Data Management and Exploratory Data Analysis
CSC8631 Coursework

Max Piotrowicz

13 November 2019

```
## Project name: Exploration_and_analysis

## Loading project configuration

## Autoloading packages

##  Loading package: reshape2

## Loading required package: reshape2

##  Loading package: plyr

## Loading required package: plyr

##  Loading package: tidyverse

## Loading required package: tidyverse

## -- Attaching packages --------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::arrange()   masks plyr::arrange()
## x purrr::compact()   masks plyr::compact()
## x dplyr::count()     masks plyr::count()
## x dplyr::failwith()  masks plyr::failwith()
## x dplyr::filter()    masks stats::filter()
## x dplyr::id()        masks plyr::id()
## x dplyr::lag()       masks stats::lag()
## x dplyr::mutate()    masks plyr::mutate()
## x dplyr::rename()    masks plyr::rename()
## x dplyr::summarise() masks plyr::summarise()
## x dplyr::summarize() masks plyr::summarize()

##  Loading package: stringr

##  Loading package: lubridate

## Loading required package: lubridate

##
## Attaching package: 'lubridate'

## The following object is masked from 'package:plyr':
##
##     here

## The following object is masked from 'package:base':
##
##     date
```

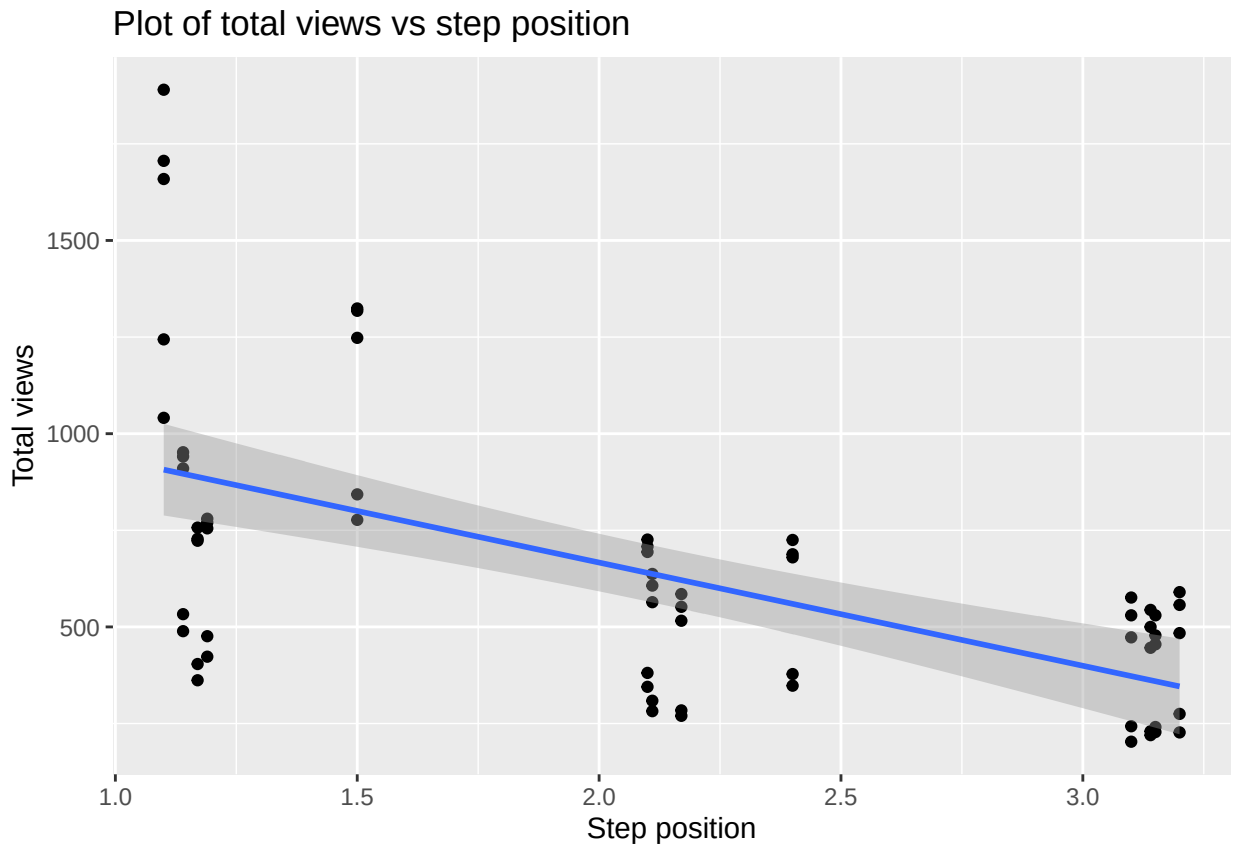
```
##  Loading package: ggplot2
##  Loading package: dplyr
##  Loading package: readr
## Autoloading helper functions
##  Running helper script: globals.R
##  Running helper script: helpers.R
## Autoloading data
##  Loading cached data set: cyber.security.1.question.response
##  Loading cached data set: cyber.security.2.question.response
##  Loading cached data set: cyber.security.3.archetype.survey.responses
##  Loading cached data set: cyber.security.3.question.response
##  Loading cached data set: cyber.security.3.video.stats
##  Loading cached data set: cyber.security.4.archetype.survey.responses
##  Loading cached data set: cyber.security.4.question.response
##  Loading cached data set: cyber.security.4.video.stats
##  Loading cached data set: cyber.security.5.archetype.survey.responses
##  Loading cached data set: cyber.security.5.question.response
##  Loading cached data set: cyber.security.5.video.stats
##  Loading cached data set: cyber.security.6.archetype.survey.responses
##  Loading cached data set: cyber.security.6.question.response
##  Loading cached data set: cyber.security.6.video.stats
##  Loading cached data set: cyber.security.7.archetype.survey.responses
##  Loading cached data set: cyber.security.7.question.response
##  Loading cached data set: cyber.security.7.video.stats
## Munging data
##  Running preprocessing script: 01-A.R
```

