code.R

rstudio-user

2021-11-12

```
#Package installation
install.packages("dplyr", dependencies=TRUE, INSTALL_opts = c('--no-lock'))
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Data loading
data <- read.csv("cyber-security-1_question-response.csv")
head(data,10)
```

```
##                            learner_id quiz_question  question_type
## 1  77454a73-6b8b-46a2-8dee-35f36b6c4fc1         1.7.1 MultipleChoice
## 2  77454a73-6b8b-46a2-8dee-35f36b6c4fc1         1.7.1 MultipleChoice
## 3  a4fa6f89-a596-4d00-9397-420a348c398d         1.7.1 MultipleChoice
## 4  a4fa6f89-a596-4d00-9397-420a348c398d         1.7.1 MultipleChoice
## 5  a4fa6f89-a596-4d00-9397-420a348c398d         1.7.1 MultipleChoice
## 6  f27eec8c-eaf1-4e6a-90f0-d6d5b653285d         1.7.1 MultipleChoice
## 7  f27eec8c-eaf1-4e6a-90f0-d6d5b653285d         1.7.1 MultipleChoice
## 8  a4fa6f89-a596-4d00-9397-420a348c398d         1.7.1 MultipleChoice
## 9  a4fa6f89-a596-4d00-9397-420a348c398d         1.7.1 MultipleChoice
## 10 dce6f379-73d1-4968-a650-70d67cefd952         1.7.1 MultipleChoice
##    week_number step_number question_number response cloze_response
## 1            1           7               1      1,2             NA
## 2            1           7               1    1,2,3             NA
## 3            1           7               1    1,2,3             NA
## 4            1           7               1      1,2             NA
## 5            1           7               1      2,3             NA
## 6            1           7               1    1,2,3             NA
## 7            1           7               1    1,2,3             NA
## 8            1           7               1      2,3             NA
## 9            1           7               1    1,2,3             NA
## 10           1           7               1    1,2,3             NA
```

```
##               submitted_at correct
## 1  2016-07-06 10:37:05 UTC   false
## 2  2016-07-06 10:57:05 UTC    true
## 3  2016-07-11 09:09:50 UTC    true
## 4  2016-07-11 09:10:05 UTC   false
## 5  2016-07-11 09:10:18 UTC   false
## 6  2016-07-27 10:37:26 UTC    true
## 7  2016-07-27 10:37:31 UTC    true
## 8  2016-08-03 15:19:39 UTC   false
## 9  2016-08-03 15:20:10 UTC    true
## 10 2016-08-15 09:41:33 UTC    true
```

```
tail(data,10)
```

```
