Notes for countries histogram

Christian Fox

28 November 2020

I decided to use data for the students detected country.

```
# add "head(cse$detected_country)" here.
setwd("H:/cyber-security/data/FutureLearn MOOC Dataset") # this is running from uni laptop
cse = read.table('cyber-security-1_enrolments.csv', header=TRUE, sep=',') # creates table from excel do
cse.df = data.frame(cse) # turns table to data frame and stores in environment
# could make column width smaller here.
head(cse.df) # showing first 6 rows
```

Notes on data:

To be able to easily manipulate the detected_country column of this data I named the variable and set this coulmn as a data frame.

```
# first create data frame with countries and frequency:
list_of_countries = cse$detected_country
# now find the 'count'/frequency of each country
count.df = as.data.frame(table(list_of_countries))
```

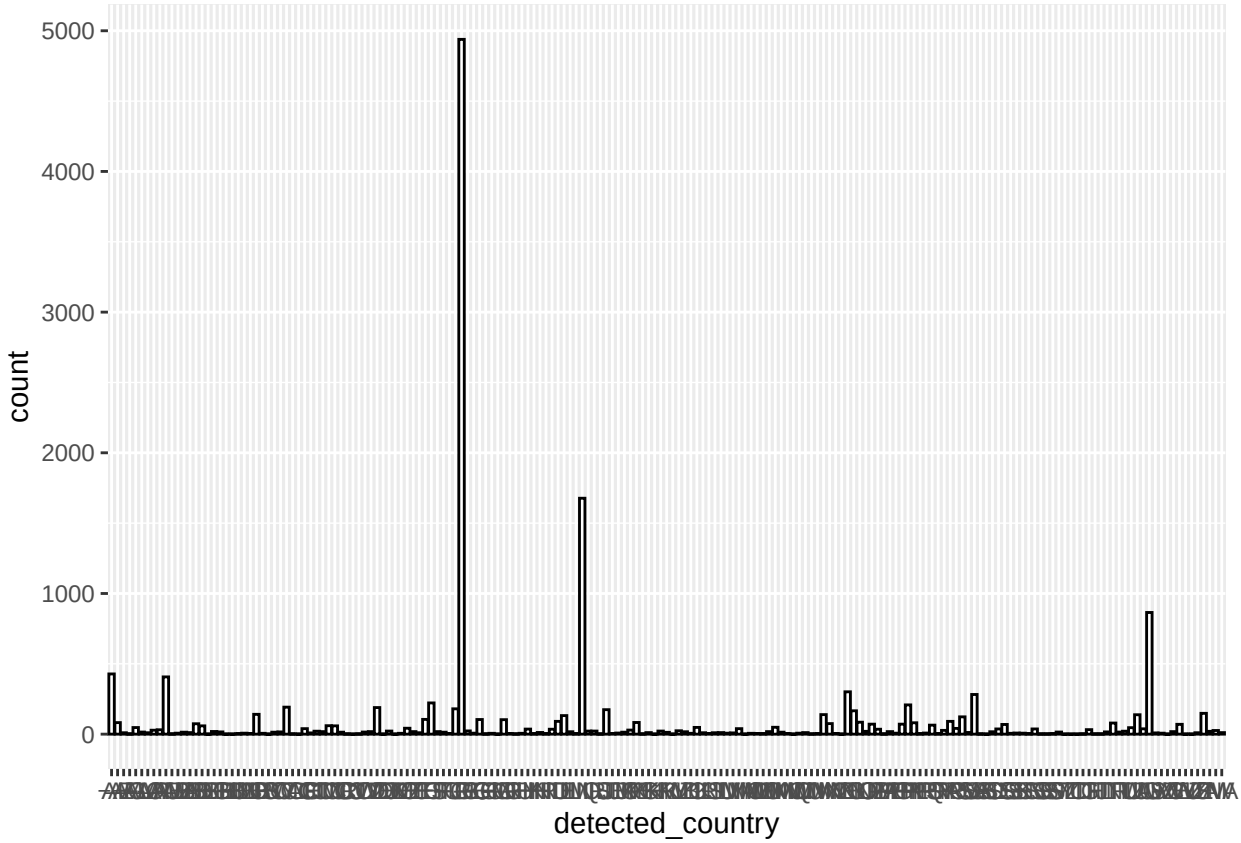
CRISP-DM methodology update

A further undetsranding of the data has been achieved by munging the data. Therefore, I was subconsciously alternating from data understanding to data preperation.

