csc8631 Report

Yahan Wang 200784463

11/23/2021

Business Understanding

The task of the business understanding phase is to articulate the goals and requirements of data mining from a business perspective and translate them into specific data mining problems. The objective of this paper is to measure, collect, analyse and report data about learners and their environment in order to understand and optimise their learning environment. For learners who are less motivated to attend classes, this may be influenced by a number of factors such as gender, difficulty of the course taken, country, major, and age. The key to optimising its learning environment is therefore to portray the factors that influence learner motivation based on learner characteristics and other supplementary data sources (e.g. access to on-campus facilities, Virtual Learning Environment (VLE) and Re-Cap visits, and student welfare referrals).

Data Understanding

Data quality checking and initial characterisation, which is essentially a process of data collection and familiarisation, Found in practice, age range of learners, gender, highest level of education, employment status, region of employment, country, , survey responses, leaving survey responses, step activity, question responses, video statistics . . . . etc,can affect learners' motivation to learn. The data used in this paper are derived from a school's student information management warehouse, with detailed records of learner characteristics and other complementary data sources (e.g. use of on-campus facilities, Virtual Learning Environment (VLE) and Re-Cap access, and student welfare referrals). In order to facilitate student management, schools set up learner ids for each student, record enrolment information (e.g. gender, country, major, age, etc.) and keep records of step activities, question answers, etc. This paper will divide students by the number of questions they answer, the difference in the number of questions answered by students is more significant, so this paper will model the factors that influence the number of questions answered by students to be explored

Data preparation.

The data preparation phase covers all the work involved in constructing the final dataset (for modelling analysis) from the raw rough data, including steps such as data cleaning and variable selection. ## Selecting data Task:First select the seven enrolments form.This is because the seven forms contain a lot of information about the learner, such as age range, gender, highest level of education, employment status, region of employment, country, etc.This may have an impact on the number of questions answered, the number of step activities, and the number of videos viewed by learners.Seven question response forms were then selected, which allowed for statistics on the number of responses per learner and the number and frequency of responses per question.Then select the seven Step activity tables, which will give you an idea of the number of people who completed and did not complete the activity at each step.Based on these video stats tables, you can see the popularity of each of the step videos.All this data has a crucial impact on portraying the factors that influence learners' motivation and optimising its learning environment. Output:In the seven enrolments tables, select age_range, highest_education_level, employment_status, employment_area,country and detected_country as the data to be used later, because these data can be used to explore the correlation with learner motivation, the enrolled at and unenrolled at cannot be used to explore learner motivation, and all the roles are learners, so they are removed. The seven question response forms leave the learner id and quiz_question, and correct.Because the question type is only one type of question that is MultipleChoice. The week number, step number and question number are all included in the quiz_question, The answer in the response is used to determine whether the correct is TRUE or FALSE, and the final result only depends on the correct so all this data needs to be removed. The seven Step activity tables need to leave the learner_id and step columns, and also need to leave first_visited_at, last_completed_at to determine if the learner is responding to the answer, based on the observation that as long as first_visited_at, last_completed_at both information is available. In the question response form there must be a response and a correct(TRUE or FALSE).However, the information on week_number and step_number is contained in the step so these data need to be deleted. ## Describe data

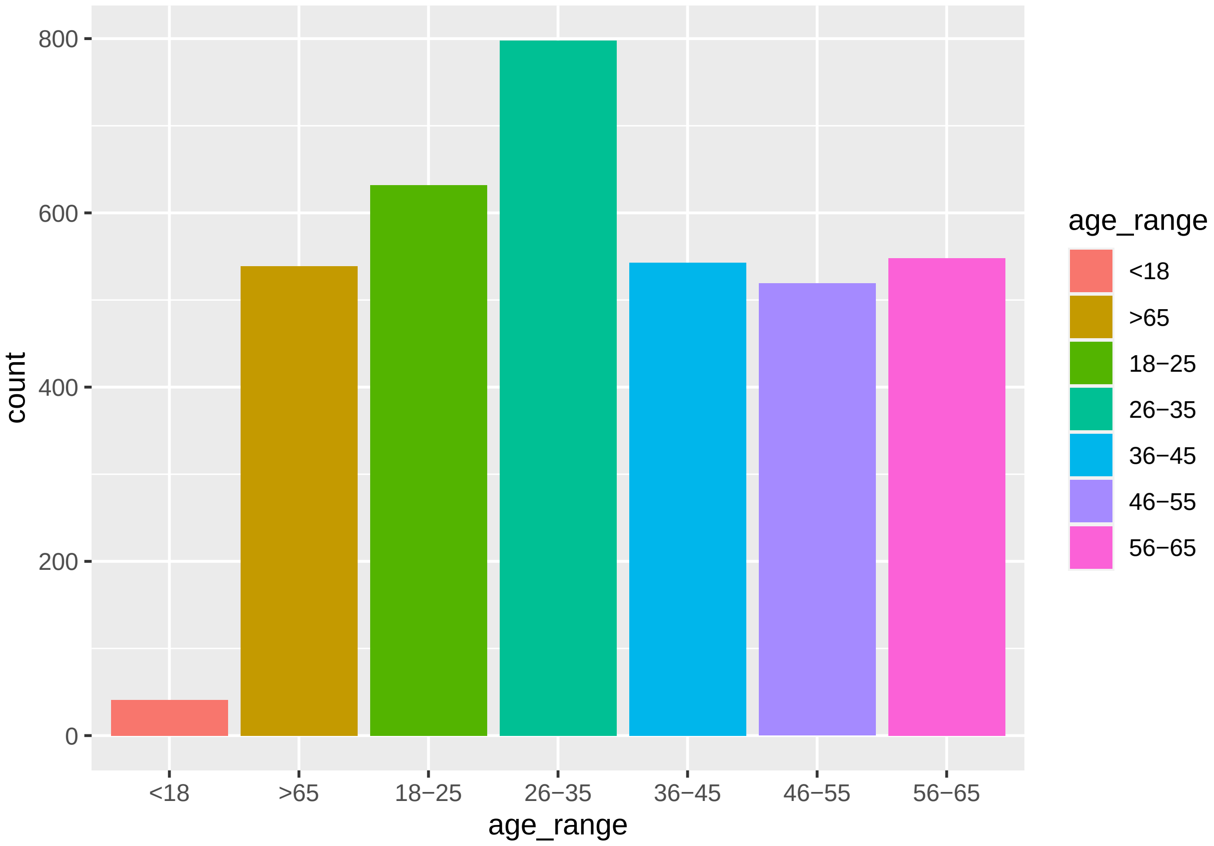
```
str(cyber.security.enrolments,vec.len  =1)
```

```
## tibble [35,225 x 13] (S3: tbl_df/tbl/data.frame)
##  $ learner_id             : chr [1:35225] "160d6600-ea0e-4568-bfa9-5d7cd5b8e61b" ...
##  $ enrolled_at            : chr [1:35225] "2016-08-10 14:28:49 UTC" ...
##  $ unenrolled_at          : chr [1:35225] "" ...
##  $ role                   : chr [1:35225] "learner" ...
##  $ fully_participated_at  : chr [1:35225] "" ...
##  $ purchased_statement_at : chr [1:35225] "" ...
##  $ gender                 : chr [1:35225] "Unknown" ...
##  $ country                : chr [1:35225] "Unknown" ...
##  $ age_range              : chr [1:35225] "Unknown" ...
##  $ highest_education_level: chr [1:35225] "Unknown" ...
##  $ employment_status      : chr [1:35225] "Unknown" ...
##  $ employment_area        : chr [1:35225] "Unknown" ...
##  $ detected_country       : chr [1:35225] "GB" ...
```

