dataset B summary

Peder M. Isager

11/1/2019

```r
data.B <- readRDS("../processed_data/dataset_B.rds")

## summary of key variables

key.vars <- c("PY", "study_number", "sample_size", "coder", "resolver", "excluded", "TC", "RV")

summary(data.B[, key.vars])
```
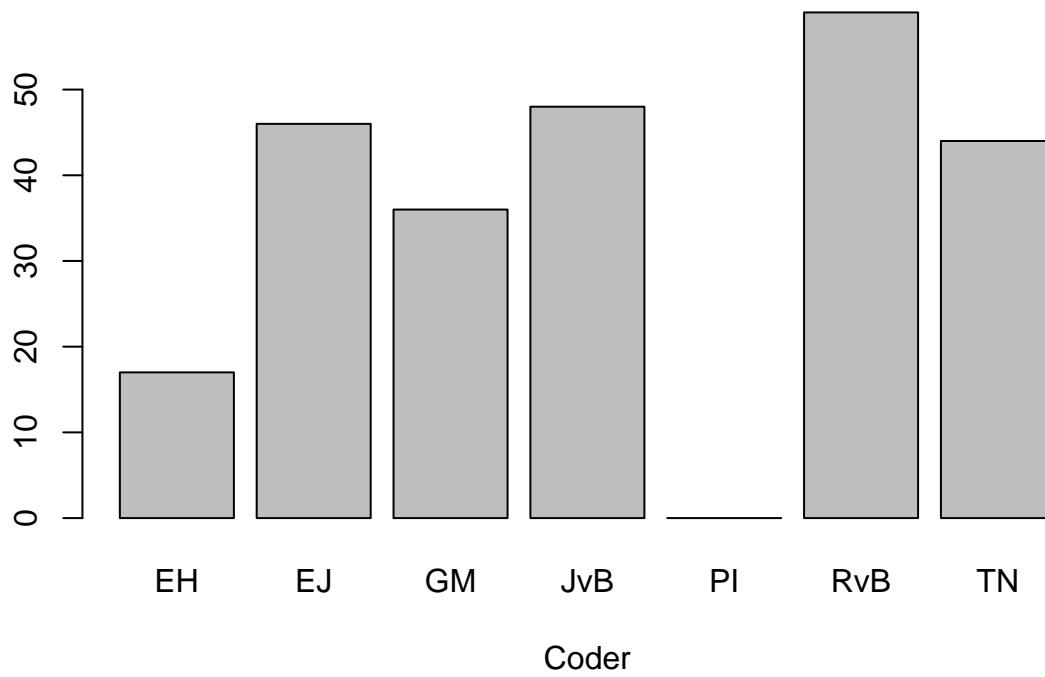
```
##       PY        study_number  sample_size       coder     resolver
##  Min.   :2009   1   :223     Min.   :   6.00   EH :17   AV  : 19
##  1st Qu.:2013   2   : 19     1st Qu.:  18.00   EJ :46   PI  : 26
##  Median :2015   3   :  4     Median :  24.00   GM :36   NA's:205
##  Mean   :2014   4   :  2     Mean   :  43.09   JvB:48
##  3rd Qu.:2016   5   :  1     3rd Qu.:  39.00   PI : 0
##  Max.   :2018   6   :  0     Max.   :1235.00   RvB:59
##  NA's   :1      NA's:  1                       TN :44
##     excluded        TC              RV
##  Min.   :0    Min.   :  0.00   Min.   :0.00000
##  1st Qu.:0    1st Qu.:  2.00   1st Qu.:0.02778
##  Median :0    Median :  9.00   Median :0.08772
##  Mean   :0    Mean   : 18.19   Mean   :0.17381
##  3rd Qu.:0    3rd Qu.: 25.75   3rd Qu.:0.20106
##  Max.   :0    Max.   :116.00   Max.   :4.22222
##                                NA's   :1
```
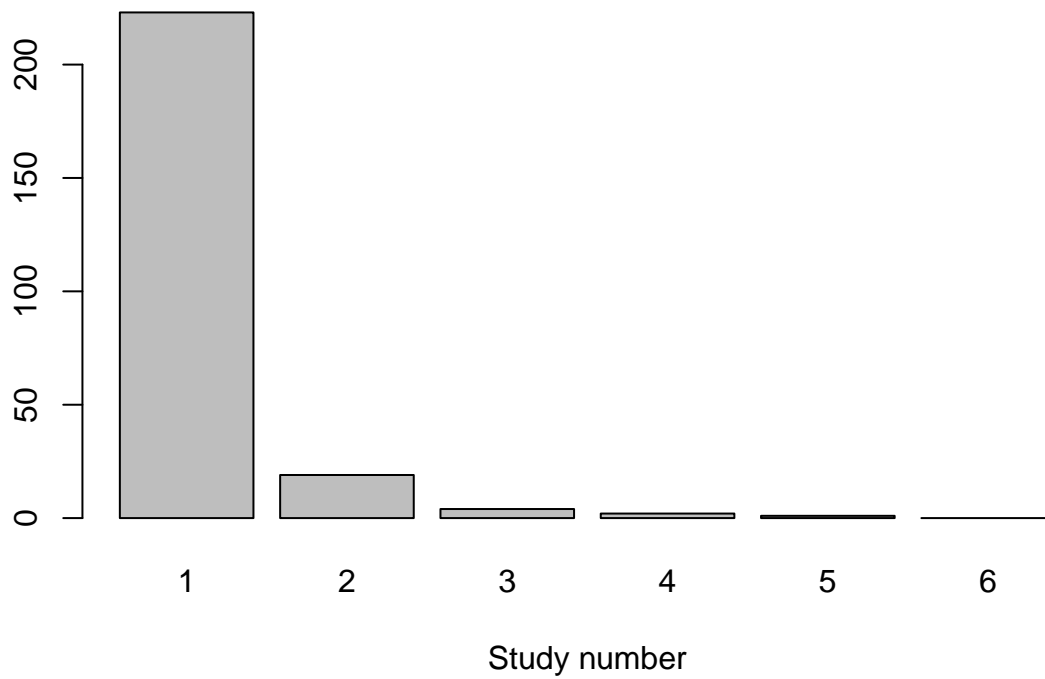
```r
## visualization of key variables

coder.freq <- table(data.B$coder)
barplot(coder.freq, xlab = "Coder")  # Plot frequency of coders
```
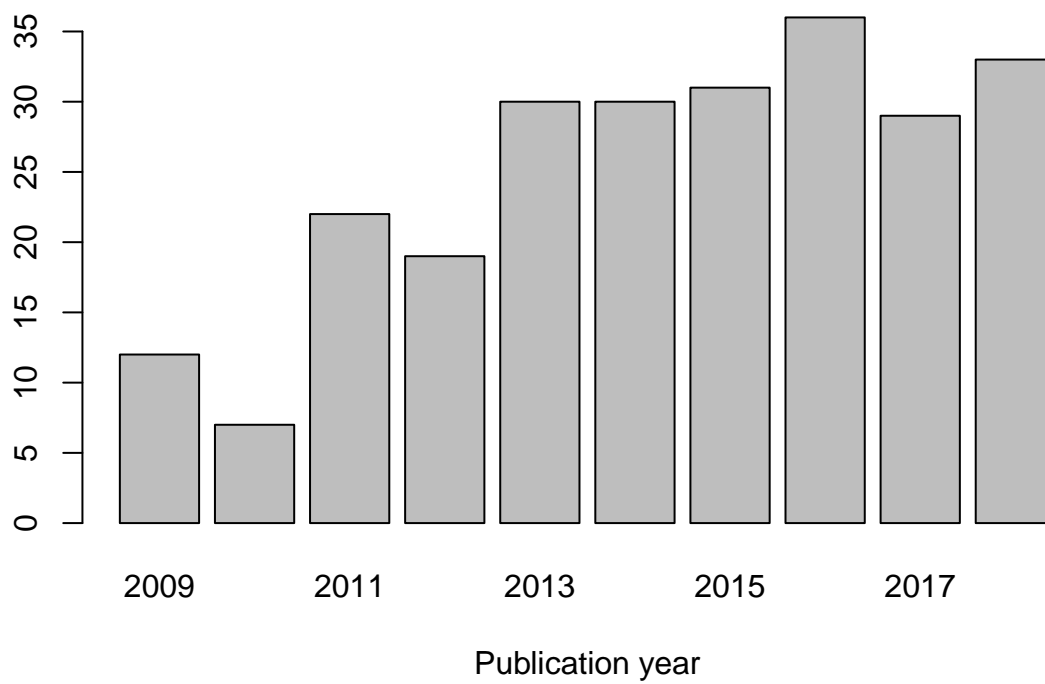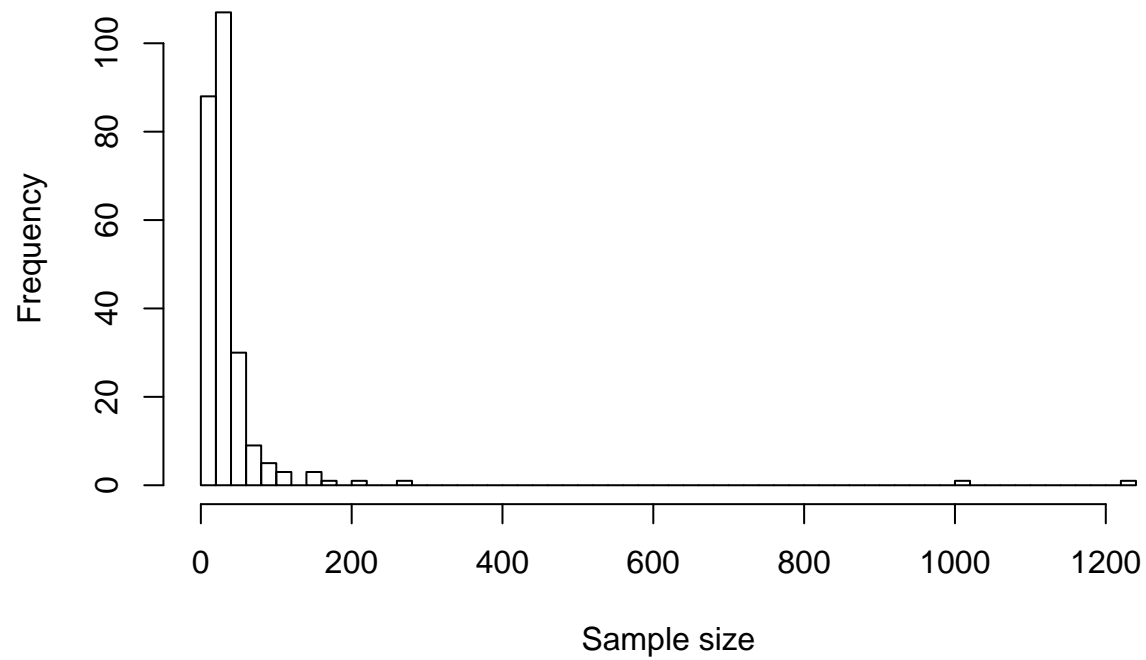