##                                learner_id quiz_question  question_type
## 76993                                          3.18.9 MultipleChoice
## 76994                                          3.18.9 MultipleChoice
## 76995 a1ad8719-444c-4012-a09a-f31c8fee955c        3.18.9 MultipleChoice
## 76996 a1ad8719-444c-4012-a09a-f31c8fee955c        3.18.9 MultipleChoice
## 76997 a1ad8719-444c-4012-a09a-f31c8fee955c        3.18.9 MultipleChoice
## 76998 62445c4c-d6a8-4bfd-8565-fccb203240e9        3.18.9 MultipleChoice
## 76999 62445c4c-d6a8-4bfd-8565-fccb203240e9        3.18.9 MultipleChoice
## 77000 b38fda59-ad46-411d-8eb1-6ac8722248a2        3.18.9 MultipleChoice
## 77001 b38fda59-ad46-411d-8eb1-6ac8722248a2        3.18.9 MultipleChoice
## 77002 b38fda59-ad46-411d-8eb1-6ac8722248a2        3.18.9 MultipleChoice
##       week_number step_number question_number response cloze_response
## 76993           3          18               9    1,2,3             NA
## 76994           3          18               9  1,2,3,4             NA
## 76995           3          18               9      3,4             NA
## 76996           3          18               9    1,3,4             NA
## 76997           3          18               9      1,3             NA
## 76998           3          18               9      1,3             NA
## 76999           3          18               9  1,2,3,4             NA
## 77000           3          18               9      1,3             NA
## 77001           3          18               9    1,2,3             NA
## 77002           3          18               9      2,3             NA
##                   submitted_at correct
## 76993 2016-10-08 13:44:44 UTC   false
## 76994 2016-10-08 13:45:04 UTC    true
## 76995 2016-10-08 14:11:07 UTC   false
## 76996 2016-10-08 14:11:10 UTC   false
## 76997 2016-10-08 14:11:19 UTC   false
## 76998 2016-10-08 16:58:45 UTC   false
## 76999 2016-10-08 16:59:51 UTC    true
## 77000 2016-10-08 19:39:24 UTC   false
## 77001 2016-10-08 19:39:37 UTC   false
## 77002 2016-10-08 19:39:40 UTC   false
```

```
dim(data)
```

```
## [1] 77002    10
```

```
str(data)
```

```
## 'data.frame':    77002 obs. of  10 variables:
##  $ learner_id     : chr  "77454a73-6b8b-46a2-8dee-35f36b6c4fc1" "77454a73-6b8b-46a2-8dee-35f36b6c4fc
```

```
##  $ quiz_question  : chr  "1.7.1" "1.7.1" "1.7.1" "1.7.1" ...
##  $ question_type  : chr  "MultipleChoice" "MultipleChoice" "MultipleChoice" "MultipleChoice" ...
##  $ week_number    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ step_number    : int  7 7 7 7 7 7 7 7 7 7 ...
##  $ question_number: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ response       : chr  "1,2" "1,2,3" "1,2,3" "1,2" ...
##  $ cloze_response : logi  NA NA NA NA NA NA ...
##  $ submitted_at   : chr  "2016-07-06 10:37:05 UTC" "2016-07-06 10:57:05 UTC" "2016-07-11 09:09:50 UTC
##  $ correct        : chr  "false" "true" "true" "false" ...
```

```
# Dropping unnecessary columns
data[,c("cloze_response","question_type","submitted_at")] <- list(NULL)
colnames(data)
```

