CSC8631 Report

Yimiao Wang

11/25/2021

## Data Merge and Cleaning

Use the rbind() function to merge the data sets into a table according to different names, and use the duplicated function to delete the duplicate values in the data, and use the na.omit() function to delete rows with NA values in the data. For archetype Survey responses and enrolment data delete rows with duplicate values of learner id.

## Exploratory data analysis

### Enrolments

First use the 7 tables of enrolments for data visualization analysis,load the dplyr package,delete the rows with unknown and “-” in the data set,and extract the gender,age\_range,country,highest\_education,employment\_status,and employee\_area in the table,respectively use data visualization processing and analyze the size of variables between different students.

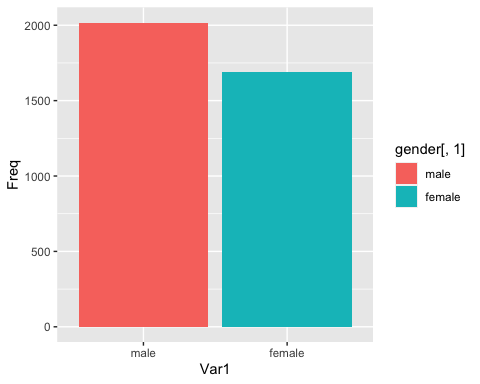
##   
## 载入程辑包：'dplyr'

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

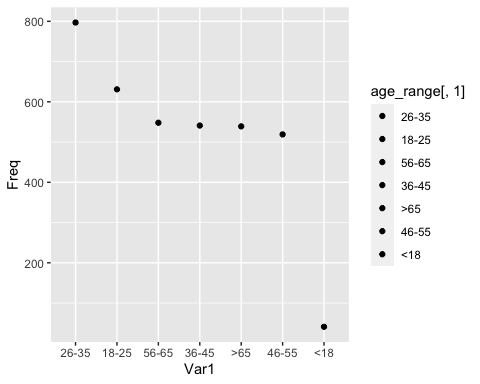
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

From the gender variable, it can be seen that in all enrolments data, there are more males than females. Draw the composition of different age groups. It can be seen from the figure that most students are between 26 to 35 years old, and With the increase of age, the number of students decreases to a certain extent. It is worth noting that the number of people younger than 18 years old is at least lower than that of middle school. Therefore, it can be clearly seen from the educational distribution chart that the highest degree of most students is a bachelor degree, and the second most number is a master degree. Among all scholars, the lowest degree is lower than that of middle school, and the number of college degree is the least.

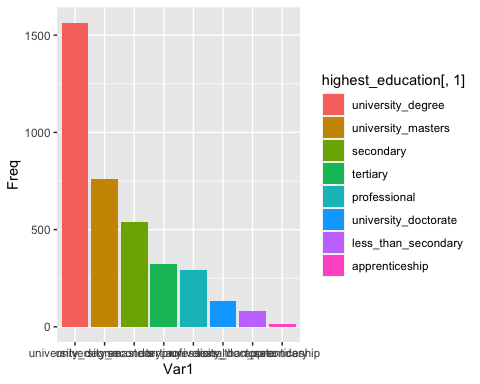
library(ggplot2)  
library(fmsb)  
#Visualize gender,agerange,education,detected country,employment  
ggplot(gender,aes(x=Var1,y=Freq,fill=gender[,1],group=factor(1)))+geom\_bar(stat="identity")



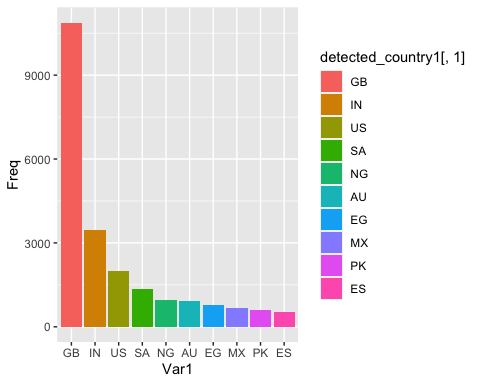
ggplot(age\_range,aes(x=Var1,y=Freq,fill=age\_range[,1],group=factor(1)))+geom\_point(stat="identity")



ggplot(highest\_education,aes(x=Var1,y=Freq,fill=highest\_education[,1],group=factor(1)))+geom\_bar(stat="identity")



Since the scholars come from different countries and regions, a histogram is made to check the distribution of the countries. From the figure, it can be seen that most of the scholars are from the United Kingdom, and the number of other countries is significantly smaller. The number of people from the three countries of MX, PK, and ES is the least.