For this coursework assignment we were given a dataset relating to a massive open online course. The aim of the courework was to create an exploratory pipeline following the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. CRISP-DM comprises six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. It is an approach to data mining that "loops back", i.e. when the evaluation phase of a given cycle is finished and new insights are acquired (even if these are all negative results, it still provides information after all), we start again with the business understanding phase which allows us to ask some new questions and proceed with a new route of exploration.

Cycle 1

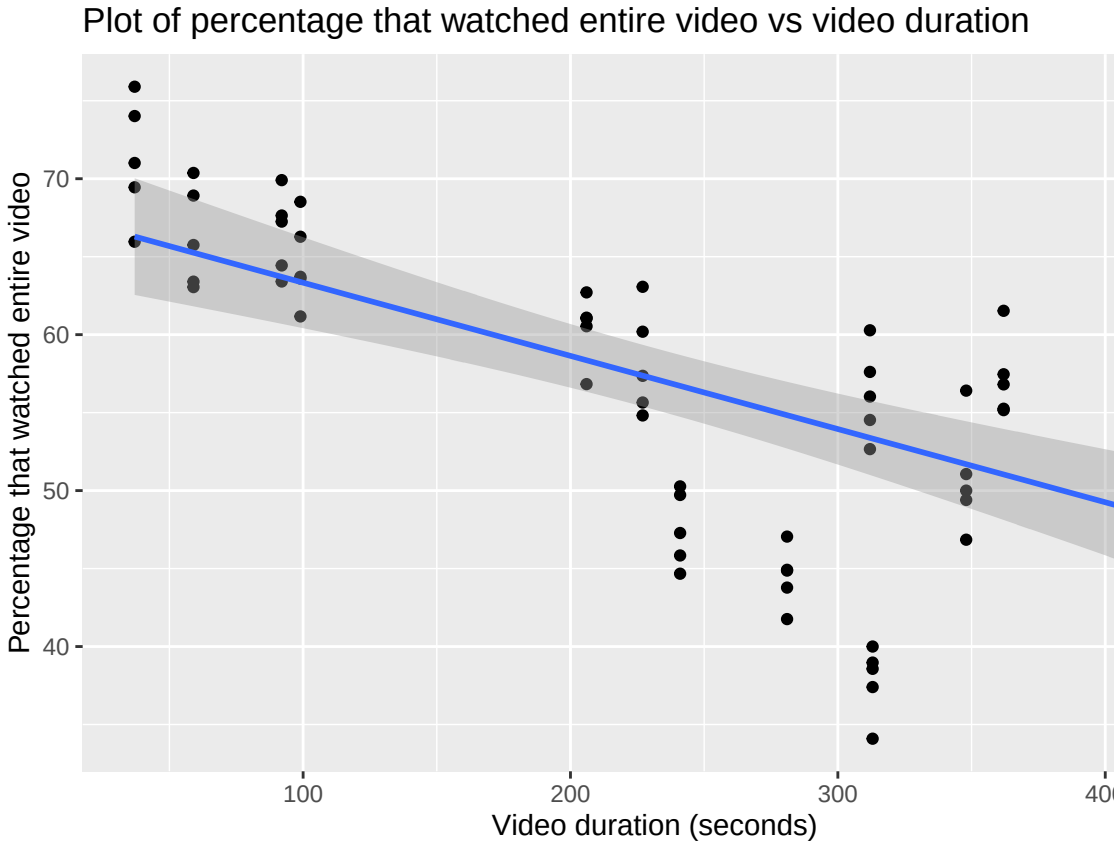
Business understanding:
The data from the course comes from an online course on cyber security over seven runs of the course. Various different datasets are gathered as part of the process of running this course, and it is from this data that an attempt to draw useful information will be made. The people running this course will want to know where they can improve the course. There are certain obvious issues that could be addressed, such as looking