```
## [1] "learner_id"      "quiz_question"   "week_number"     "step_number"
## [5] "question_number" "response"        "correct"
```

```
# How many unique elements are present in each column
for(i in colnames(data)){
  print(sum(!duplicated(data[i])))
}
```

```
## [1] 3410
## [1] 22
## [1] 3
## [1] 5
## [1] 9
## [1] 32
## [1] 2
```

```
# How many empty cells are present in each column
for(i in colnames(data)){
  print(sum(data[i]==""))
}
```

```
## [1] 401
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
```

```
# Dropping empty cells
data[data==""]<-NA
data<-data[complete.cases(data),]
sum(data=="")
```

```
## [1] 0
```

```
# which question is mostly correct in between all quizzes
data %>%
  group_by(question_number, correct) %>%
  summarise(a_sum=sum(correct=="true",
                      correct=="false"))
```

```
## `summarise()` has grouped output by 'question_number'. You can override using the `.groups` argument
```

```
## # A tibble: 18 x 3
```

```
## # Groups:   question_number [9]
##    question_number correct a_sum
##              <int> <chr>   <int>
##  1               1 false    9290
##  2               1 true     9147
##  3               2 false    3585
##  4               2 true     8133
##  5               3 false    8960
##  6               3 true     7517
##  7               4 false     644
##  8               4 true     4729
##  9               5 false    1115
## 10               5 true     4581
## 11               6 false    5661
## 12               6 true     4316
## 13               7 false    1317
## 14               7 true     1423
## 15               8 false     984
## 16               8 true     1394
## 17               9 false    3096
## 18               9 true      709
```

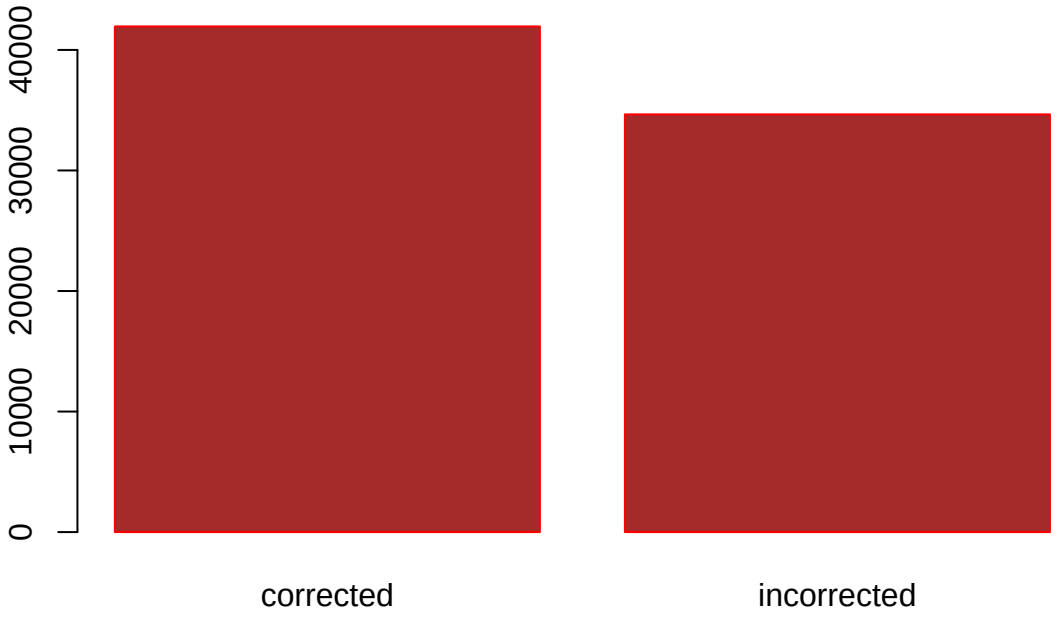
```
# question 1 is mostly correct



# Total number of Corrected questions and incorrected questions
paste("Total number of corrected question:",p=sum(data$correct=="true"))
```

```
## [1] "Total number of corrected question: 41949"
```

```
paste("Total number of incorrected question:",p=sum(data$correct=="false"))
```

```
## [1] "Total number of incorrected question: 34652"
```

```
H <- c(sum(data$correct=="true"),sum(data$correct=="false"))
M <- c("corrected","incorrected")
# Plot the bar chart
barplot(H,names.arg=M,xlab="",ylab="",col="brown",
        main="Total number of Corrected questions and incorrected questions",border="red")
```




```
# Top 10 high scorers (learning_id)
df1<- data.frame(data)
df2<-subset(df1, correct=="true")
z<-df2 %>%
  group_by(learner_id, correct=="true") %>%
  summarise(true_sum=sum(correct=="true",
                         correct=="false"))
```