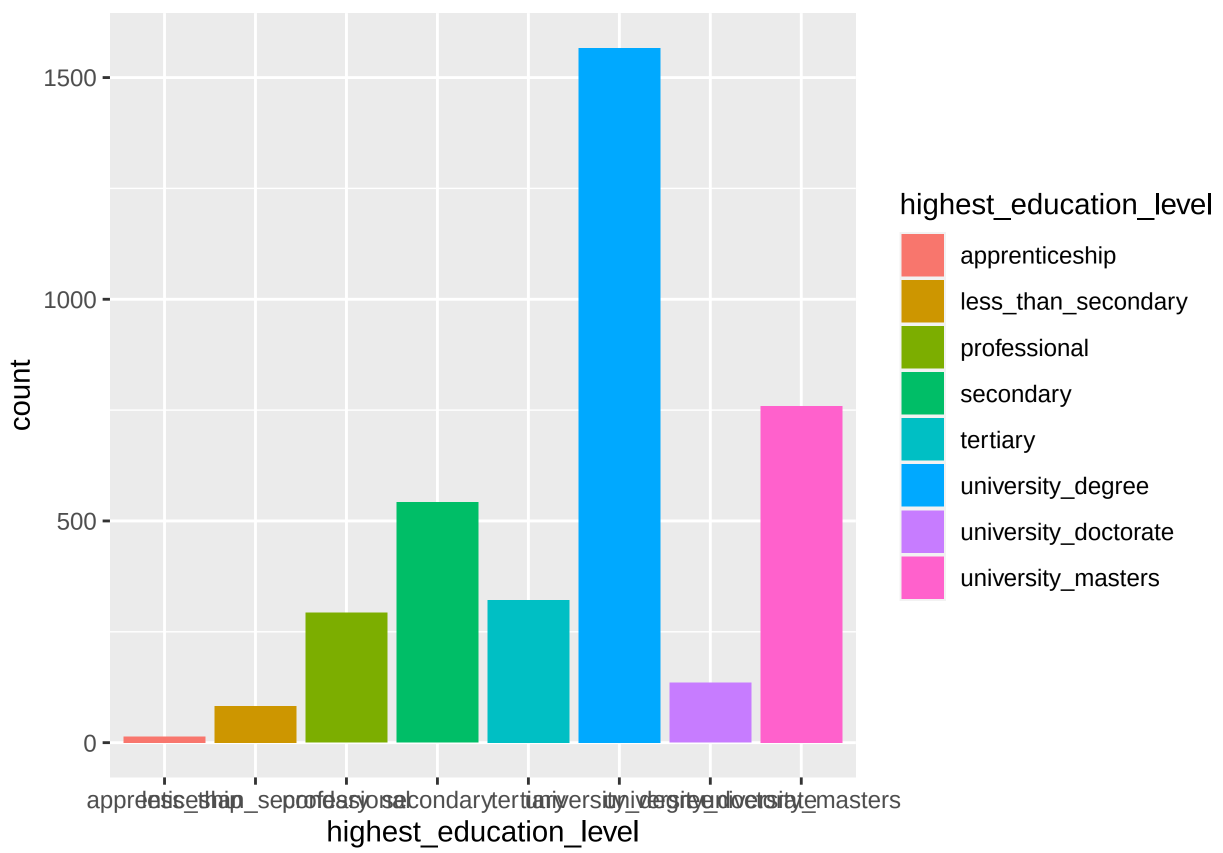
Including Plots

You can also embed plots, for example:

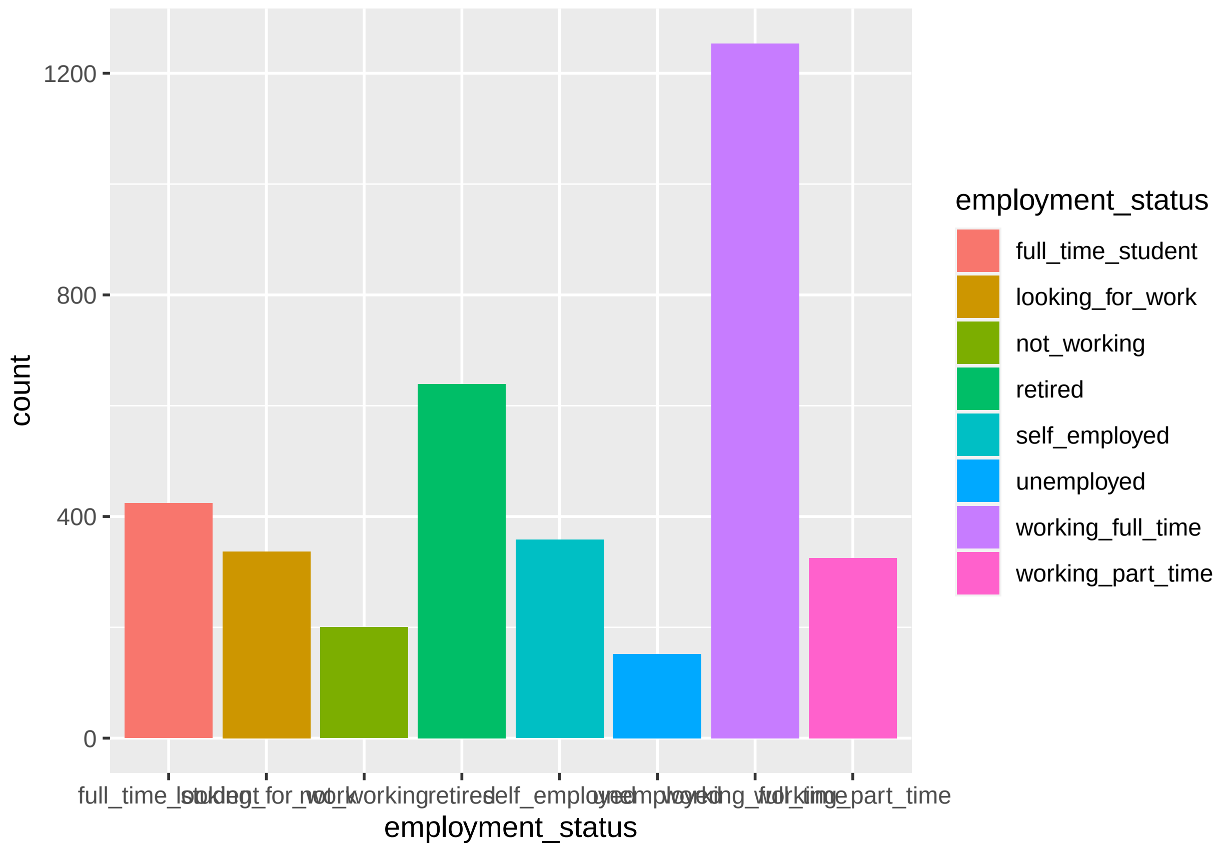
```
ggplot(age_range,aes(x=age_range, fill=age_range))+geom_bar()
```



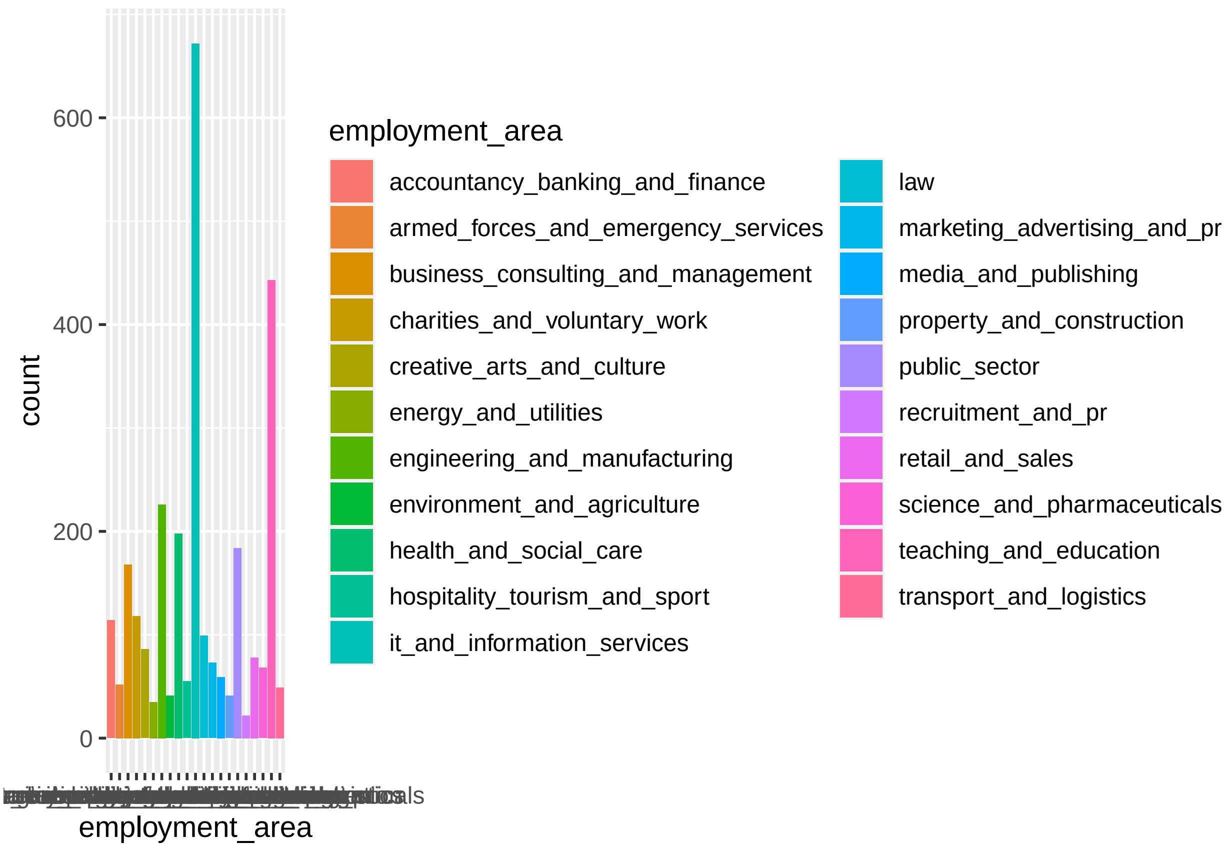
```
ggplot(highest_education_level,aes(x=highest_education_level, fill=highest_education_level))+geom_bar()
```