at the data for dropout rates or scoring on test questions. However, by performing some exploratory data analysis there is the potential to discover uknown correlations or tendencies that can provide valuable insight to improving the learning experience.
The first dataset I will look at is the video statistics data. The video data has the potential to be very useful, this is an online course and thus watching the videos is one of the primary modes of learning the course content. If it is found that some of the videos are unpopular or that videos are not being watched at all, it would probably indicate that students are not learning the course material as thoroughly as they could be. Looking at the video duration, total views, as well as the number of viewers that watched a given percentage of the video are the analyses I will focus on. Some simple plots will potentially reveal a wealth of information regarding the success of the video format as pedagogical tool.
Data Understanding:
The data for the video statistics is relatively easy to understand as it anonymised and such rather than having data for each individual student is just contains summary statistics for the entire cohort in question. Although, this is somewhat unfortunate from an exploratory standpoint as it doesn't allow any comparison of video statistics for a given learner to, for example, their test scores. Each section of the course has a video associated with it (e.g. Welcome to Week 2: payment security) and for each video statistics relating to video duration, total views, total downloads, total transcript views, the percentage of students who watched various percentages of the video, the viewing device, as well as the breakdown for different geographical areas that watched the videos. This dataset is only available for the 3rd to 7th run of the course.
Data Preparation:
In this case not a huge amount of data preprocessing was required, as the data is already in a format to allow the analysis that I wanted to perform. In the munging section of the project template library I simply combined the datasets from the 3rd to the 7th run of the course and thus perform my analysis on the combined data across many runs of the course, this will hopefully increase the reliability of any enquires and the associated infernces made from the data. A cursory glance through each of the 5 data-sets allowed me to see that there were no missing values. Modeling:



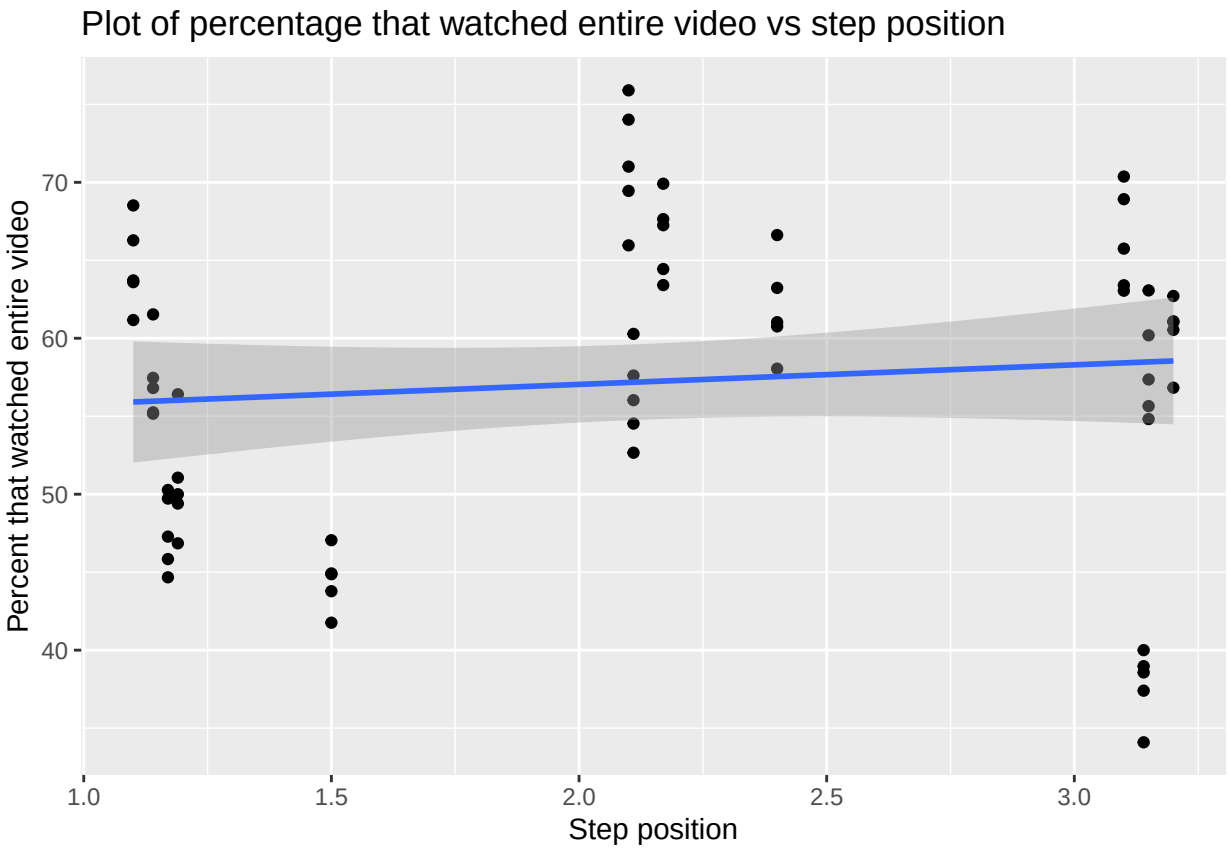
Step progression denotes at what stage of the course the video is at. Clearly, as the course progresses, less and

less people are watching the videos. The correlation coefficient is -0.59, where positive one or negative one would



indicate perfect correlation.

There is a negative correlation between video length and the percentage of people who wacthed the entire video. The correlation coefficient is, coincidentally, -0.59 again. However, there is a clear outlier with the longest video. This video is: The evolving arms race of payment security.



Evaluation: Clearly, as the course progresses, the number of poeple watching the video drastically decreases. This is quite problematic, presumably a large amount of effort was put into creating these videos and if people aren't watching them this is a considerable waste of resources. We would expect the number to decrease somewhat, as people tend to be most motivated at the beginning of a course and some people will drop out. However, from the graph of total views vs step position it is clear that the mean number of people watching the video at the end is approximately half of what is was at the beginning. This is not a very high success rate.
In addition to this, there is a negative correlation between video length and percentage of poeple who have watched the entire video. However, there is an outlier, the longest video, titled: The evolving arms race of payment security. Maybe having an exciting title such as this provides a strong incentive for people to watch the video, and then to continue watching the video till the end.

Deployment: The analysis in this cycle has revealed that video duration and the step at which the video is played has a significant influence on the number of viewers. To address this, the length of video could be decreased. The same amount of material could be covered, but maybe in a format where the length of each video is shorter but with more videos overall.