```
## `summarise()` has grouped output by 'learner_id'. You can override using the `.groups` argument.
```

```
z[order(-z$true_sum),]
```

```
## # A tibble: 3,387 x 3
## # Groups:   learner_id [3,387]
##    learner_id                           `correct == "true"` true_sum
##    <chr>                                <lgl>                   <int>
##  1 2f483763-596f-4ede-ba44-20937ce7eda9 TRUE                       24
##  2 4ee40d44-310f-4db3-b279-ee88f1afb7be TRUE                       24
##  3 623cc9d9-d068-4230-b024-80f349027b05 TRUE                       24
##  4 f8861608-c0bc-4e00-9bdd-fd0e48b5a68a TRUE                       24
##  5 6681caaf-8256-4a24-8566-6fcc4a32c570 TRUE                       23
##  6 7ec35392-2049-46f8-8db1-e3171df6e98f TRUE                       23
##  7 83ee7ae7-f769-4336-a19d-aea944464536 TRUE                       23
##  8 ce69eaa9-43b2-4740-807b-b41213b0d7f4 TRUE                       23
##  9 d153f4ef-c4c2-4505-bf98-f8daea81546a TRUE                       23
## 10 d5014389-d81f-44e0-aba8-4f579a38ebf0 TRUE                       23
## # ... with 3,377 more rows
```

```
# Which quiz question is most correctly marked
df1<- data.frame(data)
df2<-subset(df1, correct=="true")
z<-df2 %>%
  group_by(quiz_question, correct) %>%
  summarise(true_sum=sum(correct=="true",
```

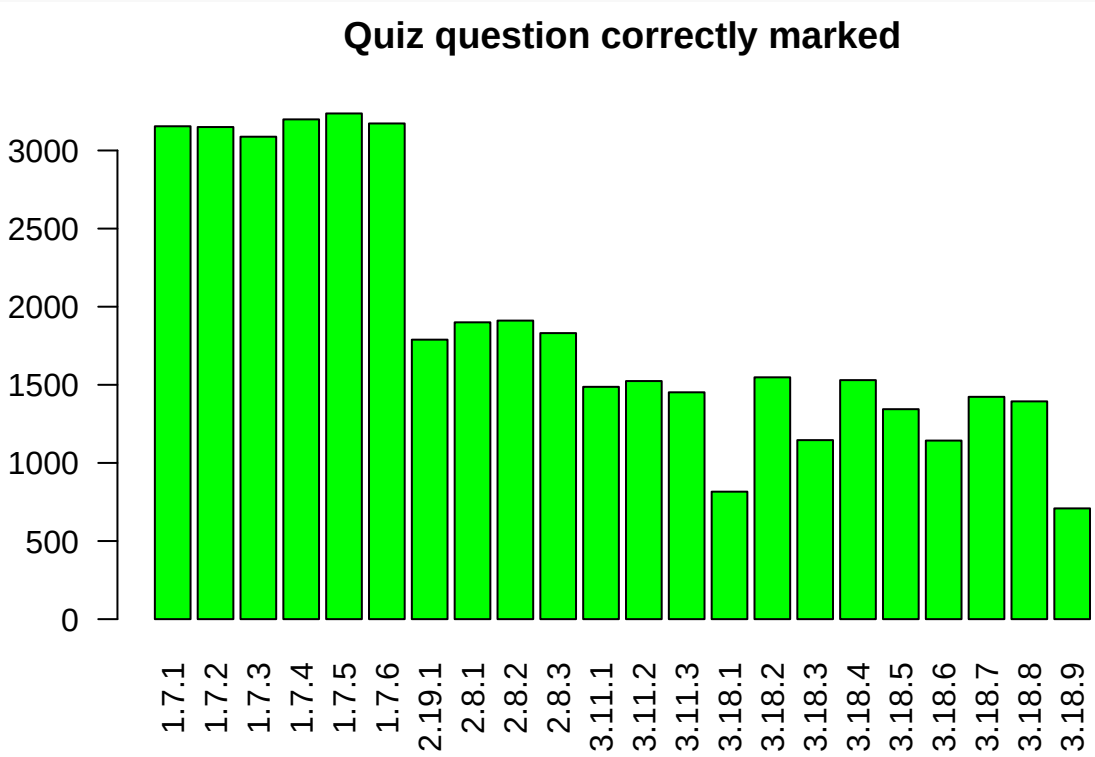
```
                        correct=="false"))
```

