totalTF
## S = 80164, p-value = 0.02375
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.243699
```

- Total number of key developers

```
distinct_tfs <- sqldf("select full_name, count(distinct login) total
                       from tf_ds
                       group by full_name
                       order by 2 desc")

nrow(distinct_tfs)
```

```
## [1] 737
```

```
head(distinct_tfs, 50)
```

```
##                                  full_name total
## 1                          ansible/ansible    23
## 2                        gitlabhq/gitlabhq    20
## 3                    elastic/elasticsearch    17
## 4                    kubernetes/kubernetes    17
## 5                           python/cpython    15
## 6                   apache/incubator-mxnet    13
## 7                    facebook/react-native    13
## 8                          pytorch/pytorch    13
## 9                              rails/rails    13
## 10                        jedi4ever/veewee    12
## 11                             php/php-src    11
## 12                             spree/spree    10
## 13                           FFmpeg/FFmpeg     9
## 14                          Microsoft/CNTK     9
## 15                       aspnet/AspNetCore     9
## 16                           dotnet/corefx     9
## 17                        emberjs/ember.js     9
## 18                         github/linguist     9
## 19                               golang/go     9
## 20                               moby/moby     9
## 21                          saltstack/salt     9
## 22                         Bash-it/bash-it     8
## 23                     WordPress/WordPress     8
## 24                            apache/kafka     8
## 25             chriskempson/tomorrow-theme     8
## 26                   cockroachdb/cockroach     8
## 27                          facebook/folly     8
## 28                     mesosphere/marathon     8
## 29                       puppetlabs/puppet     8
## 30                  robbyrussell/oh-my-zsh     8
## 31                         twitter/finagle     8
## 32                RaRe-Technologies/gensim     7
## 33                        fzaninotto/Faker     7
## 34                    geekcomputers/Python     7
## 35                     influxdata/influxdb     7
## 36                             nodejs/node     7
## 37                           opencv/opencv     7
## 38                         palantir/tslint     7
## 39                      rubocop-hq/rubocop     7
## 40                        twitter/scalding     7
## 41 windows-toolkit/WindowsCommunityToolkit     7
## 42                        Microsoft/vscode     6
## 43                     PaddlePaddle/Paddle     6
## 44                           Theano/Theano     6
## 45                               akka/akka     6
## 46                         angular/angular     6
## 47                           apache/thrift     6
## 48                             apple/swift     6
## 49                               chef/chef     6
## 50                              dotnet/cli     6
```

```
nrow(distinct_tfs[distinct_tfs$total>1, ])
```

```
## [1] 397
```

(RQ1) How common are women key developers in OSS projects?

We answer this research question using an exploratory data analysis. We first report the characteristics of the projects (see table bellow).

```
ds_summary <- sqldf("select language, full_name, lines, num_contributors, forks, watchers, count(distin
                    from tf_ds
                    group by language, full_name, lines, num_contributors, forks, watchers")
```

```
nrow(ds_summary)
```

[1] 737

```
ds_summary_language <- sqldf("select language as 'Prog. Language',
                                     avg(lines) as 'Average number of lines of code',
                                     avg(num_contributors) as 'Average number of contributors',
                                     avg(forks) as 'Average number of forks' ,
                                     avg(watchers) as 'Average number of watchers',
                                     avg(num_tf) as 'Average number of key developers'
                              from ds_summary
                              group by language
                              order by 1")