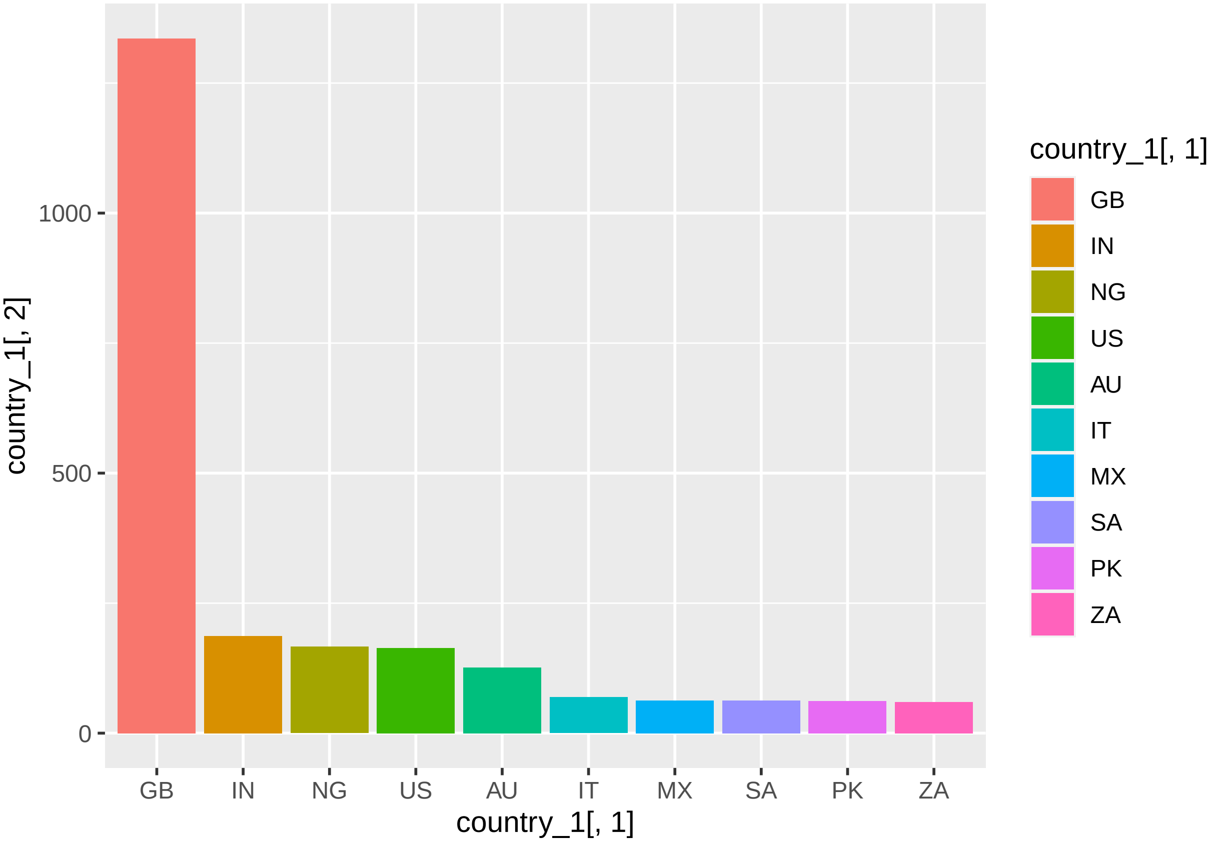
```
ggplot(employment_status,aes(x=employment_status, fill=employment_status))+geom_bar()
```



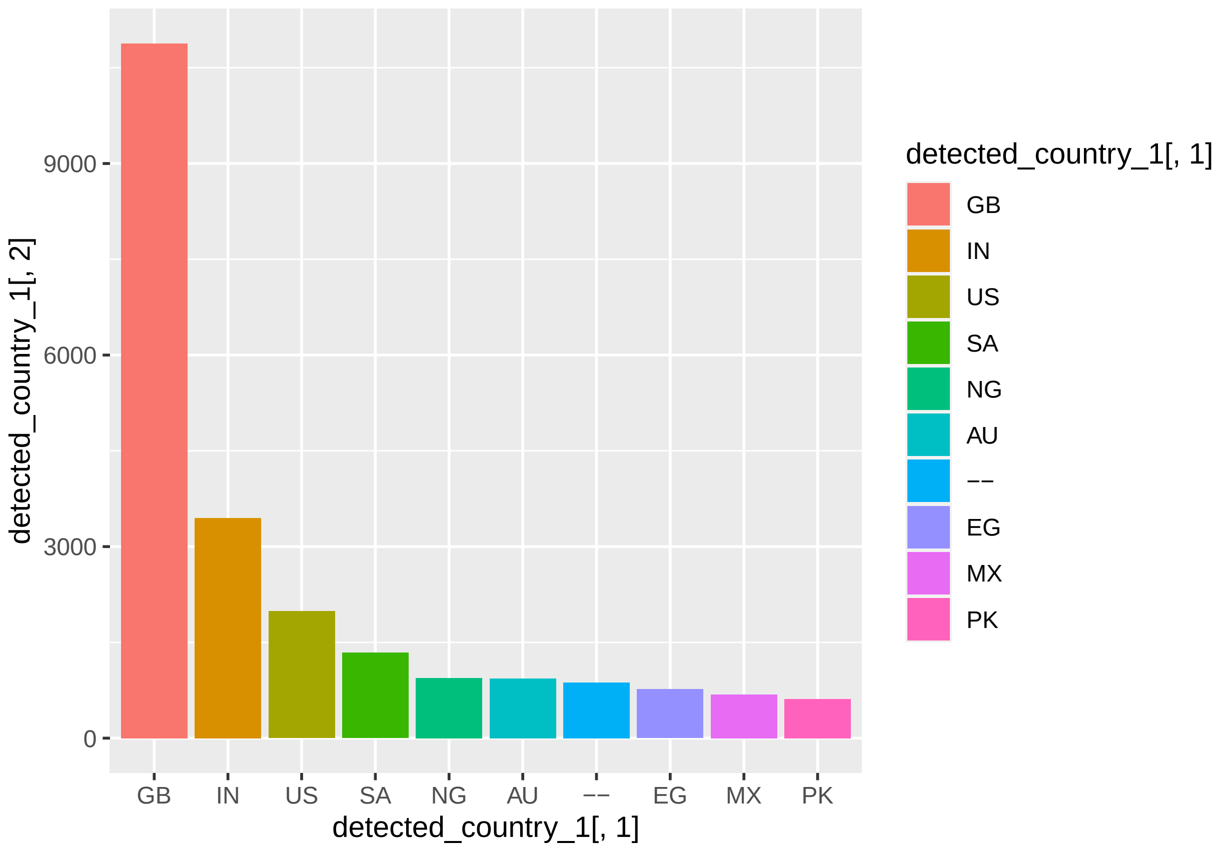
```
ggplot(employment_area,aes(x=employment_area, fill=employment_area))+geom_bar()
```



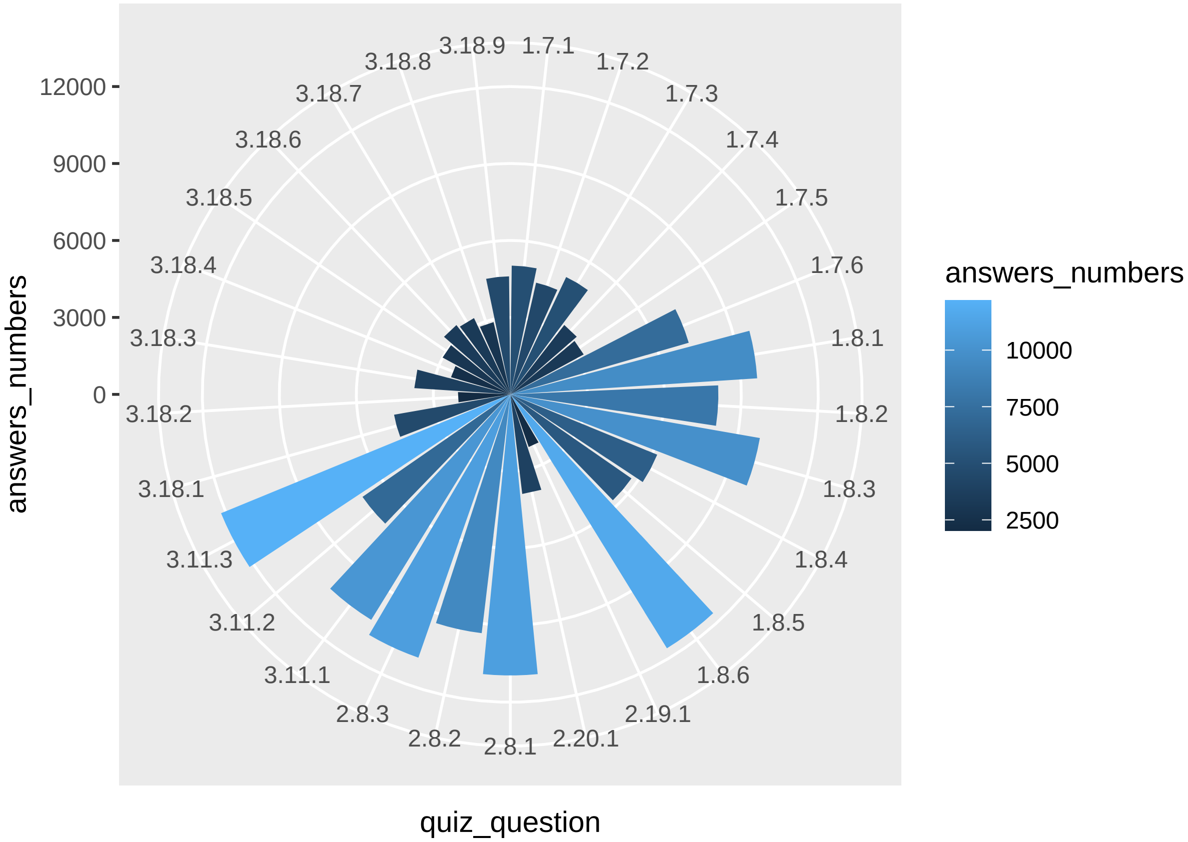
```
country_1 = data.frame(country_1)
ggplot(country_1,aes(x=country_1[,1],y=country_1[,2],fill=country_1[,1])) +
  geom_bar(stat = "identity")
```