There are many professional workers among all learners, so visualize the employment area and analyze the work fields of different learners. From the histogram, it can be seen that the number of people engaged in it and information services accounts for most of the proportion, while those engaged in teaching and education the number is second, and the occupation with the least number is recruitment and pr.analyze the employment status of different scholars, load the fmsb program package, and make a radar chart. From the radar chart, it can be intuitively seen that the largest employment status is full-time employment, of which the second number is retired retirees,the number of unemployed is the least.

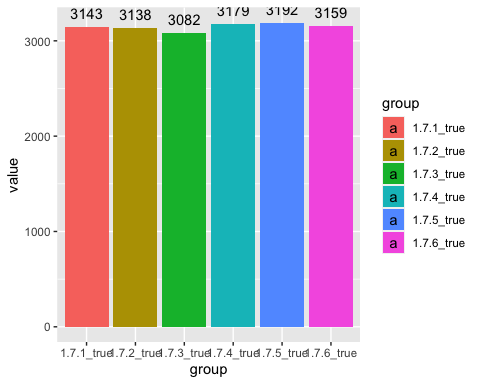
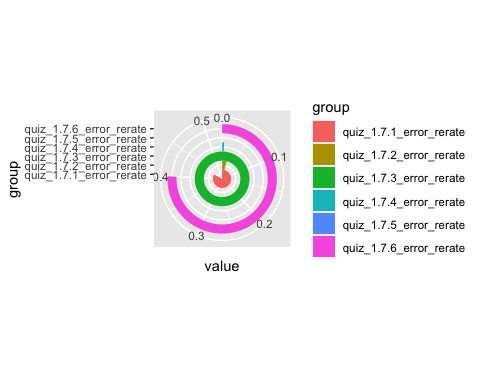
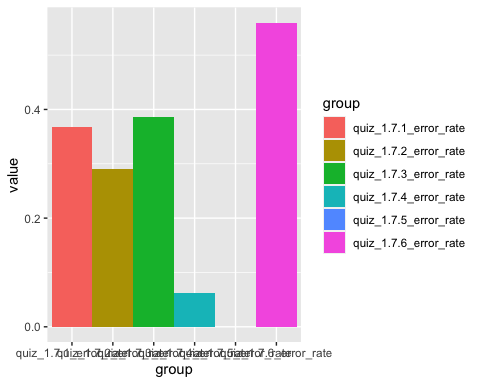
### Step activity

In order to be able to know the time required for students to do a question, use the data in the step activity to calculate the time difference,first delete the data with null values in the last\_completed\_at data set, take question 1.1 as the main analysis object,first filter out all answers to the 1.1 question People, extract the 5th and 6th columns, namely the time of the first visit and the completion time of the student, use the difftime() function to calculate the time difference, where as.POSIXct is used to specify the format of the year, month, day, time and time zone, and the apply() function is used to find the largest time difference With the minimum time difference, the maximum value is 56516879 seconds in line 58, which takes 15 hours to complete, and the minimum value is 1 second. Most of the time difference is between [500,1000], indicating that the problem of 1.1 there are some more difficult questions.Then extract a question in 1.1, take 1.14 as an example, and also calculate the time difference. The maximum value is 41476865 seconds, and the minimum value is also 1 second.

|  |  |  |  |
| --- | --- | --- | --- |
|  | interval | max | min |
| 1.1 | [1000,4000] | 15h | 1s |
| 1.14 | [500,1000] | 14h | 1s |
| :—: | :—: | :—: | :—: |