print(xtable(ds_summary_language), type="html")
```

Prog. Language

Average number of lines of code

Average number of contributors

Average number of forks

Average number of watchers

Average number of key developers

1

C

241308.02

232.54

2350.66

9825.80

2.00

2

C#

453339.08

161.11

1656.71

5909.29

2.24

3

C++

633316.91

253.04

3323.22

13645.39

3.01

4

CSS

120447.17

108.91

2313.78

13286.96

1.57

5

Go

489145.35

282.82

2520.96

15314.19

2.78

6

Java

275866.11

217.61

5640.75

16677.05

2.43

7

JavaScript

202729.63

505.96

7611.28

42959.09

2.54

8

Objective-C

165398.92

102.00

1650.27

8826.46

1.62

9

PHP

108836.46

289.46

2007.04

9518.15

1.87

10

Python

162283.81

493.00

4397.00

18717.02

3.29

11

Ruby

99639.68

601.07

2261.10

10418.13

3.10

12

Scala

100119.61

157.12

978.74

3646.35

2.42

13

Shell

66032.21

225.71

1843.12

12075.75

2.88

14

Swift

33857.72

114.69

1347.94

10810.28

1.69

15

TypeScript

245799.41

321.70

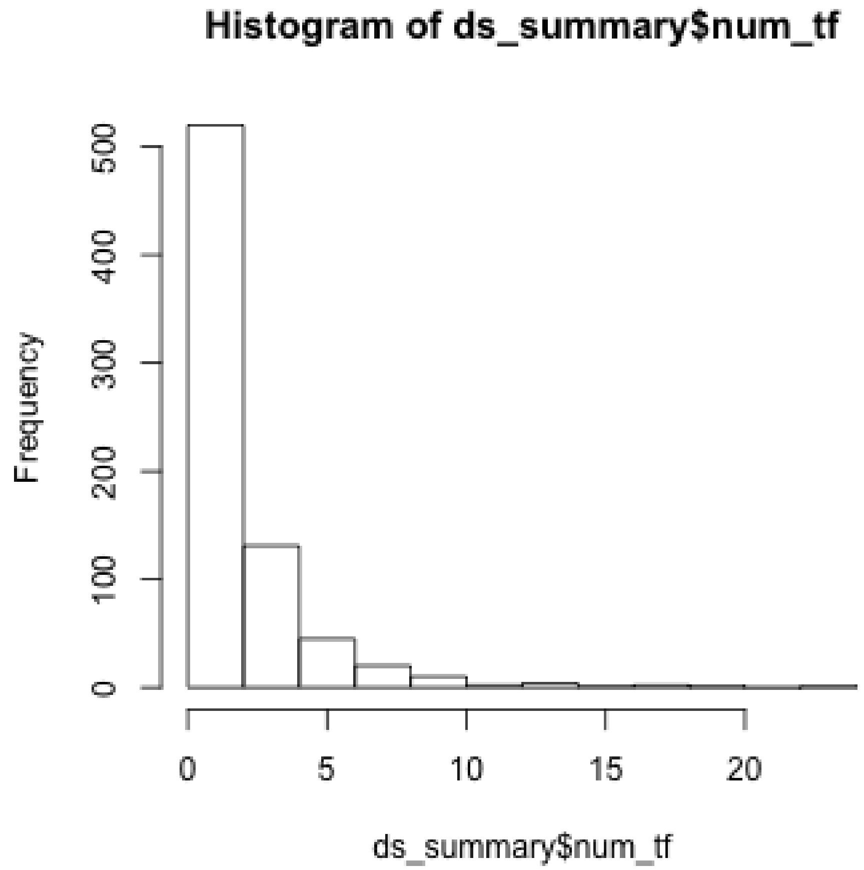
2228.66

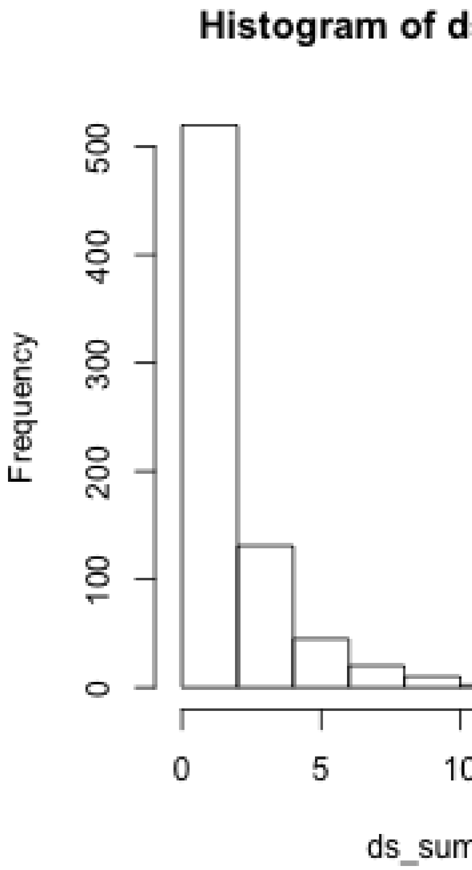
12903.53

2.00

Next we present a histogram with the number of key developers per project.

```r
plot(hist(ds_summary$num_tf))
```





Let's try to understand the women participation on OSS projects.

