CSC8631Report

Yang Cong

2021/11/23

1. Analyse the gender, age range and country of learners

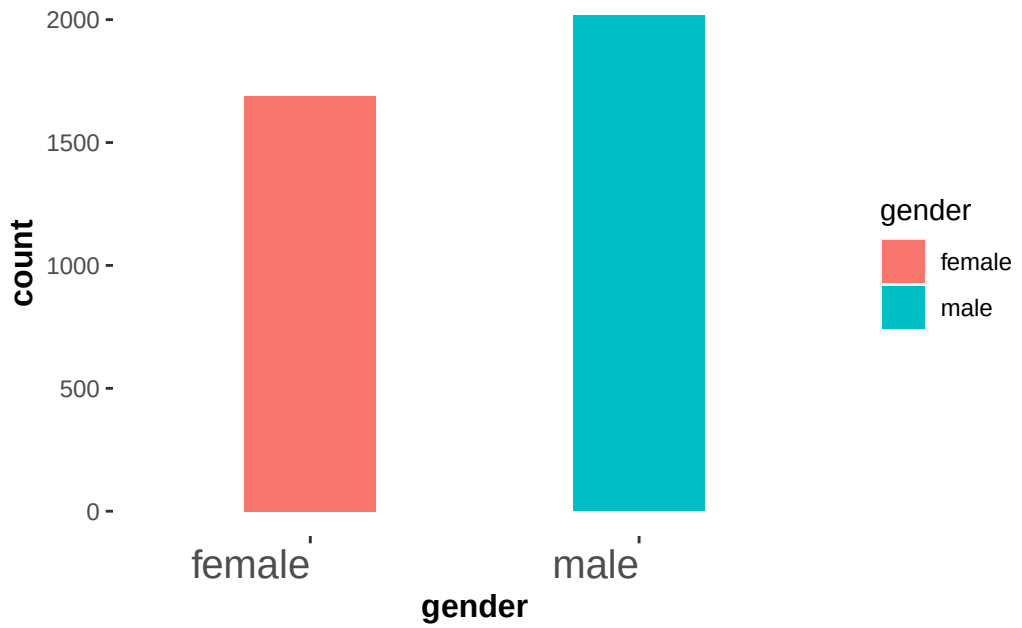
Firstly, we would like to get an overall picture of the gender and age distribution of the learners in order to get a general idea of the profile of our subjects.

Analyse the gender

When making statistical distinctions by gender, we first eliminate all Unknown values. Using codes:

```
enr_gender = filter(enrolments, gender == "male" | gender == "female" )
```

And then we can easily draw the bar charts. From the plot we can see that there are more male learners than female.