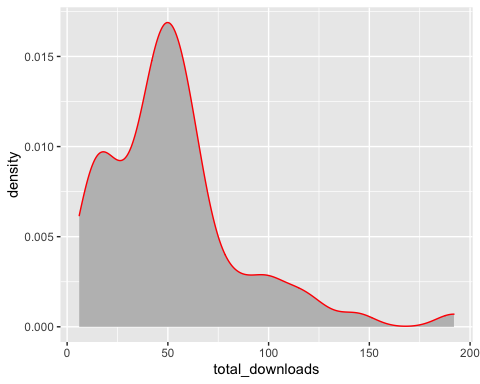
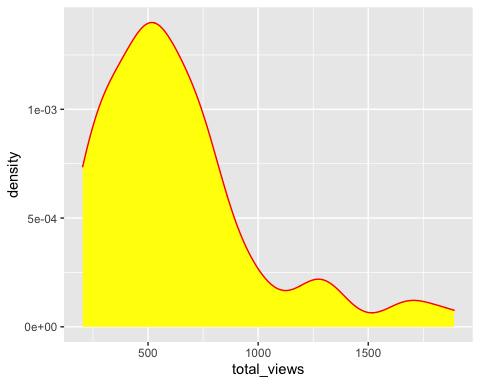
### Question response

In order to get a clearer understanding of students answering,integrating the 7 tables of question response, delete all empty values in the 8th column, and delete rows with NA in the table to extract the number of people who answered 1.7.1-1.7.6. Since there will be duplicate students doing 1.7.1 and the answer is TRUE in the data set, the duplicate value is deleted, and the number of people who answered the question correctly is 3143. Similarly, count the number of people whose answer is wrong, and calculate Calculate the total number of people who answered the question and calculate the total error rate. Among the number of people who answered incorrectly, there are some people who repeatedly answered incorrectly. Use the total number of wrong answers minus the number of wrong answers to get the number of repeated mistakes, and repeat the wrong answers. Divide the number of people by the total number of incorrect answers to get the repeat error rate. Questions with a high repeat error rate indicate that they are more difficult and need to be explained.

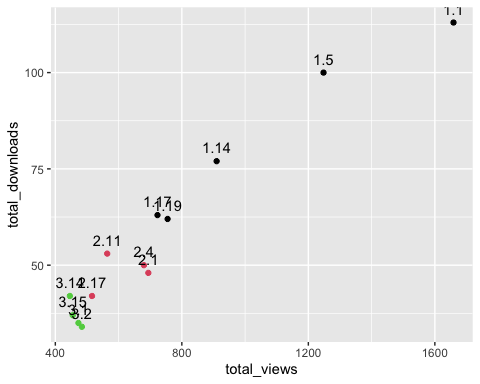
 Therefore, in order to see the total error rate, repeated error rate of each question more intuitively, draw a table, a pie chart, and a histogram. It can be seen that 1.7.5 and 1.7.2 have the lowest repeat error rate, and the number of correct answers is also High, and 1.7.3 is the question with the highest repeat error rate, indicating that it is necessary to focus on the question.

### Video stats

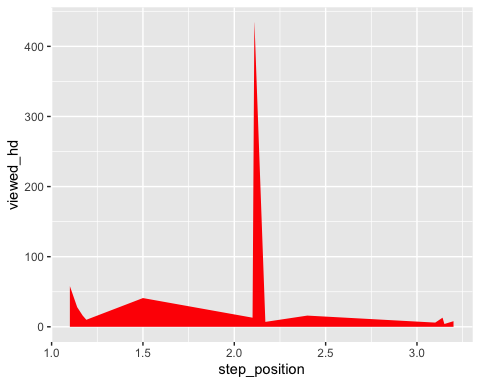
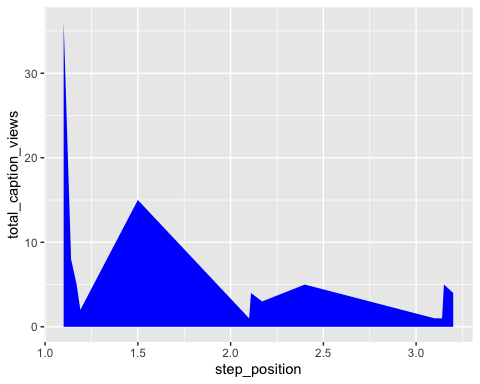
In order to be able to visually see the positive status of students learning, the video status data set is merged. First, the data is sorted, and all columns with 0 are deleted to obtain a data set with 65 rows and 23 columns. Use ggplot to make a density map from the video. It can be seen from the density graph of the total playback volume that the video playback volume of each section has the highest density at 500, indicating that the average number of views is about 500. From the density graph of the number of video downloads, it can be seen that the number of video downloads in each section is About 50 times.