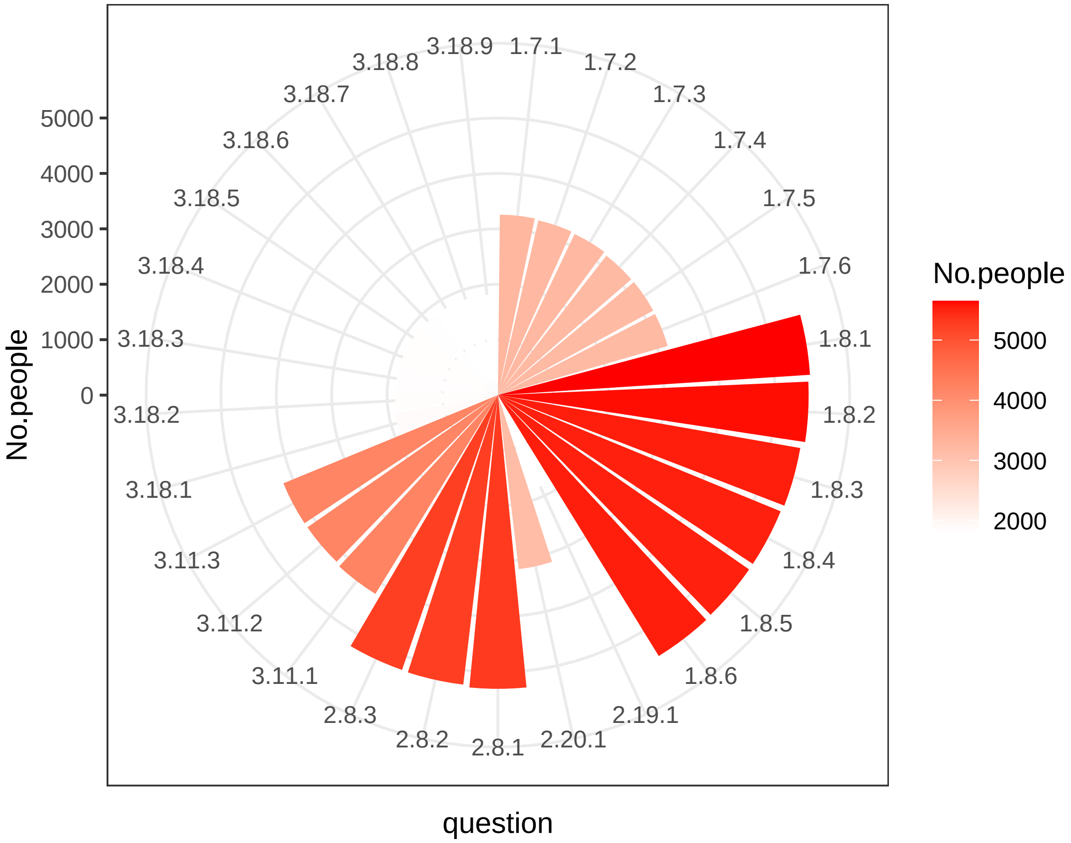


```
detected_country_1 = data.frame(detected_country_1)
ggplot(detected_country_1,aes(x=detected_country_1[,1],y=detected_country_1[,2],fill=detected_country_1
  geom_bar(stat = "identity")
```