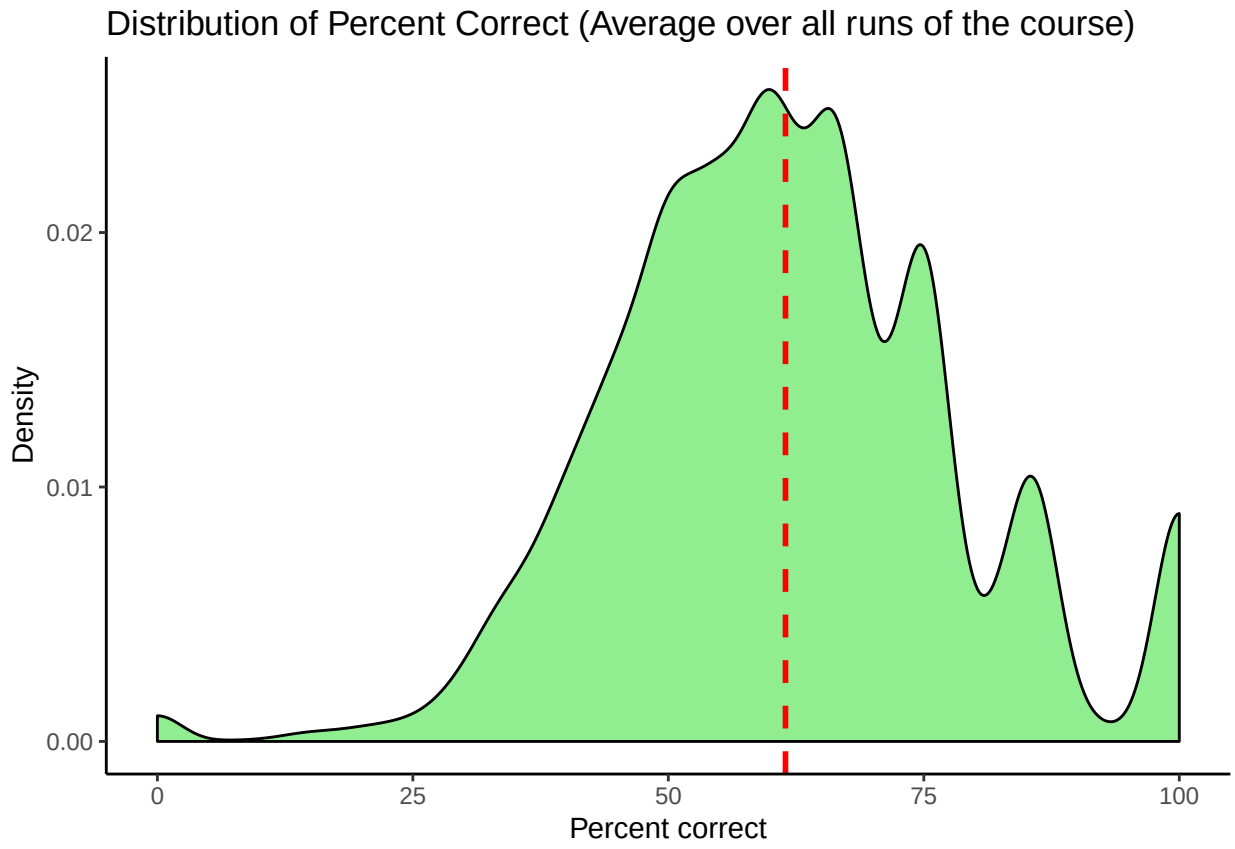
Cycle 2

Business Understanding: Having looked at the video statistics I decided to look at a completely different dataset, namely the question responses. This could be useful to the business to see what the distribution of scores is as well as seeing how many questions are actually answered. For example, if it turns out that many students are not even bothering to answer questions it might be a good idea to completely overhaul the system in which questions to ask, such as making it a compulsory part of the assignment rather than optional. Data understanding:
The data files on question response are available for every single year of the porgram. Each response given by any indivudual to a multiple choice question is given by a true/false value along with the learner identification.