```
all_commiters_ds["gender_final"] <-
  ifelse(trim(as.character(all_commiters_ds$gender)) == "", as.character("unisex"), as.character(all_co

tf_ds["gender_final"] <-
  ifelse(as.character(tf_ds$gender) != as.character(tf_ds$gender2), "unissex",  as.character(tf_ds$gend

# note. we have to remove duplicated data

t_all_commiters <- sqldf("select name, gender_final, count(*) summaryCommiters
                          from all_commiters_ds
                          group by name, gender_final
                          order by 3 desc")


t_tf <- sqldf("select login, user, gender_final, count(*) summaryTF
                    from tf_ds
                    group by login, user, gender_final
                    order by 3 desc")
```

```r
sqldf("select gender_final, count(*) total from all_commiters_ds group by gender_final")
```

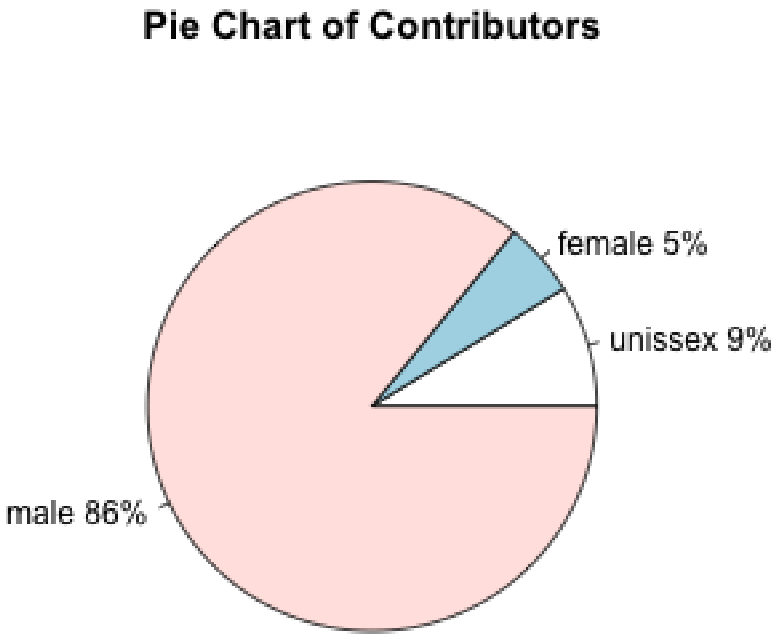
```
##   gender_final  total
## 1       female  12987
## 2         male 208384
## 3       unisex  21250
```

```r
sqldf("select gender_final, count(*) total from tf_ds group by gender_final")
```

```
##   gender_final total
## 1       female    45
## 2         male  1762
## 3      unissex   195
```

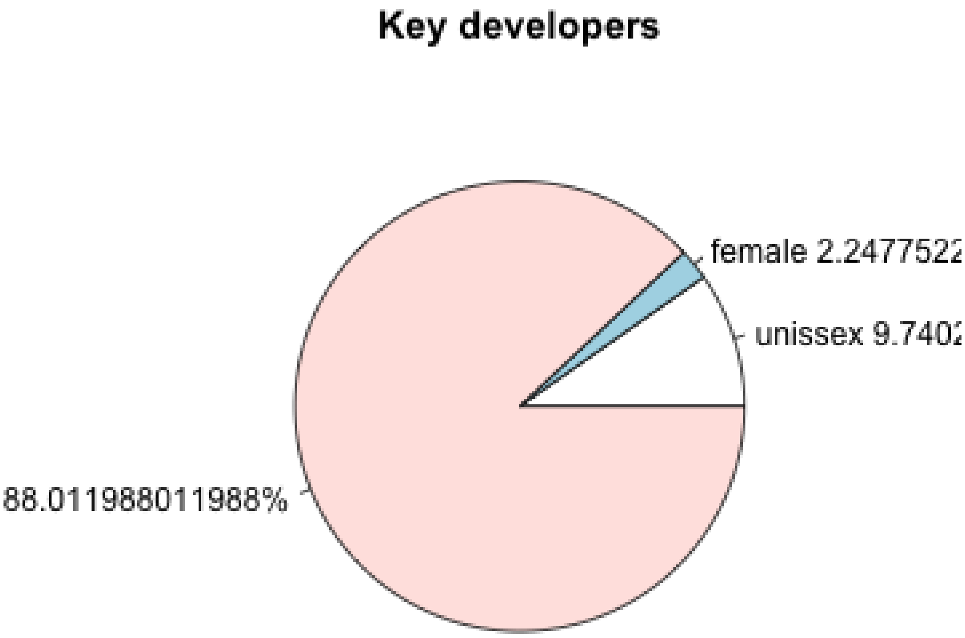
```r
slices <- c(21250, 12987, 208384)
lbls <- c("unissex", "female", "male")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels

pie(slices, labels=lbls,explode=0.1,main="Pie Chart of Contributors")
```

```
## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in title(main = main, ...): "explode" is not a graphical parameter
```