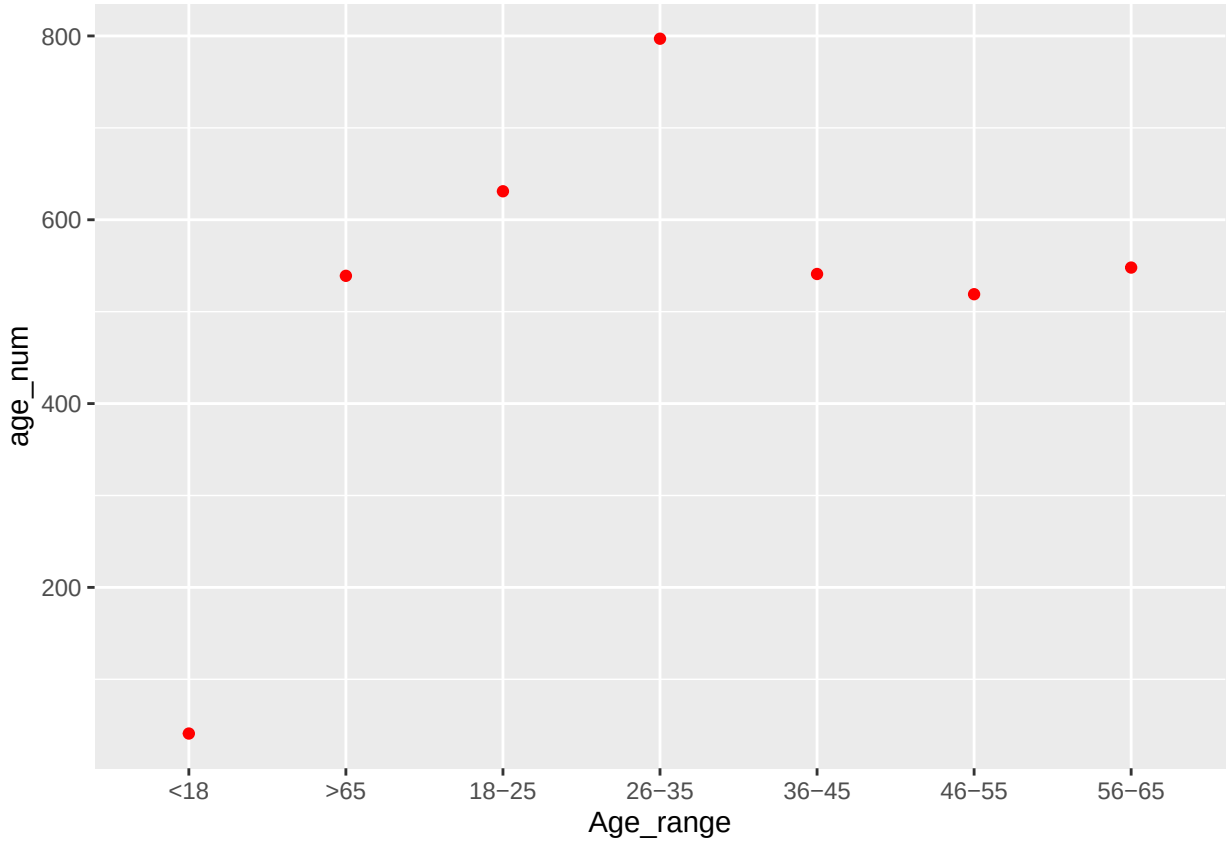
Analyse the age range

Same as analyse the gender, we should firstly remove the Unknown data.

And to reduce the amount of arithmetic, I first simplified the data frame by extracting only the useful data. This operation will be used frequently later in the data analysis and is useful for improving the performance of the analysis.

```
enr_age = select(enr_age, learner_id, age_range)
```

The graph shows that the age distribution of the learners is very wide, but most of the students are already adults. We can therefore assume that they are more mentally mature.



Analyse the country

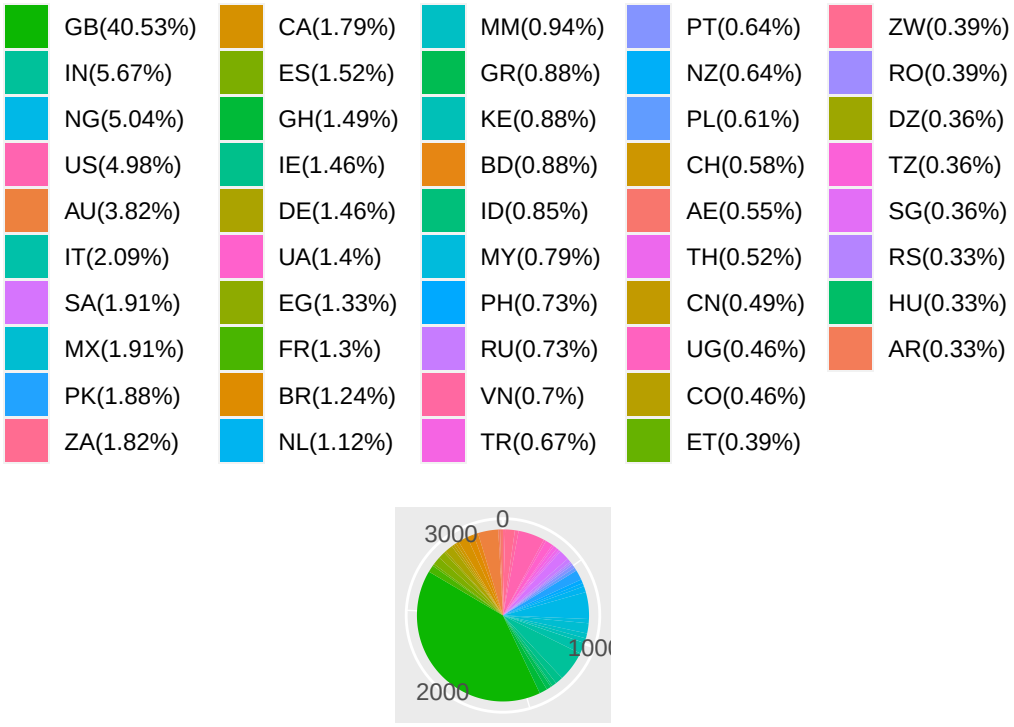
It is interesting to analyse the nationality of the learners, we can derive their learning habits from their nationality, which is related to the culture of their country.