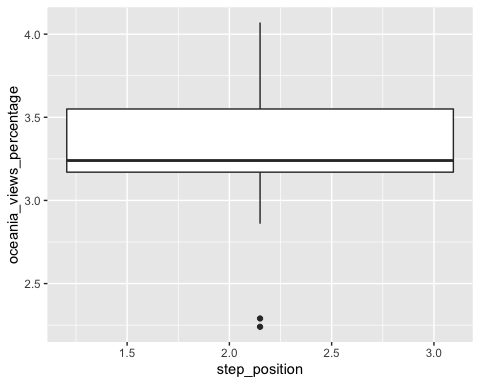
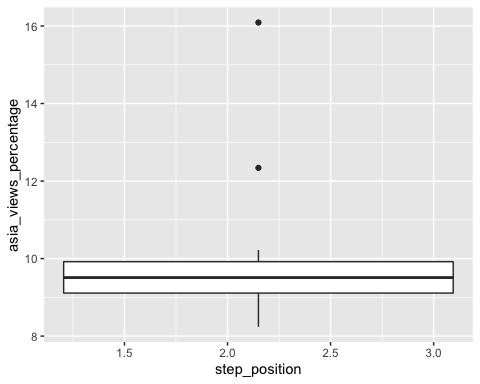
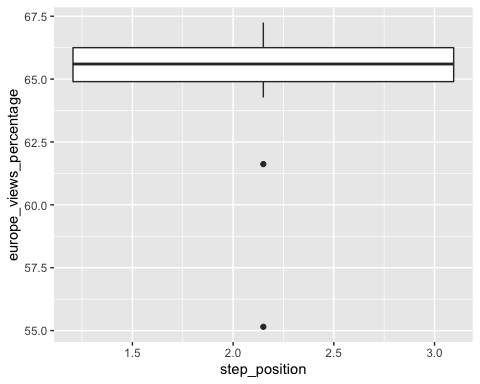
In order to further study the relationship between variables, take rows 1 to 13 that is the data in the first table, load the reshape2 program package to refit the data set, take 1 to 4 columns, and make a scatter plot, from the scatter plot It can be seen that there is a linear relationship between the total number of videos played and the number of downloads. From the scattered point distribution, it can be seen that the most viewed and downloaded chapters of the video are 1.1, followed by 1.5, and the third chapter is generally very small, indicating that students enthusiasm for learning will be reduced later.



Making an area chart for whether learner can understand the video content without adding subtitles. It can be seen from the figure that section 1.1 and section 1.5 have the most views, indicating that these two video contents are very Easy to understand. As for the requirements for high-definition video, it can be seen in the area chart that section 2.11 uses high-definition viewing the most, indicating that the 2.11 video definition needs to be improved.



People who watch the video are in Europe, with a median of about 65.5%, indicating that the average view of people in Europe is 65.5%. The proportion of viewing in Asia and Oceania has decreased significantly, with the medians being 9.75% and 3.25% respectively.



## Model establishment

### video stats least squares regression model

To model the video stats, first do the principal component analysis on the data, delete the data in the second column and columns 15 to 23, select the data in rows 14 to 26 for principal component analysis, first calculate the correlation between the variables, It can be seen that there is a strong correlation between the first three variables. Use the var function to obtain the sample covariance matrix and extract the diagonal elements. The total coefficient of variation is calculated to be 181379.7.