```
slices <- c(195, 45, 1762)
lbls <- c("unissex", "female", "male")
pct <- slices/sum(slices)*100
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels

pie(slices, labels=lbls,explode=0.1,main="Key developers")
```

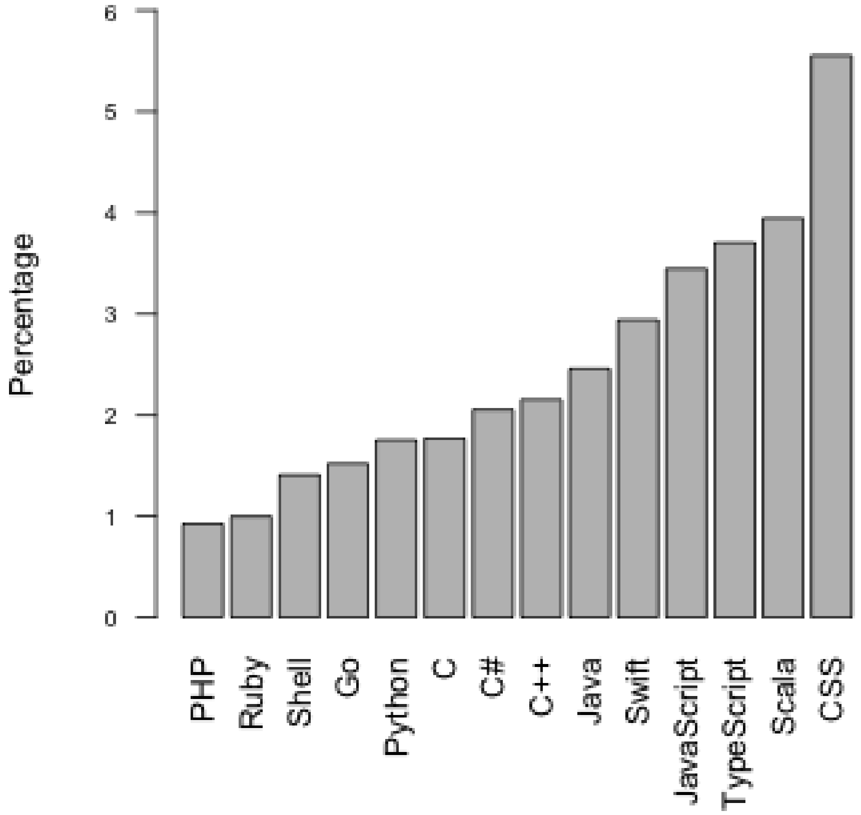
```
## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in text.default(1.1 * P$x, 1.1 * P$y, labels[i], xpd = TRUE, adj =
## ifelse(P$x < : "explode" is not a graphical parameter

## Warning in title(main = main, ...): "explode" is not a graphical parameter
```



That is, the percentage of women key developers is smaller then the percentage of women, when considering all contributors of the projects.

Now, lets group the key developers by language and gender.

```
tf_language_female_only <- sqldf("select language, count(*) total_female
                             from tf_ds
                             where gender_final = 'female'
                             group by language");

tf_language <- sqldf("select language, count(*) total
                             from tf_ds
                             group by language")

res <- merge(tf_language, tf_language_female_only)
res["percent"] <- res$total_female * 100 / res$total

res <- res[order(res$percent), ]

barplot(res$percent, names.arg = res$language, las=2, ylab="Percentage", cex.axis = 0.7, ylim=range(pre
```



```
# dotplot(reorder(res$language, res$percent)~res$language)

# dotplot(res$percent~res$language, las=2)

# dotchart2(res$percent, labels=res$language, las=2, horizontal=F, sort.=T)
```

How many projects have at least one women key developer?

```
sqldf("select count(distinct full_name) total_projects
                              from tf_ds
                              where gender_final = 'female'");
```

```
##   total_projects
## 1             37
```

```
sqldf("select count(distinct full_name) total_projects
                              from tf_ds");
```

```
##   total_projects
## 1            737
```

Table 1.