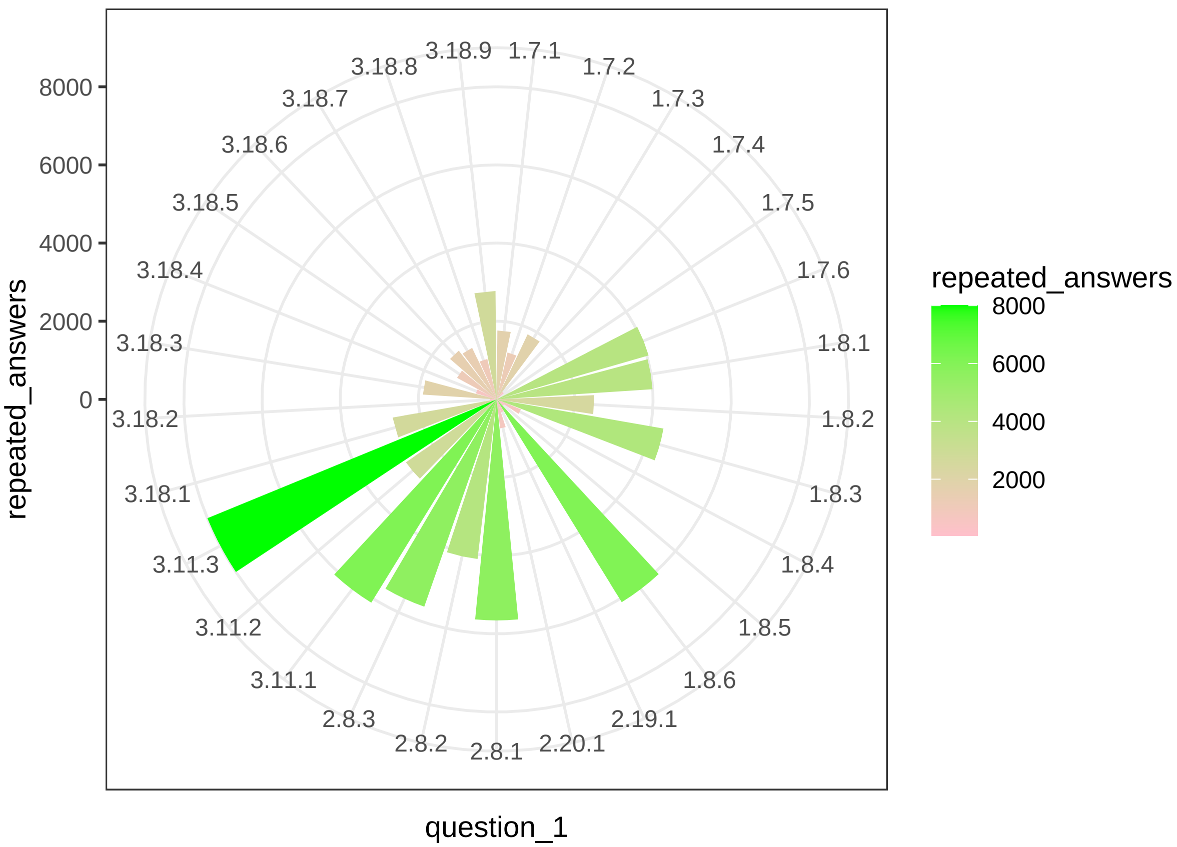


Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

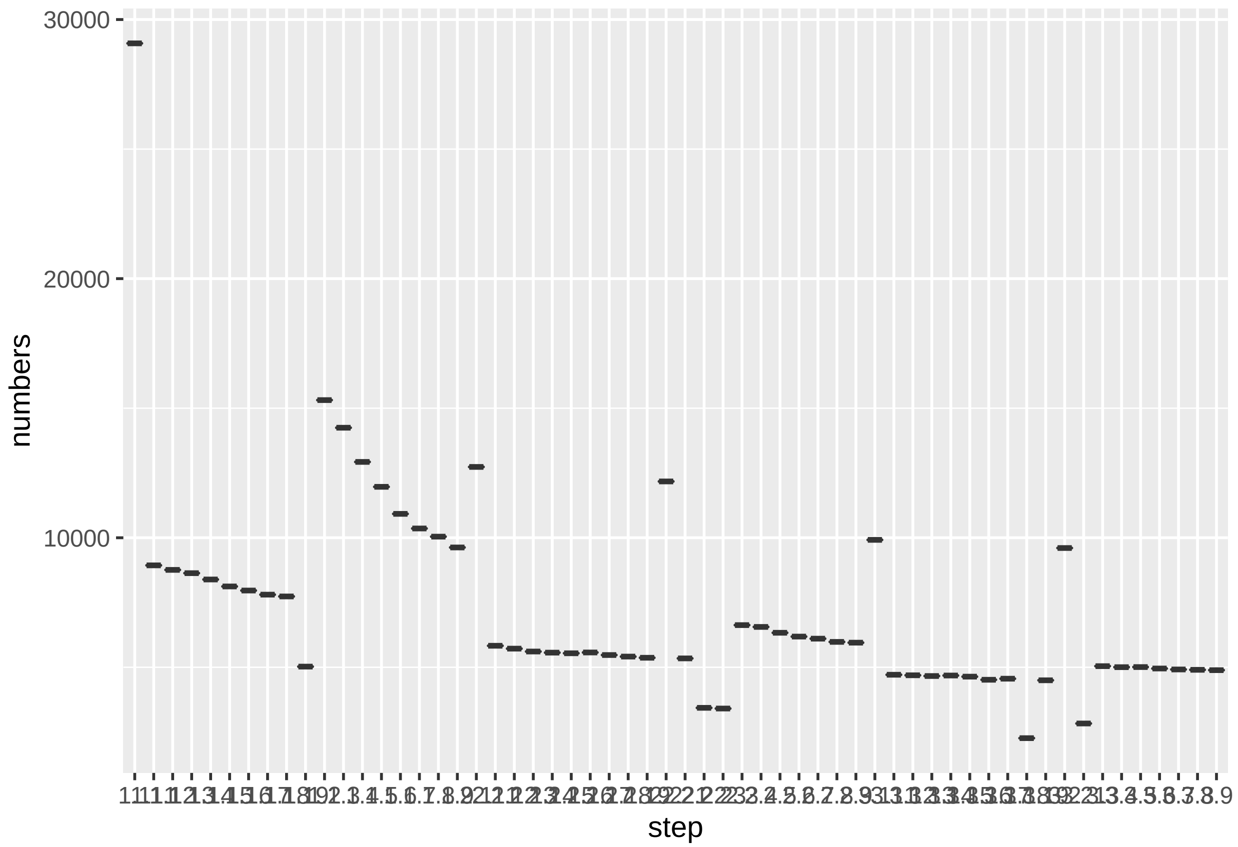

answers_numbers
12000
9000
6000
3000
0
3.18.9
1.7.1
1.7.2
1.7.3
1.7.4
1.7.5
1.7.6
1.8.1
1.8.2
1.8.3
1.8.4
1.8.5
1.8.6
2.19.1
2.20.1
2.8.1
2.8.2
2.8.3
3.11.1
3.11.2
3.11.3
3.18.1
3.18.2
3.18.3
3.18.4
3.18.5
3.18.6
3.18.7
3.18.8
quiz_question
answers_numbers
10000
7500
5000
2500



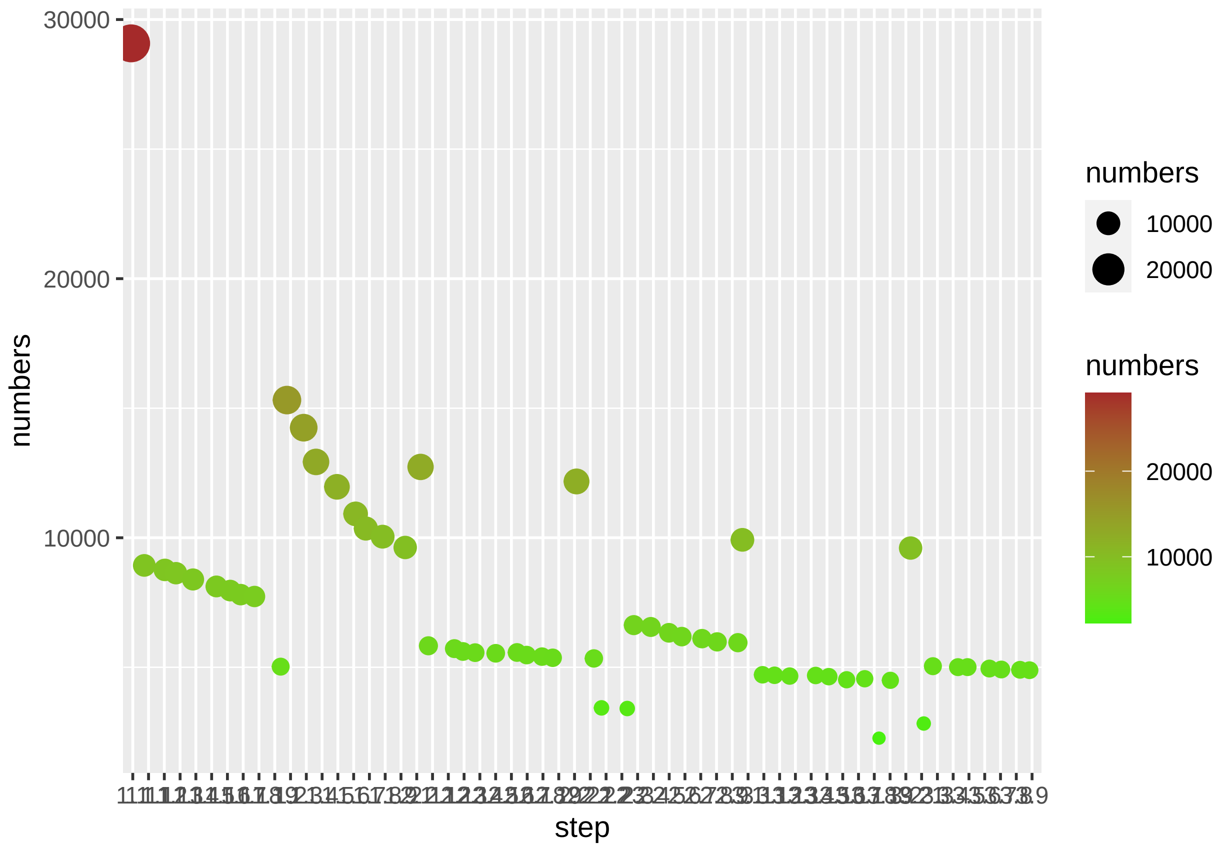
No.people
5000
4000
3000
2000
1000
0
question
3.18.9
1.7.1
1.7.2
1.7.3
1.7.4
1.7.5
1.7.6
1.8.1
1.8.2
1.8.3
1.8.4
1.8.5
1.8.6
2.19.1
2.20.1
2.8.1
2.8.2
2.8.3
3.11.1
3.11.2
3.11.3
3.18.1
3.18.2
3.18.3
3.18.4
3.18.5
3.18.6
3.18.7
3.18.8
No.people
5000
4000
3000
2000



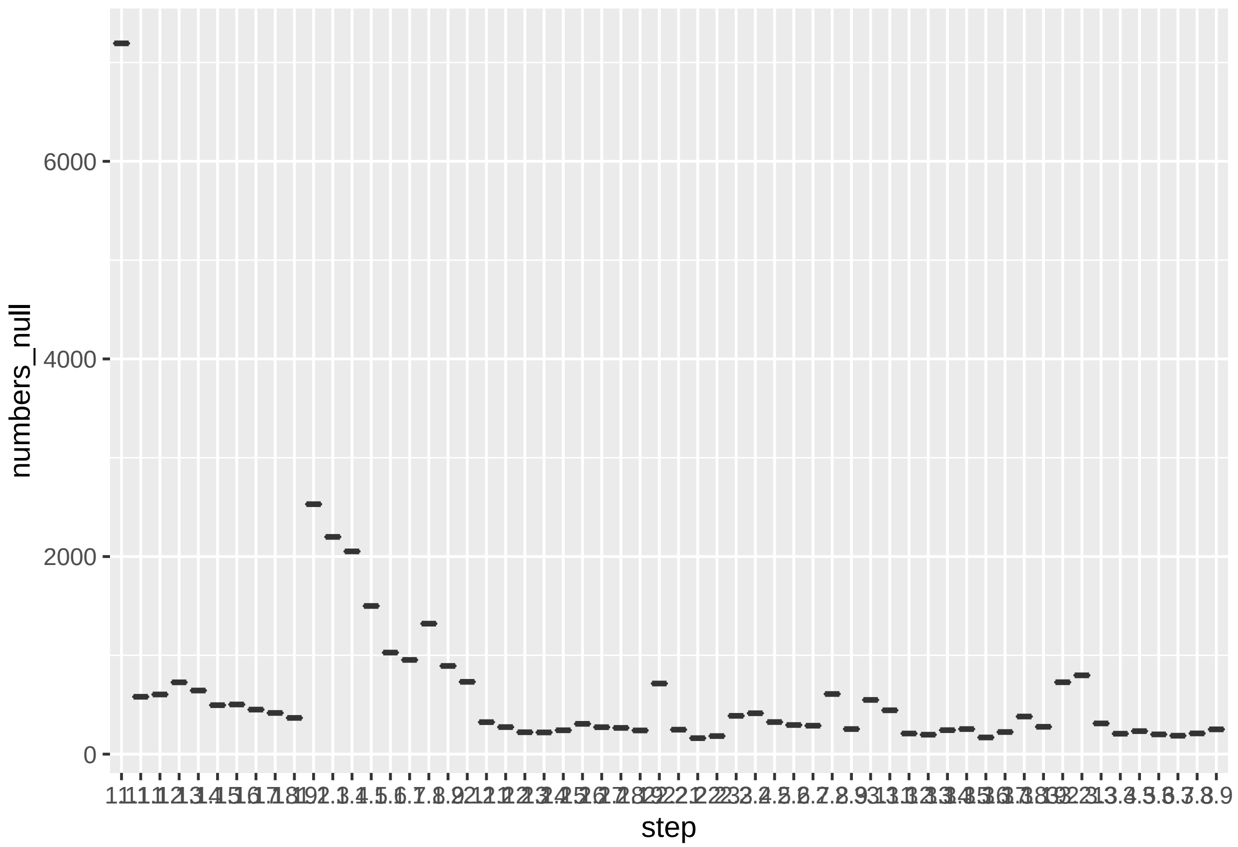
```
ggplot(t_2, aes(x = step, y = numbers)) + geom_boxplot()
```