In fact there are so many countries that it is difficult for us to show them all in one map. We have therefore eliminated the countries with a small number of learners so that we can easily focus on the more valuable objects of analysis.

```
enr_country_num = filter(enr_country_num, country_num>10)
```

If I used a simple bar chart, some of the countries would stand out too much and detract from the overall graph, so I have chosen to use a fan chart. To give a more visual representation of the number of learners in each country, I have also calculated the percentage of learners and labelled it on the graph.

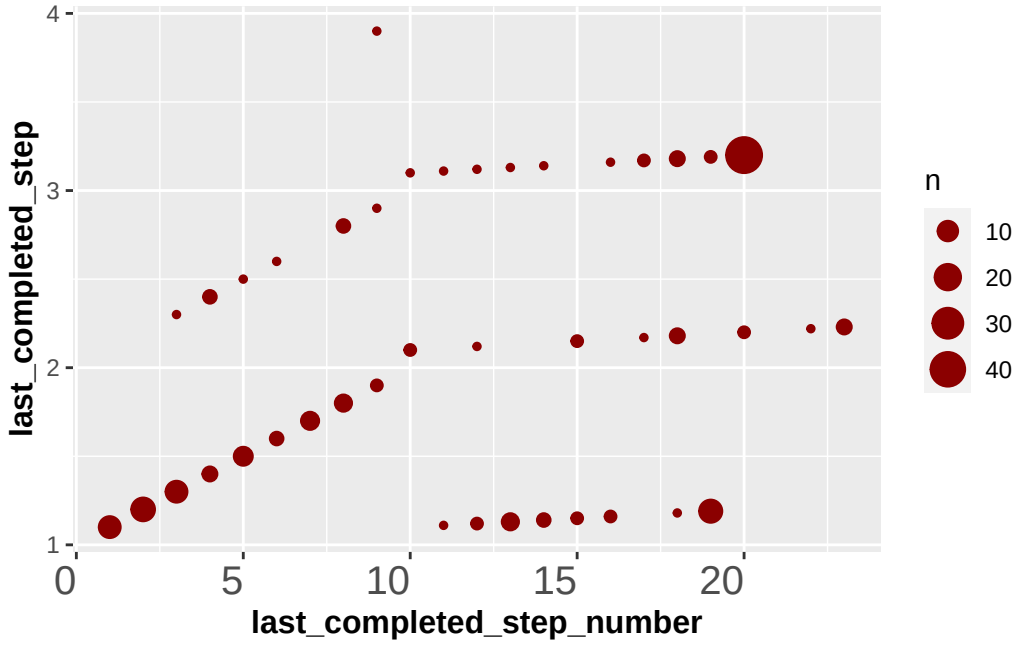
The graph shows that most of the students are from the UK, India and Nigeria. And as non-native English learners make up a large group, I think schools should provide them with some language assistance.