```
## `summarise()` has grouped output by 'quiz_question'. You can override using the `.groups` argument.
```

```
z[order(-z$true_sum),]
```

```
## # A tibble: 22 x 3
## # Groups:   quiz_question [22]
##    quiz_question correct true_sum
##    <chr>         <chr>      <int>
##  1 1.7.5         true        3237
##  2 1.7.4         true        3199
##  3 1.7.6         true        3173
##  4 1.7.1         true        3155
##  5 1.7.2         true        3150
##  6 1.7.3         true        3088
##  7 2.8.2         true        1911
##  8 2.8.1         true        1900
##  9 2.8.3         true        1831
## 10 2.19.1        true        1789
## # ... with 12 more rows
```

```
# Plot the bar chart
barplot(z$true_sum,names.arg=z$quiz_question,las=2,xlab="",ylab="",col="green",
        main="Quiz question correctly marked",border="black")
```




```
# Which quiz question is most incorrectly marked
df1<- data.frame(data)
df2<-subset(df1, correct=="false")
z<-df2 %>%
  group_by(quiz_question, correct) %>%
  summarise(true_sum=sum(correct=="true",
```

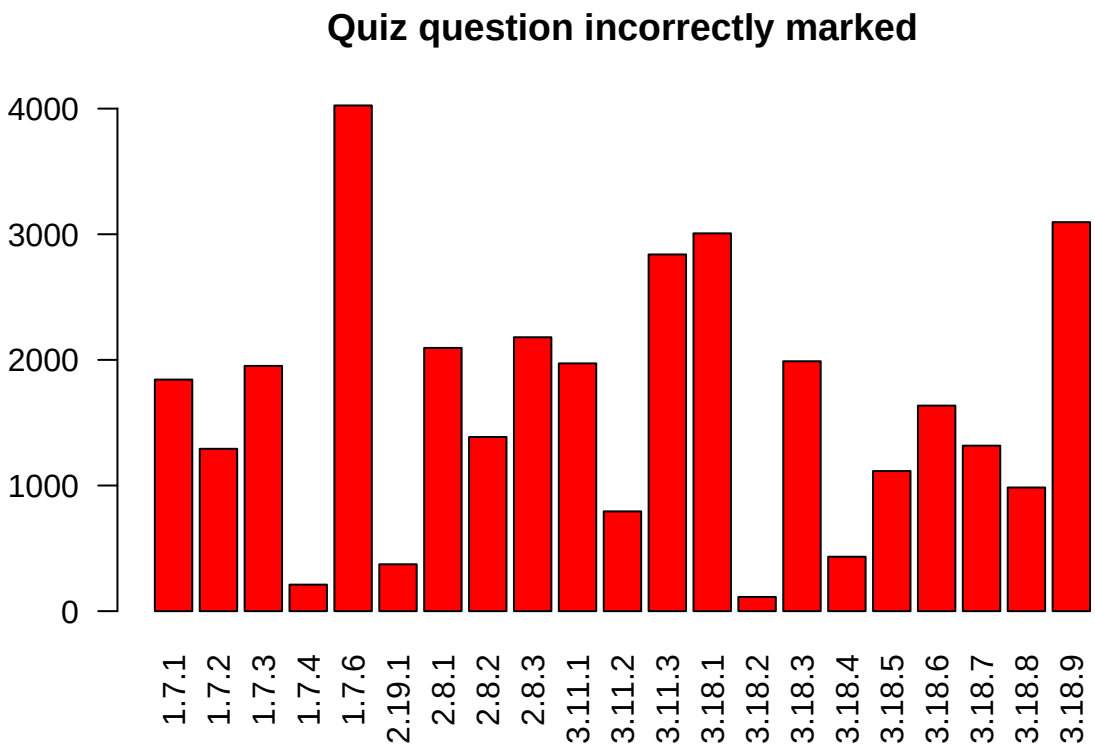
```
                          correct=="false"))
```