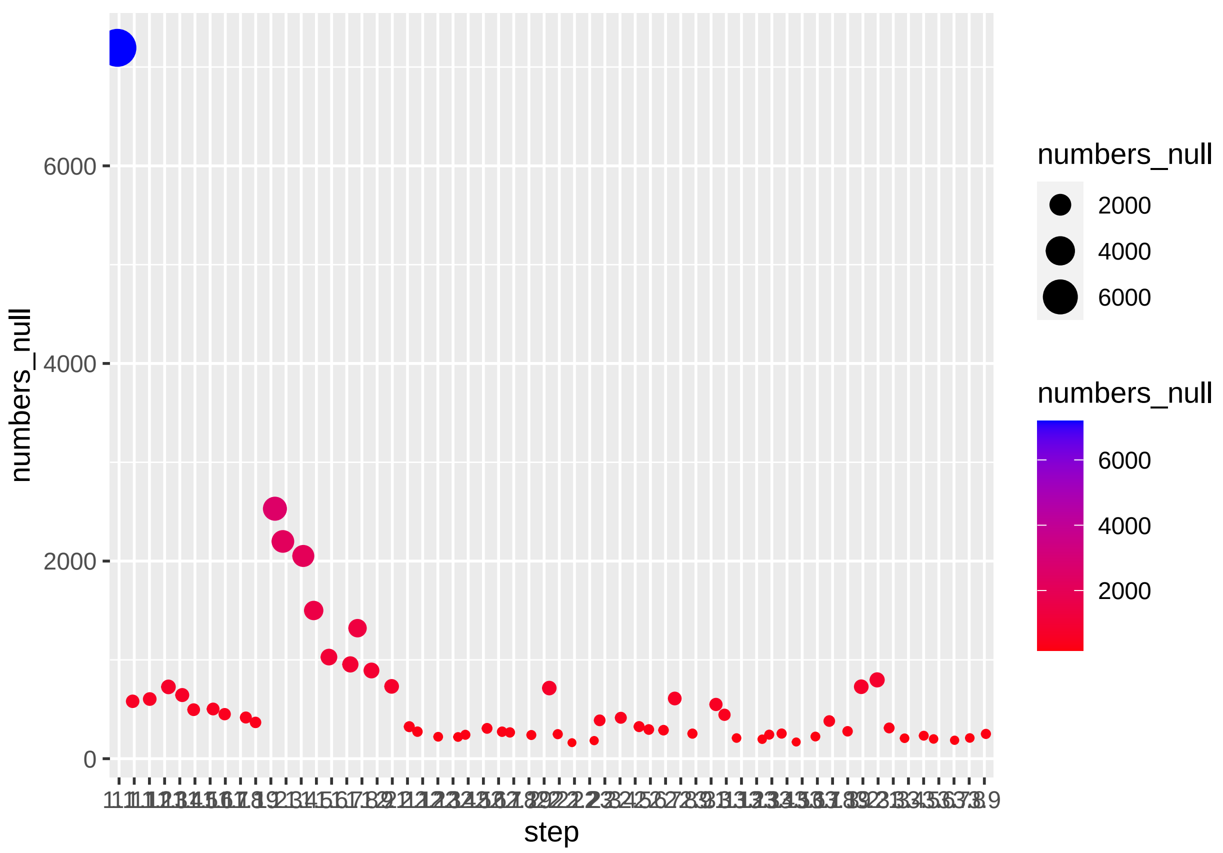
```
ggplot (t_2, aes (x = step, y = numbers, size = numbers, colour =numbers)) +
  # Scatter function: alpha sets the transparency of the scatter.
  geom_point (position = "jitter") +
  # Make the area of the scatter positively proportional to the value of the variable.
  scale_size_area () +scale_color_gradient2(low = "yellow", mid = "green", high = "brown")
```



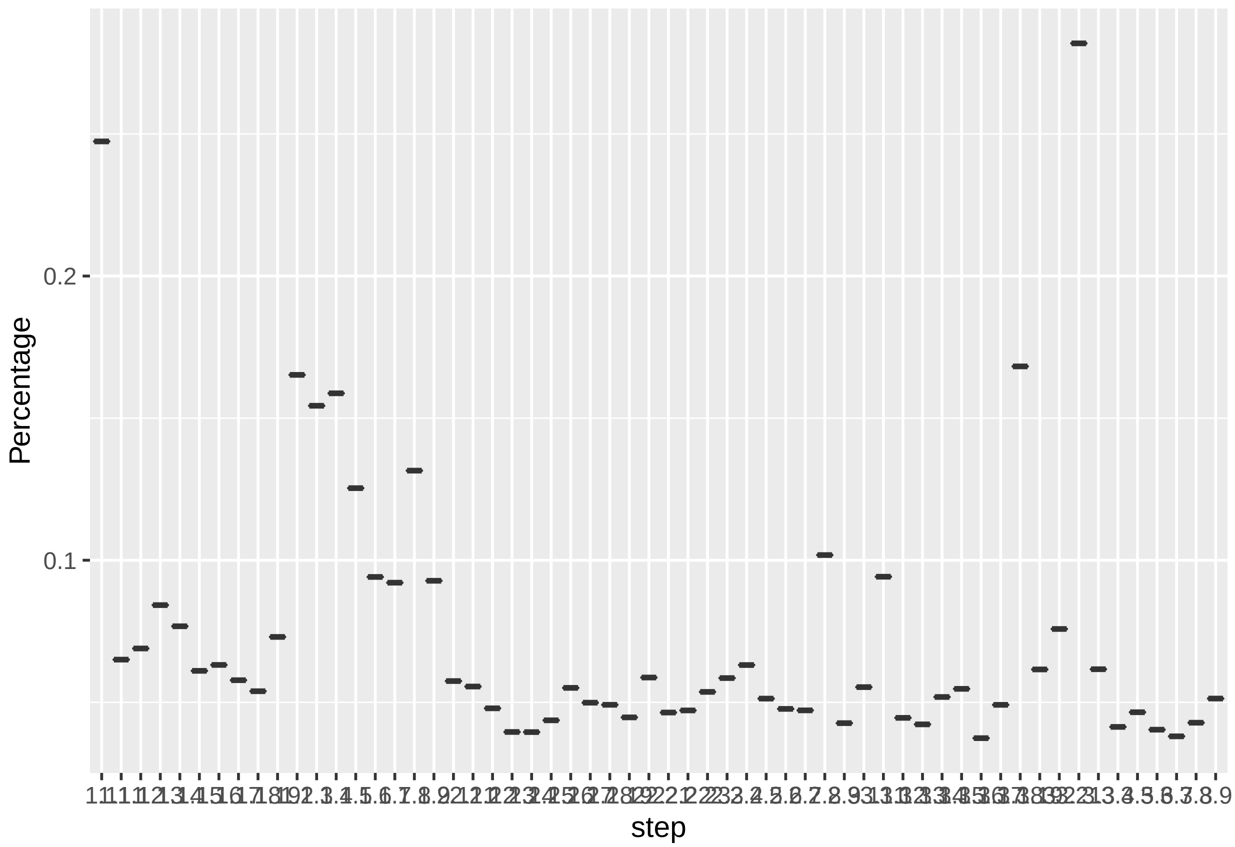
```
ggplot(t_3, aes(x = step, y = numbers_null)) +  geom_boxplot()
```