The total number of students is 14383. There are 428 students who's country is unknown. This gives a percentage of students with no detected country of 2.9757352%.

A histogram for countries was plotted:

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

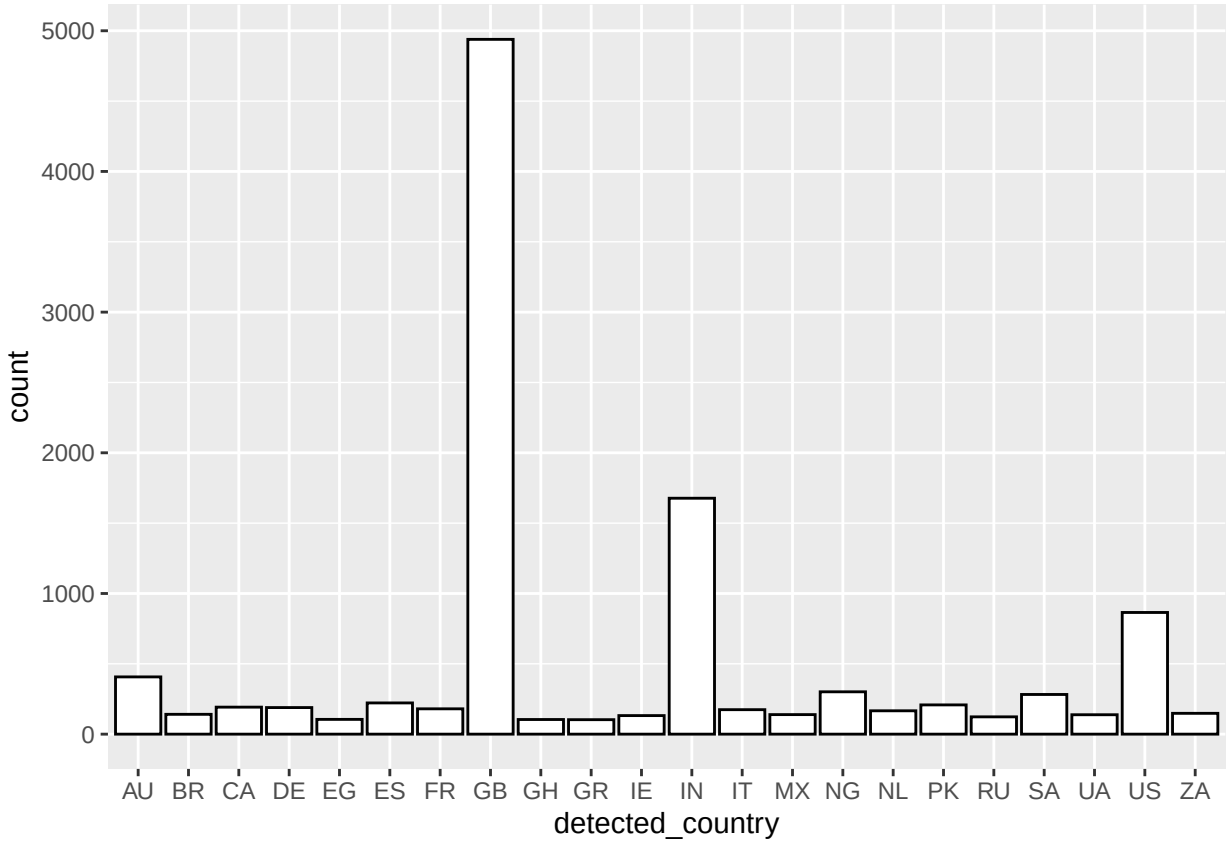


CRISP-DM methodology update

Producing a histogram was the first modelling step. After a first look at the graph, it is clear that a back-track in the steps in necessary, since the x labels are unreadable. #### Data understanding Studying the data frame of countries, it is seen that there are 183 unique countries with a large portion only having 1 or 2 students. #### Back to data preperation This lead to sorting through the countries list and a decision was made to omit countries with under 100 participating students, as well as the undetected country row.

```
# omitting the countries with < 100 students/frequency
countries_over_100 = count.df[which(count.df$Freq >= 100 & count.df$list_of_countries != "--"),]$list_o
```

The following hosogram was produced:



This plot is more readable, as well as more relevant as the undetected countries bar didnt provide much useful information.