```
## `summarise()` has grouped output by 'quiz_question'. You can override using the `.groups` argument.
```

```
z[order(-z$true_sum),]
```

```
## # A tibble: 21 x 3
## # Groups:   quiz_question [21]
##    quiz_question correct true_sum
##    <chr>         <chr>      <int>
##  1 1.7.6         false       4025
##  2 3.18.9        false       3096
##  3 3.18.1        false       3007
##  4 3.11.3        false       2839
##  5 2.8.3         false       2180
##  6 2.8.1         false       2095
##  7 3.18.3        false       1989
##  8 3.11.1        false       1972
##  9 1.7.3         false       1952
## 10 1.7.1         false       1843
## # ... with 11 more rows
```

```
# Plot the bar chart
barplot(z$true_sum,names.arg=z$quiz_question,las=2,xlab="",ylab="",col="red",
        main="Quiz question incorrectly marked",border="black")
```




```
# which step number got the most correctly marked
df1<- data.frame(data)
df2<-subset(df1, correct=="true")
z<-df2 %>%
  group_by(step_number, correct) %>%
  summarise(true_sum=sum(correct=="true",
```

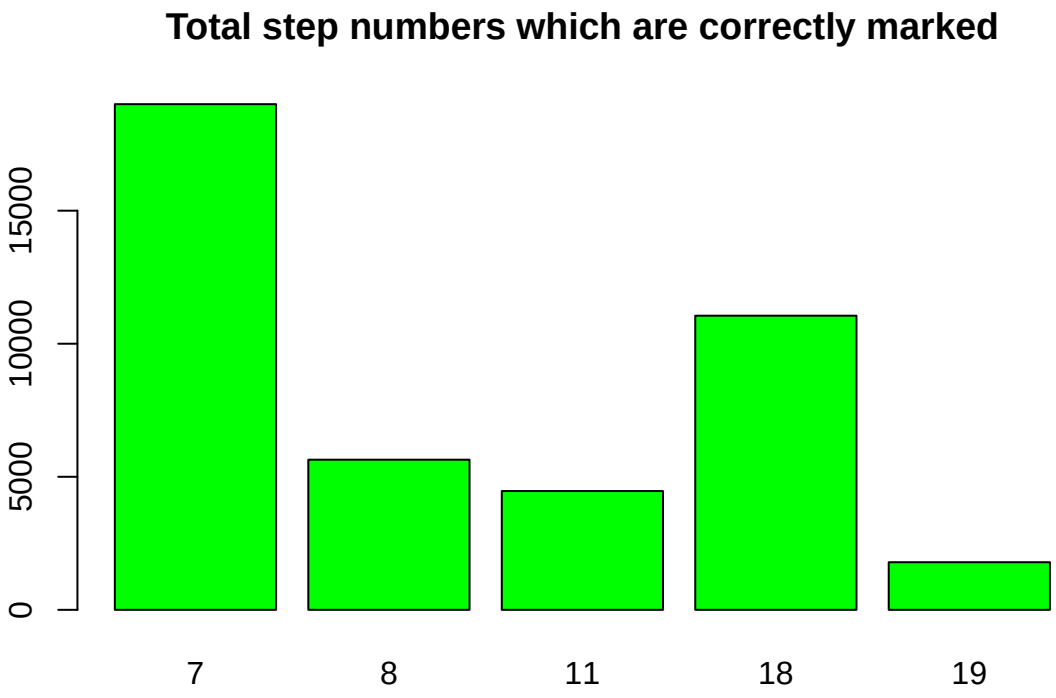
```
                          correct=="false"))
```

```
## `summarise()` has grouped output by 'step_number'. You can override using the `.groups` argument.
```

```
z[order(-z$true_sum),]
```

```
## # A tibble: 5 x 3
## # Groups:   step_number [5]
##   step_number correct true_sum
##         <int> <chr>      <int>
## 1           7 true       19002
## 2          18 true       11053
## 3           8 true        5642
## 4          11 true        4463
## 5          19 true        1789
```

```
# Plot the bar chart
barplot(z$true_sum,names.arg=z$step_number,xlab="",ylab="",col="green",
        main="Total step numbers which are correctly marked",border="black")
```




```
# which step number got the most incorrectly marked
df1<- data.frame(data)
df2<-subset(df1, correct=="false")
z<-df2 %>%
  group_by(step_number, correct) %>%
  summarise(true_sum=sum(correct=="true",
                         correct=="false"))
```

```
## `summarise()` has grouped output by 'step_number'. You can override using the `.groups` argument.
```

```
z[order(-z$true_sum),]
```

```
## # A tibble: 5 x 3
## # Groups:   step_number [5]
##   step_number correct true_sum
```

```
##         <int> <chr>      <int>
## 1          18 false      13690
## 2           7 false       9323
## 3           8 false       5661
## 4          11 false       5605
## 5          19 false        373
```

```
# Plot the bar chart
barplot(z$true_sum,names.arg=z$step_number,xlab="",ylab="",col="red",
        main="Total step numbers which are incorrectly marked",border="black")
```