##   
## 载入程辑包：'MASS'

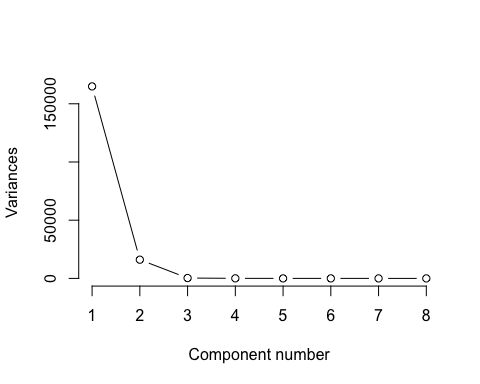
## The following object is masked from 'package:dplyr':  
##   
## select

## total\_views total\_downloads total\_caption\_views  
## total\_views 1.00000000 0.98642400 0.90324840  
## total\_downloads 0.98642400 1.00000000 0.86109753  
## total\_caption\_views 0.90324840 0.86109753 1.00000000  
## total\_transcript\_views 0.94292177 0.94308561 0.77333149  
## viewed\_hd -0.02184285 -0.01030225 0.08034915  
## viewed\_five\_percent 0.72336421 0.68836373 0.59368574  
## viewed\_ten\_percent 0.67096273 0.61282545 0.59109243  
## viewed\_twentyfive\_percent 0.28672817 0.19681943 0.38300383  
## total\_transcript\_views viewed\_hd  
## total\_views 0.94292177 -0.02184285  
## total\_downloads 0.94308561 -0.01030225  
## total\_caption\_views 0.77333149 0.08034915  
## total\_transcript\_views 1.00000000 -0.02963948  
## viewed\_hd -0.02963948 1.00000000  
## viewed\_five\_percent 0.58473554 -0.34077028  
## viewed\_ten\_percent 0.52648630 -0.50038967  
## viewed\_twentyfive\_percent 0.08773656 -0.59430671  
## viewed\_five\_percent viewed\_ten\_percent  
## total\_views 0.7233642 0.6709627  
## total\_downloads 0.6883637 0.6128255  
## total\_caption\_views 0.5936857 0.5910924  
## total\_transcript\_views 0.5847355 0.5264863  
## viewed\_hd -0.3407703 -0.5003897  
## viewed\_five\_percent 1.0000000 0.9316016  
## viewed\_ten\_percent 0.9316016 1.0000000  
## viewed\_twentyfive\_percent 0.6648653 0.8424012  
## viewed\_twentyfive\_percent  
## total\_views 0.28672817  
## total\_downloads 0.19681943  
## total\_caption\_views 0.38300383  
## total\_transcript\_views 0.08773656  
## viewed\_hd -0.59430671  
## viewed\_five\_percent 0.66486535  
## viewed\_ten\_percent 0.84240118  
## viewed\_twentyfive\_percent 1.00000000

## total\_views total\_downloads total\_caption\_views   
## 1.605838e+05 1.829077e+03 8.907692e+01   
## total\_transcript\_views viewed\_hd viewed\_five\_percent   
## 2.776692e+03 1.607273e+04 7.525924e+00   
## viewed\_ten\_percent viewed\_twentyfive\_percent   
## 9.238994e+00 1.150693e+01

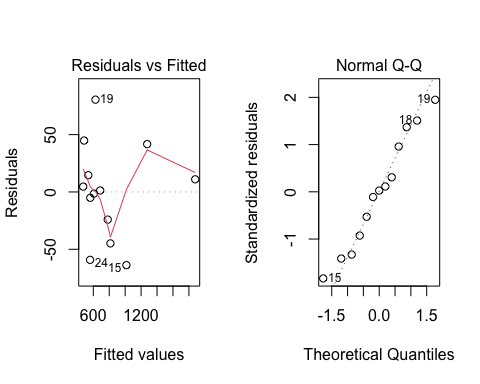
Use the prcomp() function to find the principal components and view the summary. The standard deviation of each principal component exceeds 100, and the cumulative variance contribution rate reaches 90%. It can be seen from the gravel graph that the variance of the first two principal components accounts for most of the total variance, so the first two principal components are selected. The ingredients are appropriate.