```r
studyn.freq <- table(data.B$study_number)
barplot(studyn.freq, xlab = "Study number")   # Plot frequency of study numbers (first vs. second vs. fo
```
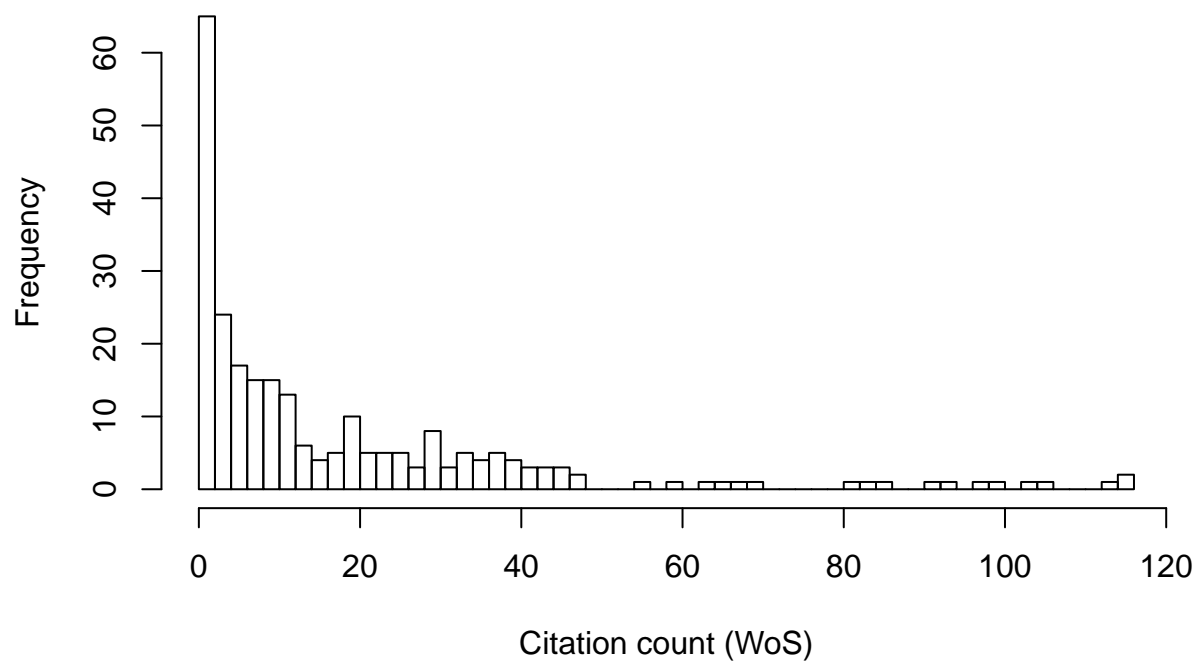
```
pubyear.freq <- table(data.B$PY)
barplot(pubyear.freq, xlab = "Publication year")  # Plot frequency of publication years
```

Publication year

```r
hist(data.B$sample_size, breaks = 50, xlab = "Sample size", main = "")  # Plot sample size distribution
```

```r
hist(data.B$TC, breaks = 50, xlab = "Citation count (WoS)", main = "")   # Plot citation count distribut
```

```r
hist(data.B$RV, breaks = 50, xlab = "Replication value (C/Y)/(N-3)", main = "")   # Plot replication val
```

Replication value (C/Y)/(N−3)