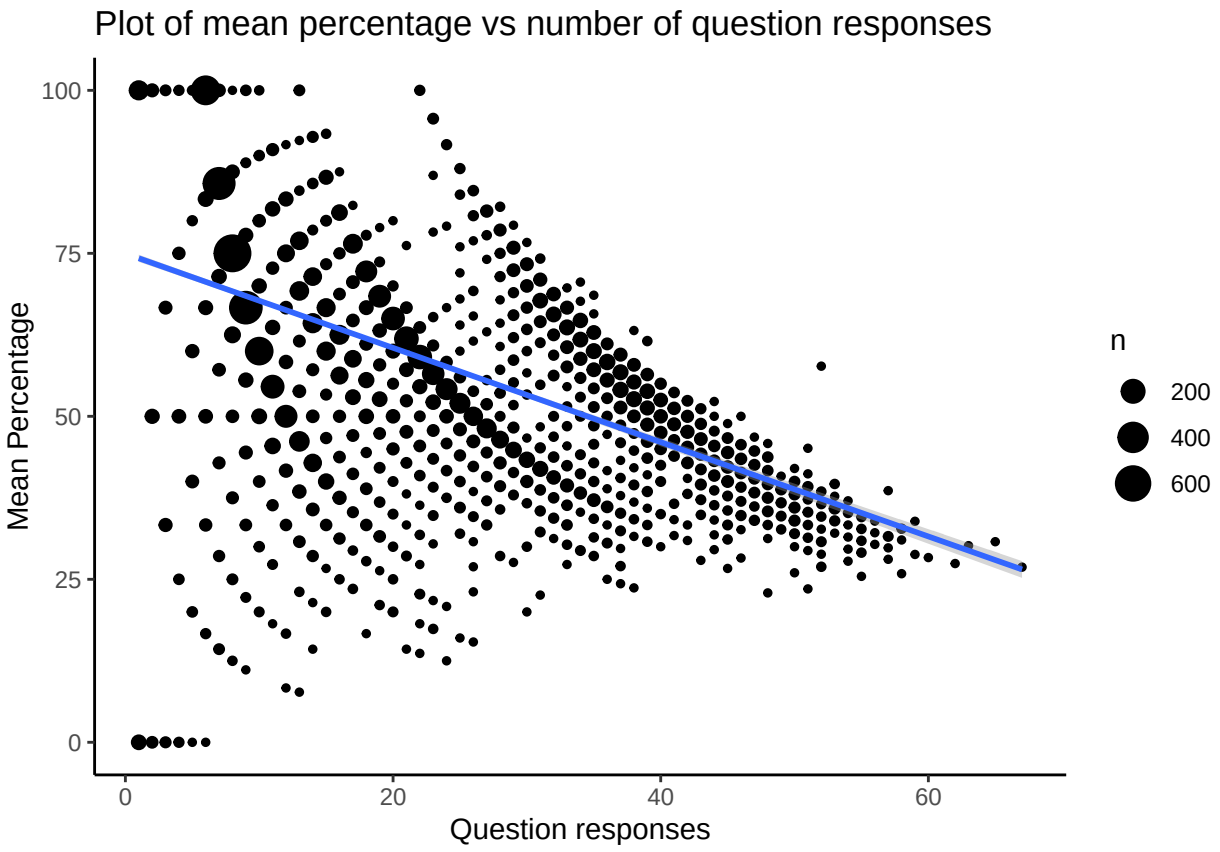
It aslo shows the question number that was answered, in this way it can be seen whether a student attempted a question multiple times. For the purposes of my analysis I will treat multiple attempts of the same question as all counting towards the score of that student (e.g. a student answering only one question but attempting it three times, and only gettting it right on the third attempt would get a score of 33%). All the question types are multiple choice. Data preparation: To prepare the data for use I created a function that creates a dataframe containing the information about each student that I'm interested in.

Modeling: I created another function which counted the number each student answered correct, wrong, percent correct, as well as number answered.

Now that a dataframe containg the relevant information has been created I can continue with the intended analysis.