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 406.1134 126.77751 17.79552 6.98204 3.21674 1.59953  
## Proportion of Variance 0.9093 0.08861 0.00175 0.00027 0.00006 0.00001  
## Cumulative Proportion 0.9093 0.99791 0.99966 0.99993 0.99998 1.00000  
## PC7 PC8  
## Standard deviation 0.7727 0.2566  
## Proportion of Variance 0.0000 0.0000  
## Cumulative Proportion 1.0000 1.0000



The next step is to fit the least squares regression model. When doing correlation analysis, it is found that there is no correlation between the following variables and the total amount of video playback. Therefore, columns 3 to 6 of the data set are selected as independent variable, total viewed as the dependent variable, check the model summary and found that the first two variables are significant, that is, the total number of videos played has a significant relationship with the number of downloads and the number of views, draw the residual diagram, and the distribution of the residual diagram on the left It is rather scattered, and most of them are randomly distributed around 0.

##   
## Call:  
## lm(formula = y ~ ., data = video\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -63.853 -24.106 1.286 14.605 80.508   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 792.85 14.14 56.064 1.14e-11 \*\*\*  
## total\_downloads 235.61 56.82 4.147 0.00322 \*\*   
## total\_caption\_views 99.36 30.16 3.294 0.01095 \*   
## total\_transcript\_views 78.46 45.48 1.725 0.12279   
## viewed\_hd -11.98 14.96 -0.801 0.44610   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 50.99 on 8 degrees of freedom  
## Multiple R-squared: 0.9892, Adjusted R-squared: 0.9838   
## F-statistic: 183.3 on 4 and 8 DF, p-value: 6.728e-08



Therefore, it can be considered that the total views volume of the video has a linear relationship with the number of downloads and the viewing time. The right panel is a quantile graph to evaluate the residuals. The normality of, most of the points are close to the unit diagonal, indicating that the residual of the linear regression model is normal.

### Question response logistic model

Perform model fitting on the question response data set. Since the data set type of the correct column is character, it needs to be converted to an integer type, and replace false with the number 2 and true with the number 1. Delete the data in the eighth column to Take the number of people who answered question 1.8.1 as an example, set a random seed, divide the data set into training set and test set according to 8:2, and delete the comma in the response column, use 1 for false, 0 for true, and correct as the response Variables and response columns are used as predictors to test the correlation between variables, fit a logistic model, and make predictions. The final test error is 0, indicating that the model has a better predictive effect.It can be seen from the figure that when the student’s answer is 123, it can be counted as correct, and all other answers are counted as wrong.

## [1] 2

## [1] 1

##   
## 0 1   
## 1104 824

##   
## Call:  
## glm(formula = correct ~ t, data = quiz\_train1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.065453 -0.003284 -0.003284 -0.003284 0.126086   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.074e+00 8.211e-04 1308.2 <2e-16 \*\*\*  
## t -8.706e-03 8.764e-06 -993.4 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.001892997)  
##   
## Null deviance: 1882.836 on 7710 degrees of freedom  
## Residual deviance: 14.593 on 7709 degrees of freedom  
## AIC: -26458  
##   
## Number of Fisher Scoring iterations: 2

##   
## 载入程辑包：'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

## 载入需要的程辑包：carData

##   
## 载入程辑包：'car'

## The following object is masked from 'package:psych':  
##   
## logit

## The following object is masked from 'package:dplyr':  
##   
## recode

## Call:corr.test(x = quiz\_train1[, 4:5])  
## Correlation matrix   
## correct t  
## correct 1 -1  
## t -1 1  
## Sample Size   
## [1] 7711  
## Probability values (Entries above the diagonal are adjusted for multiple tests.)   
## correct t  
## correct 0 0  
## t 0 0  
##   
## To see confidence intervals of the correlations, print with the short=FALSE option

## predict  
## Observed 0 1  
## 0 4444 0  
## 1 0 3267

## [1] 0

## [1] 0