2. Analyse the leaving survey response

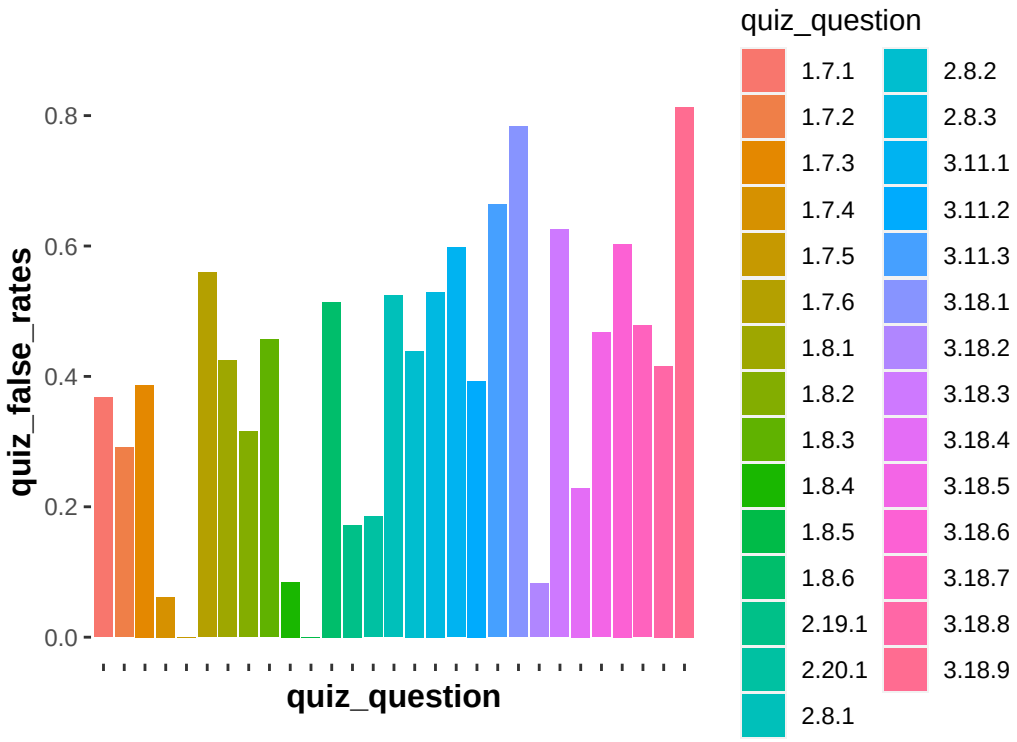
In this step I tried to analyse what stage the learners were at when they left the study.As you can see from the graph most people managed to complete all the steps after several attempts, but there were some who gave up just as soon as they started trying, I think the mindset of this group needs to be adjusted, their patience is too low.

At the same time there is another group of people who, despite many attempts, are still at a more rudimentary stage and I think this group needs some technical guidance from the school.



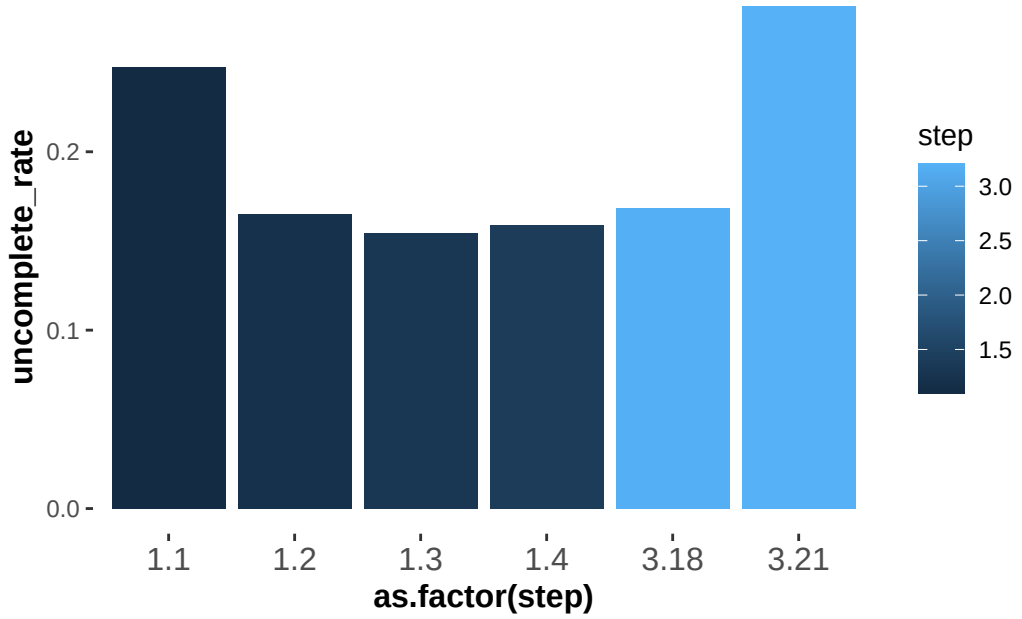
3. Analyse the false rate

In this step I counted the error rate of each question and showed the ones with the highest error rate. I believe that the teaching methods for the knowledge points covered in these questions need to be improved.