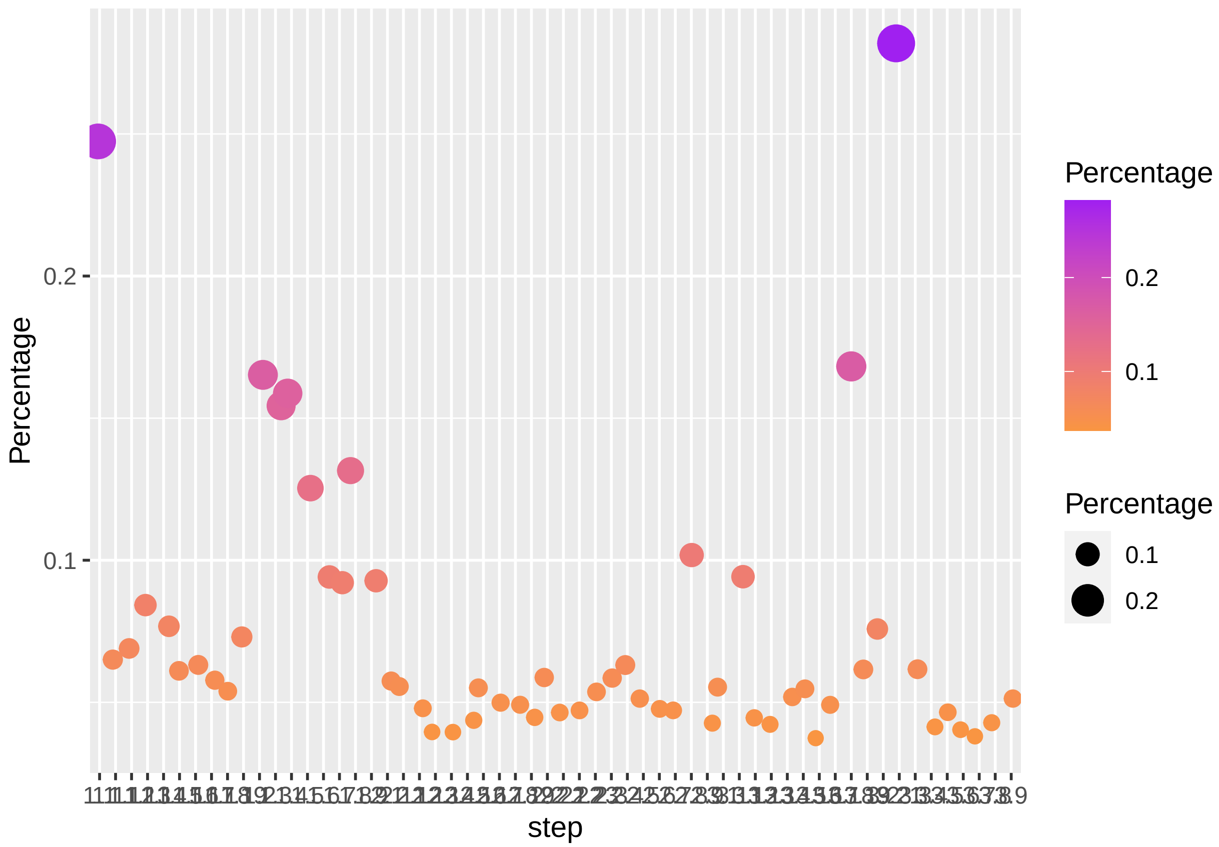
```
ggplot (t_3, aes (x = step, y = numbers_null, size = numbers_null, colour =numbers_null)) +
  #  Scatter plot function: alpha sets the scatter transparency
  geom_point (position = "jitter") +
  # Make the area of the scatter positively proportional to the value of the variable
  scale_size_area () +scale_color_gradient2(low = "pink", mid = "red", high = "blue")
```



```
ggplot(t_4, aes(x = step, y = Percentage)) + geom_boxplot()
```



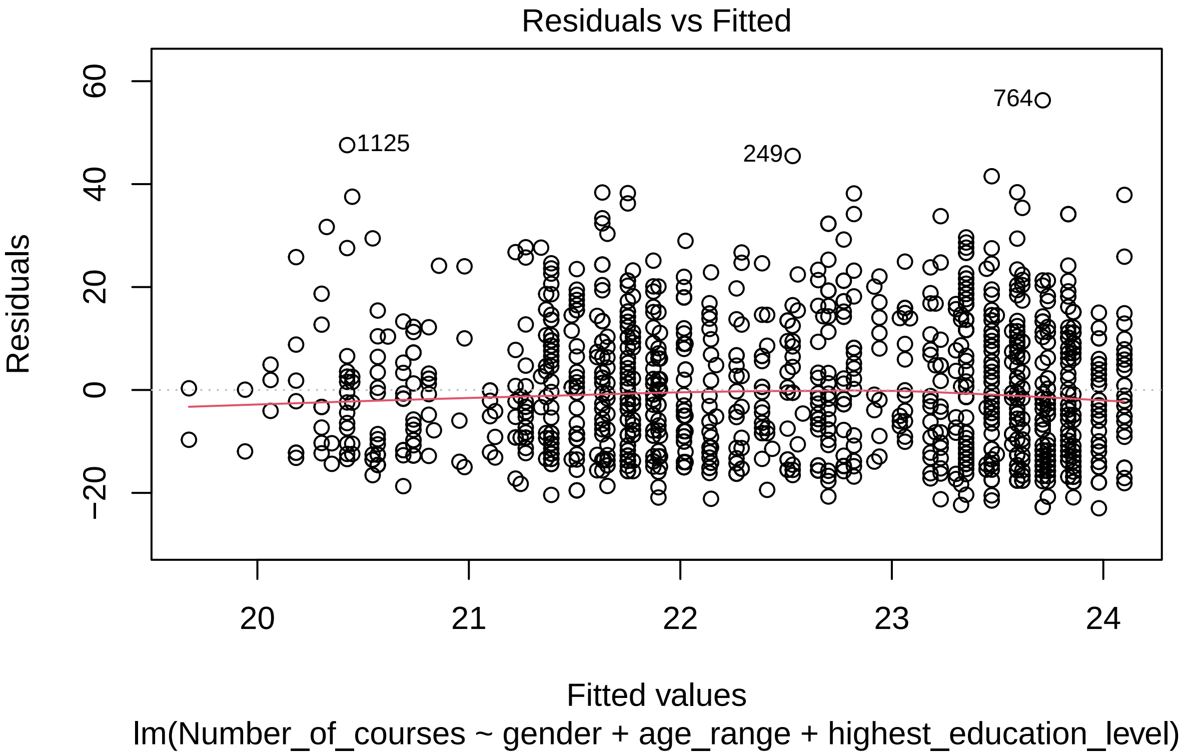
```
ggplot (t_4, aes (x = step, y = Percentage, size = Percentage, colour =Percentage)) +
  # Scatter plot function: alpha sets the scatter transparency
  geom_point (position = "jitter") +
  # Make the area of the scatter positively proportional to the value of the variable
  scale_size_area () +scale_color_gradient2(low = "blue", mid = "orange", high = "purple")
```