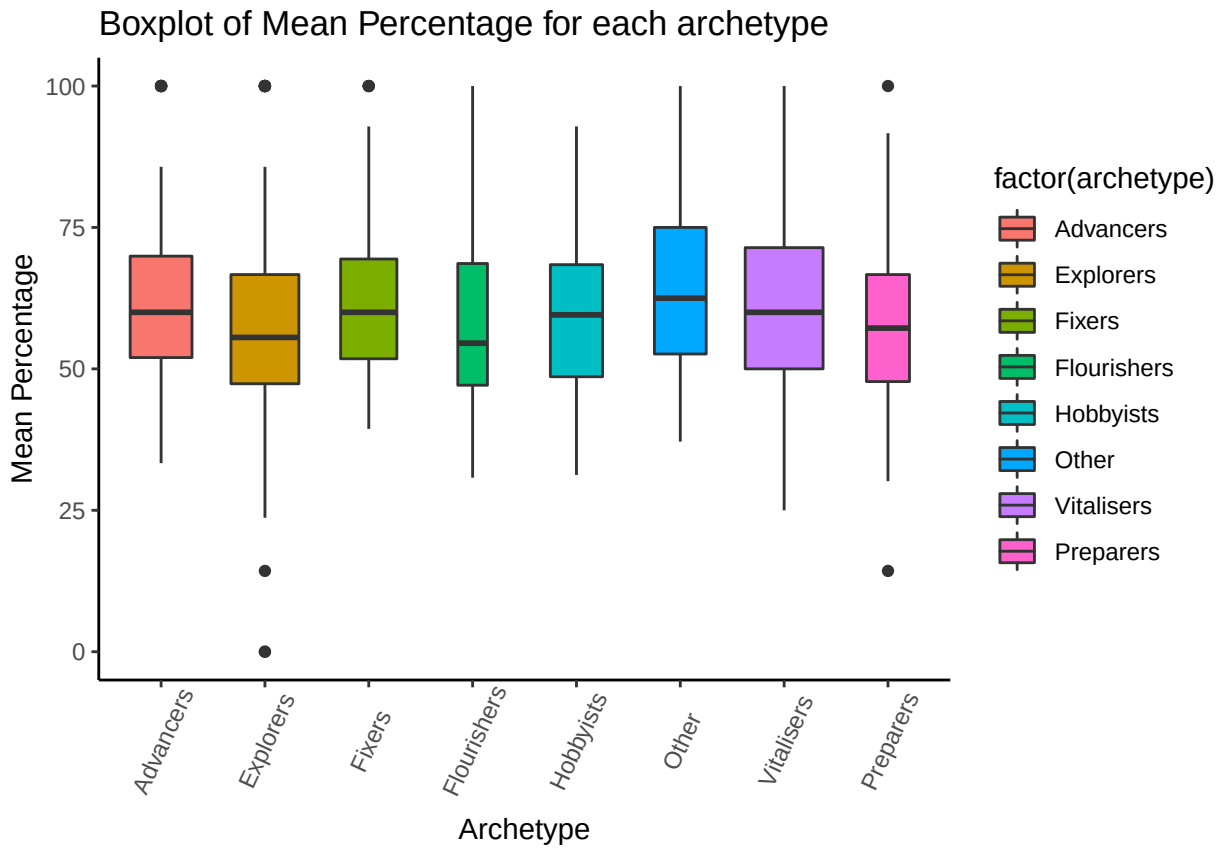
Course Run	Mean	Standard Deviation
1	60.17	16.76
2	63.00	16.74
3	63.36	16.79
4	60.49	17.80
5	61.93	17.54
6	61.81	18.22
7	62.55	18.16