```
head(ds_summary, results='asis')
```

```
##   language                                             full_name   lines
## 1        C                                    FFmpeg/FFmpeg 1179881
## 2        C                            MarlinFirmware/Marlin  117328
## 3        C SamyPesse/How-to-Make-a-Computer-Operating-System   23443
## 4        C                                  alibaba/tengine  279209
## 5        C                                allinurl/goaccess   26906
## 6        C                            beanstalkd/beanstalkd    7154
##   num_contributors forks watchers num_tf
## 1             1019  5424    14444      9
## 2              413  6910     5211      1
## 3               35  3127    17933      1
## 4               67  1985     8229      1
## 5               75   624     8676      1
## 6               45   734     5005      1
```

```
tab1<- sqldf("select language, count(distinct full_name) projects, sum(num_contributors) contributors
       from ds_summary
       group by language")

tab1_kd <- sqldf("select language, count(*) as total_kds
             from tf_ds
             group by language")

tab1_men <- sqldf("select language, count(*) as total_male
             from tf_ds
             where gender_final = 'male'
             group by language")


tab1_female <- sqldf("select language, count(*) as total_female
                 from tf_ds
                 where gender_final = 'female'
                 group by language")


tab1_unknown <- sqldf("select language, count(*) as total_unknown
                 from tf_ds
                 where gender_final = 'unissex'
                 group by language")

tab1_total_projects_with_women <- sqldf("select language, count(distinct full_name) as total_projects_wi
                 from tf_ds
                 where gender_final = 'female'
                 group by language")

tab1 <- merge(tab1, tab1_kd)

tab1 <- merge(tab1, tab1_men)


tab1 <- merge(tab1, tab1_female)

tab1 <- merge(tab1, tab1_unknown)

tab1 <- merge(tab1, tab1_total_projects_with_women)
```

```r
tab1["percentage"] <- tab1$total_projects_with_women * 100 / tab1$projects

print(xtable(tab1[,c("language", "projects", "contributors", "total_kds", "total_male", "total_female",
```

```
## % latex table generated in R 3.6.1 by xtable 1.8-4 package
## % Tue Jan  7 08:27:14 2020
## \begin{table}[ht]
## \centering
## \begin{tabular}{rlrrrrrrrr}
##   \hline
##  & language & projects & contributors & total\_kds & total\_male & total\_female & total\_unknown & t
##   \hline
## 1 & C &  50 & 11627.00 & 113 &  99 &   2 &  12 &   2 & 4.00 \\
##   2 & C\# &  63 & 10150.00 & 146 & 129 &   3 &  14 &   3 & 4.76 \\
##   3 & C++ &  67 & 16954.00 & 232 & 191 &   5 &  36 &   3 & 4.48 \\
##   4 & CSS &  23 & 2505.00 &  36 &  30 &   2 &   4 &   2 & 8.70 \\
##   5 & Go &  68 & 19232.00 & 197 & 169 &   3 &  25 &   3 & 4.41 \\
##   6 & Java &  44 & 9575.00 & 122 & 109 &   3 &  10 &   2 & 4.55 \\
##   7 & JavaScript &  67 & 33899.00 & 203 & 172 &   7 &  24 &   5 & 7.46 \\
##   8 & PHP &  46 & 13315.00 & 108 &  98 &   1 &   9 &   1 & 2.17 \\
##   9 & Python &  42 & 20706.00 & 171 & 155 &   3 &  13 &   2 & 4.76 \\
##   10 & Ruby &  60 & 36064.00 & 200 & 183 &   2 &  15 &   2 & 3.33 \\
##   11 & Scala &  57 & 8956.00 & 152 & 137 &   6 &   9 &   5 & 8.77 \\
##   12 & Shell &  24 & 5417.00 &  71 &  63 &   1 &   7 &   1 & 4.17 \\
##   13 & Swift &  36 & 4129.00 &  68 &  59 &   2 &   7 &   2 & 5.56 \\
##   14 & TypeScript &  64 & 20589.00 & 135 & 123 &   5 &   7 &   4 & 6.25 \\
##    \hline
## \end{tabular}
## \end{table}
```

```r
sum(tab1$projects)
```

```
## [1] 711
```

```r
sum(tab1$contributors)
```

```
## [1] 213118
```

```r
sum(tab1$total_kds)
```

```
## [1] 1954
```

```r
sum(tab1$total_male)
```

```
## [1] 1717
```

```r
sum(tab1$total_female)
```

```
## [1] 45
```

```r
sum(tab1$total_unknown)
```

```
## [1] 192
```

```r
sum(tab1$total_projects_with_women)
```

```
## [1] 37
```

```r
mean(tab1$percentage)
```

```
## [1] 5.240598
```

```r
sd(tab1$percentage)
```

```
## [1] 1.929702
```