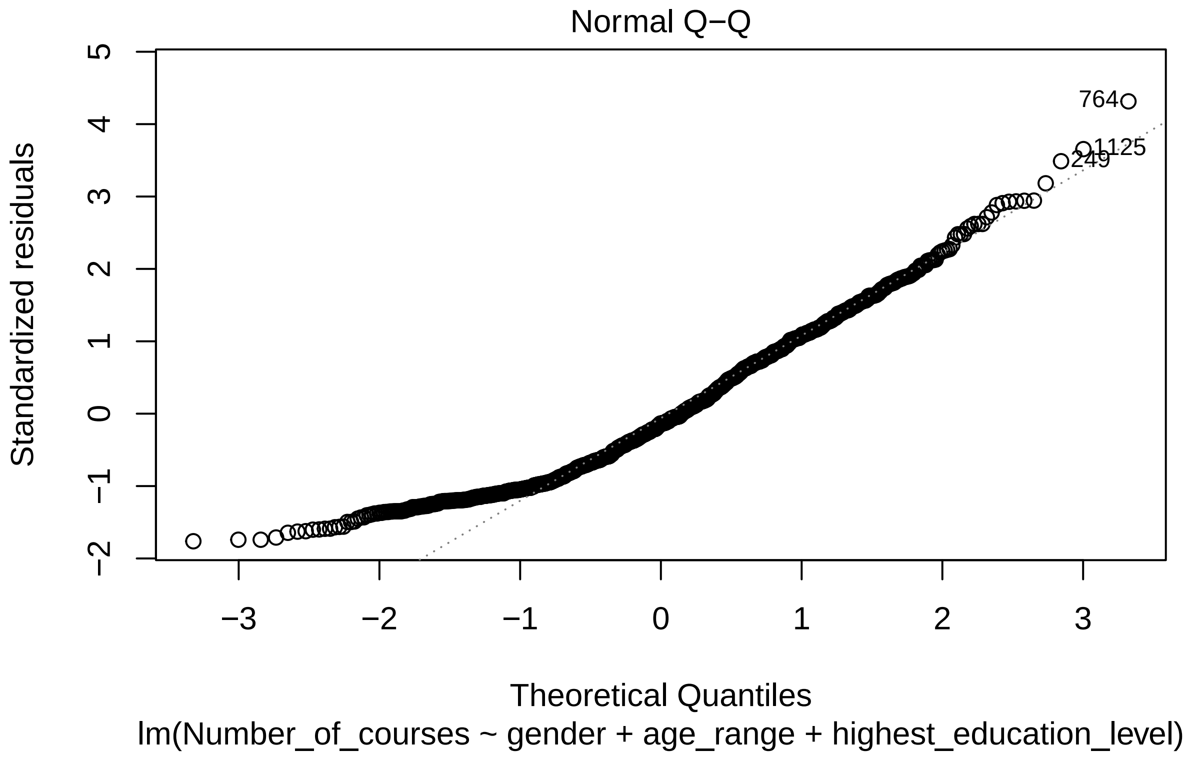


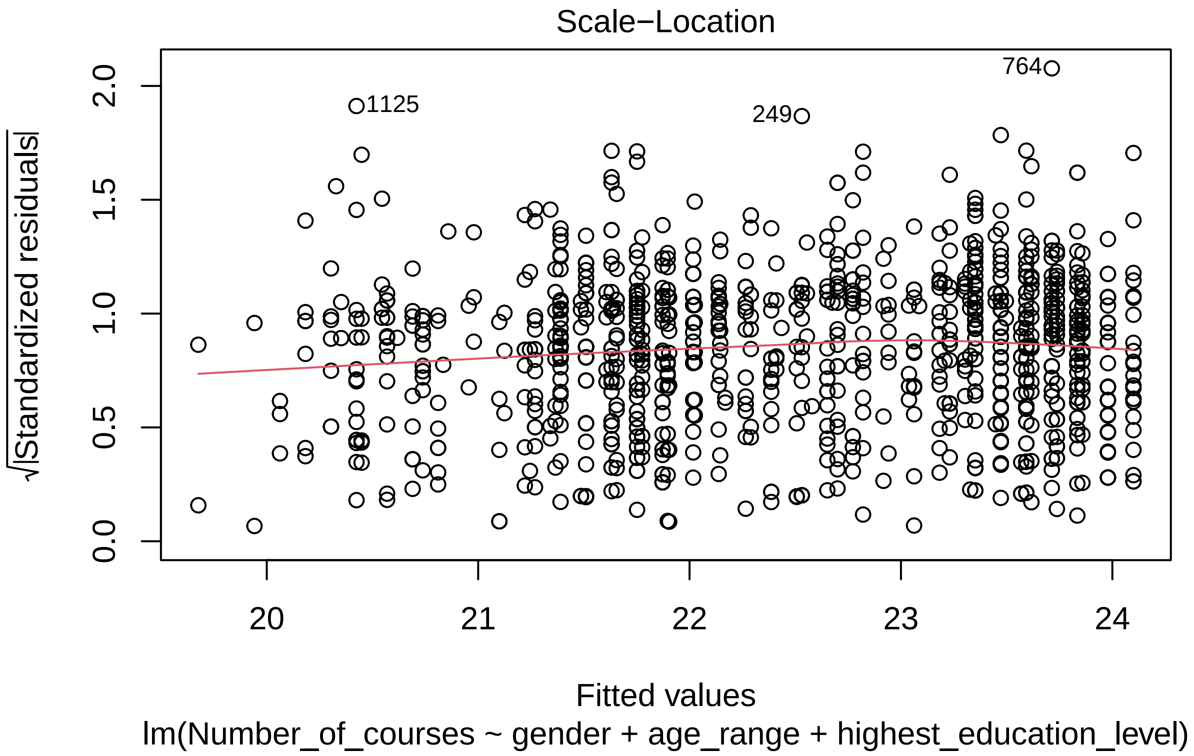
```
fit.full=lm(Number_of_courses~gender+age_range+highest_education_level,data=data1)
summary(fit.full)
```

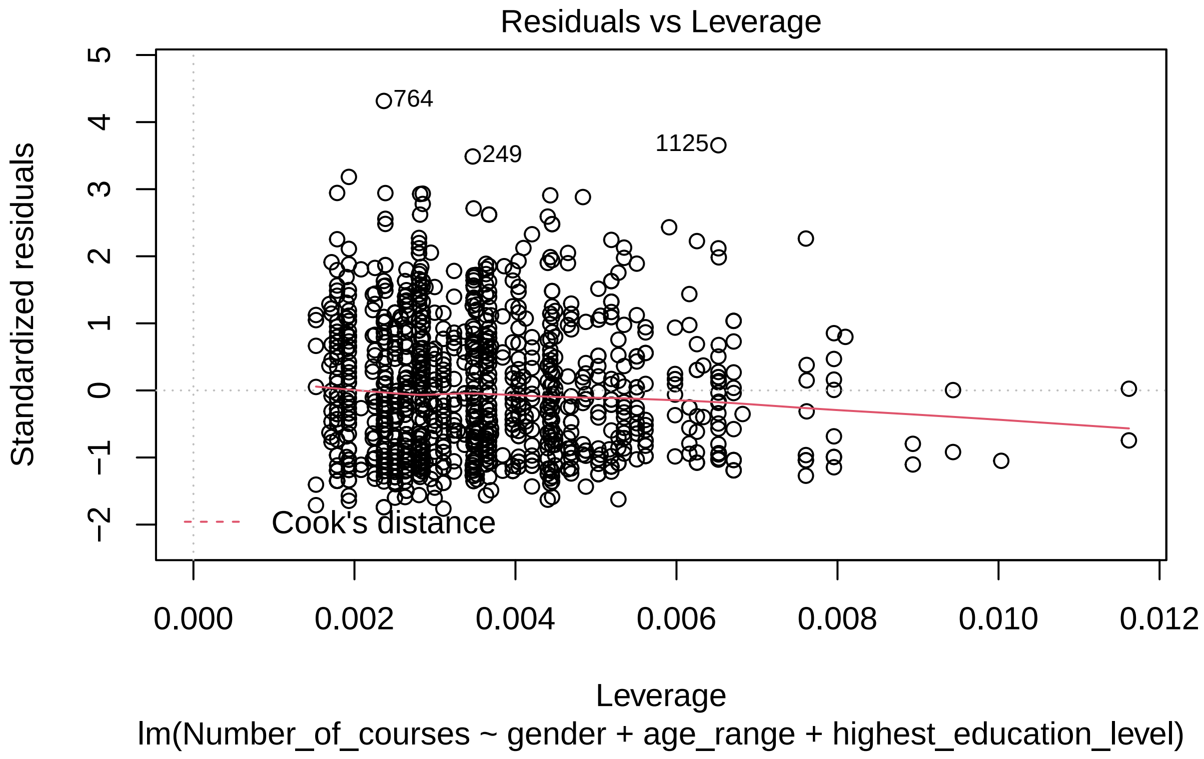
```
##
## Call:
## lm(formula = Number_of_courses ~ gender + age_range + highest_education_level,
##     data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.979 -10.840  -1.773   9.234  56.286
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              23.6412     1.2007  19.689   <2e-16 ***
## gender                   -1.9619     0.7979  -2.459   0.0141 *
## age_range                 0.1207     0.2489   0.485   0.6278
## highest_education_level  -0.2655     0.1972  -1.346   0.1785
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 13.06 on 1116 degrees of freedom
## Multiple R-squared:  0.006769,   Adjusted R-squared:  0.004099
## F-statistic: 2.535 on 3 and 1116 DF,  p-value: 0.05546
```

```
plot(fit.full)
```

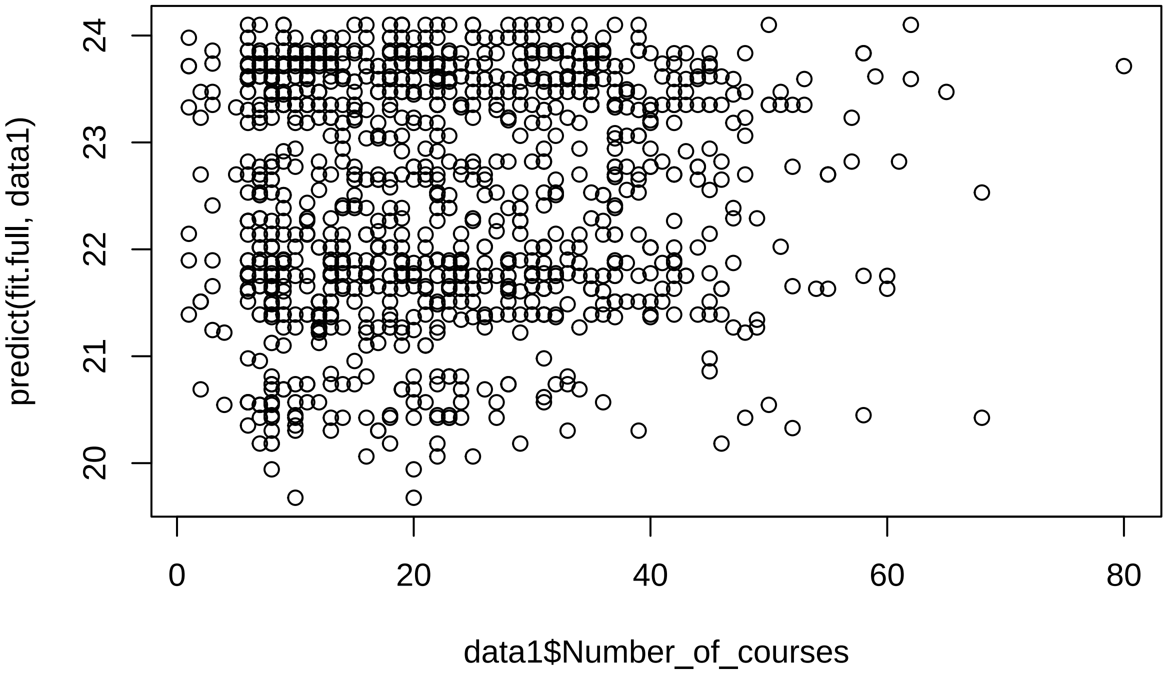



Normal Q−Q
Standardized residuals
5
4
3
2
1
0
−1
−2
764
1125
249
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Number_of_courses ~ gender + age_range + highest_education_level)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1125
249
764
20
21
22
23
24
Fitted values
lm(Number_of_courses ~ gender + age_range + highest_education_level)



```
plot(data1$Number_of_courses,predict(fit.full,data1))
```


predict(fit.full, data1)
20
21
22
23
24
0
20
40
60
80
data1$Number_of_courses