3. Analyse the activity step

In this section we look at the unfinished rate of the activity. I have counted the failure rates for each step of the campaign and listed the six steps with the highest failure rates. I think the design of these six steps may need to be improved.



4. Relationship between learners' gender/nationality/age and correctness

I had already learned about the gender, age and nationality of the learners, and now I was interested in the relationship between these attributes of the learners and their correctness.

Data preparation

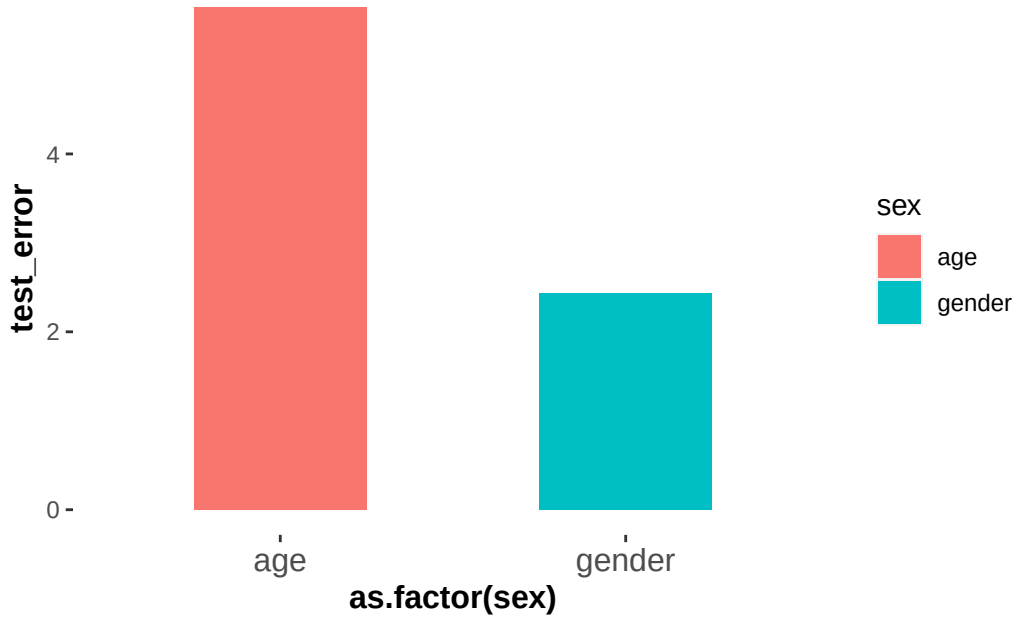
First you have to de-duplicate the ids to see how many learners there are. Then, calculate the percentage of correct answers for each learner, and remove the null value. The final data structure obtained is shown below.

```
## # A tibble: 3 x 5
##   learner_id                           gender country age_range true_rate
##   <chr>                                <chr>  <chr>   <chr>         <dbl>
## 1 fffc0222-845c-460c-859c-d31b8e492dd4 female LT      36-45         0.643
## 2 ffa400c1-5e27-45b0-a6b4-4ea1d22e2c4d female FR      >65           0.591
```

```
## 3 ffa0781f-7a10-48a2-aa7c-122062213976 female CH      56-65         0.632
```

Fit the model

Here the data set is divided into a test set and a validation set, and the model is fitted with gender and age as the corresponding variables respectively, and by comparing the error rates we can see that there is a greater association between the learners' correct rates and gender.



Study of the difference between male and female students in terms of correct answer rates

Finally I also looked at whether there was a large difference between male and female learners in terms of the percentage of questions answered correctly. As can be seen from the graph below, there is actually not much difference between males and females in terms of the percentage of questions answered correctly.