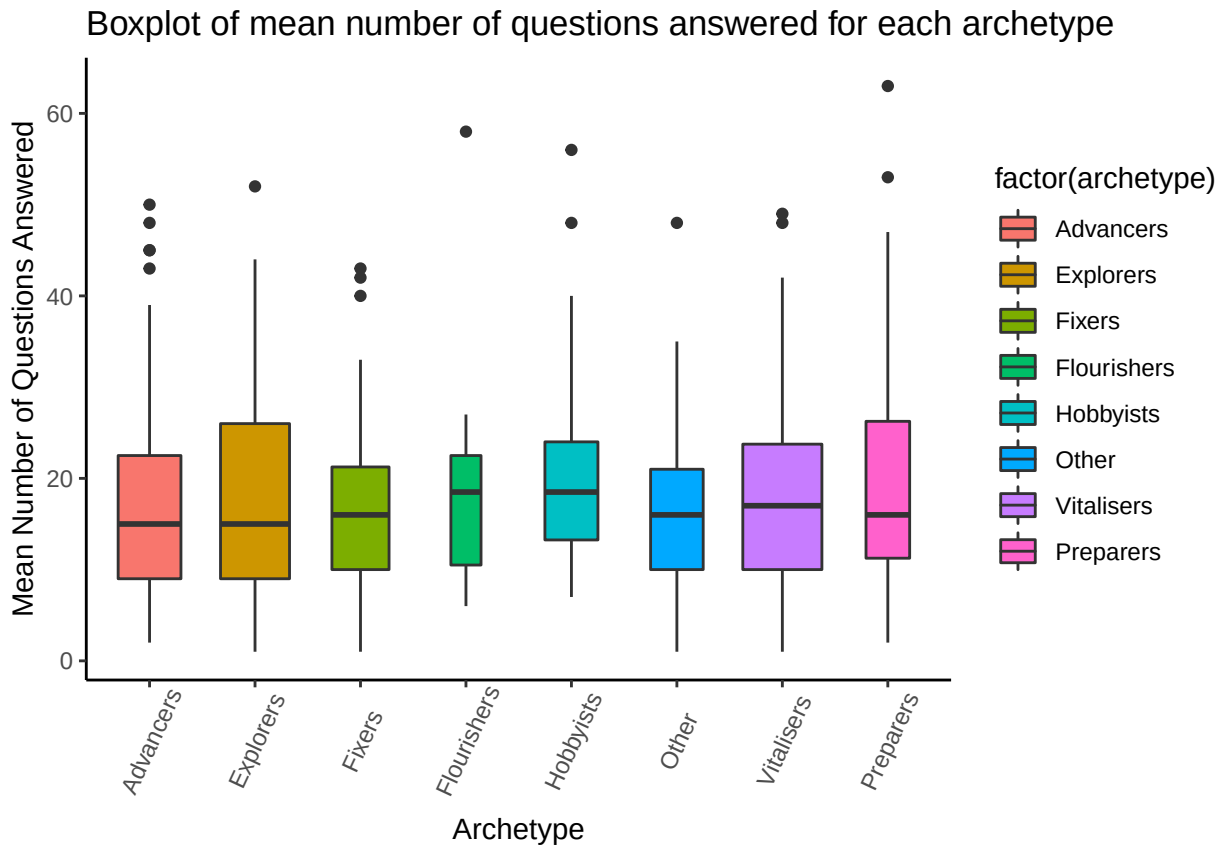

```
## [1] -0.4912035
```

Cycle 3

Business Understanding: Continuing the analysis from cycle two, I wanted to see if the learners' archetype has a large influence on their mean percentage score in the online tests. Learners can optionally be categorised according to an archetype such as 'flourisher' or 'explorer'. These supposedly can help to explain certain character traits and therefore can also potentially give clues as to what personality type flourishes the most with MOOCs. Data understanding:
The data files contain the leraner ID, the archetype assigned to the given learner and the time at which they responed. I am interested in only the learner ID and the archetype. Note these data files are far shorter than the number of learners in a given cohort, probably because only a minority of users decided to be assigned an archetype. The data is only available for run three to seven of the course. Data preparation: To prepare the data I simply removed all of the data except for the student ID and the archetype. I then merged each archetype to its corresponding student statistics dataframe from the second CRISP-DM cycle. I then merged all of these into a combined dataframe.


Boxplot of Mean Percentage for each archetype
Mean Percentage
100
75
50
25
0
Advancers
Explorers
Fixers
Flourishers
Hobbyists
Other
Vitalisers
Preparers
Archetype
factor(archetype)
Advancers
Explorers
Fixers
Flourishers
Hobbyists
Other
Vitalisers
Preparers

Modeling:


Boxplot of mean number of questions answered for each archetype
Mean Number of Questions Answered
60
40
20
0
Advancers
Explorers
Fixers
Flourishers
Hobbyists
Other
Vitalisers
Preparers
Archetype
factor(archetype)
Advancers
Explorers
Fixers
Flourishers
Hobbyists
Other
Vitalisers
Preparers