# Milestone report:
# A deeper dive into the data set

The initial idea for the project comes from thinking about motivation levels for online courses. In more detail, it leads to thinking about who are the students, gender and age distribution, background. How many of these students withdraw from the courses and is it possible to stop the trend?

To sum up, the problem is to classify students who withdraw from self-learning/online courses so that earlier intervention can prevent them from withdrawing from the course

The client will be Open University who can use the classification to increase the proportion of students who withdraw from their courses. This can be applied to other courses too with a high degree of self-learning such as online courses. It will not only make the courses more relevant and useful to the students but also increase their success rate.

Many studies have been conducted into student progress at the Open University. Kennedy and Powell (1976) reported the following after conducting research on withdrawals from the Open Univeristy:

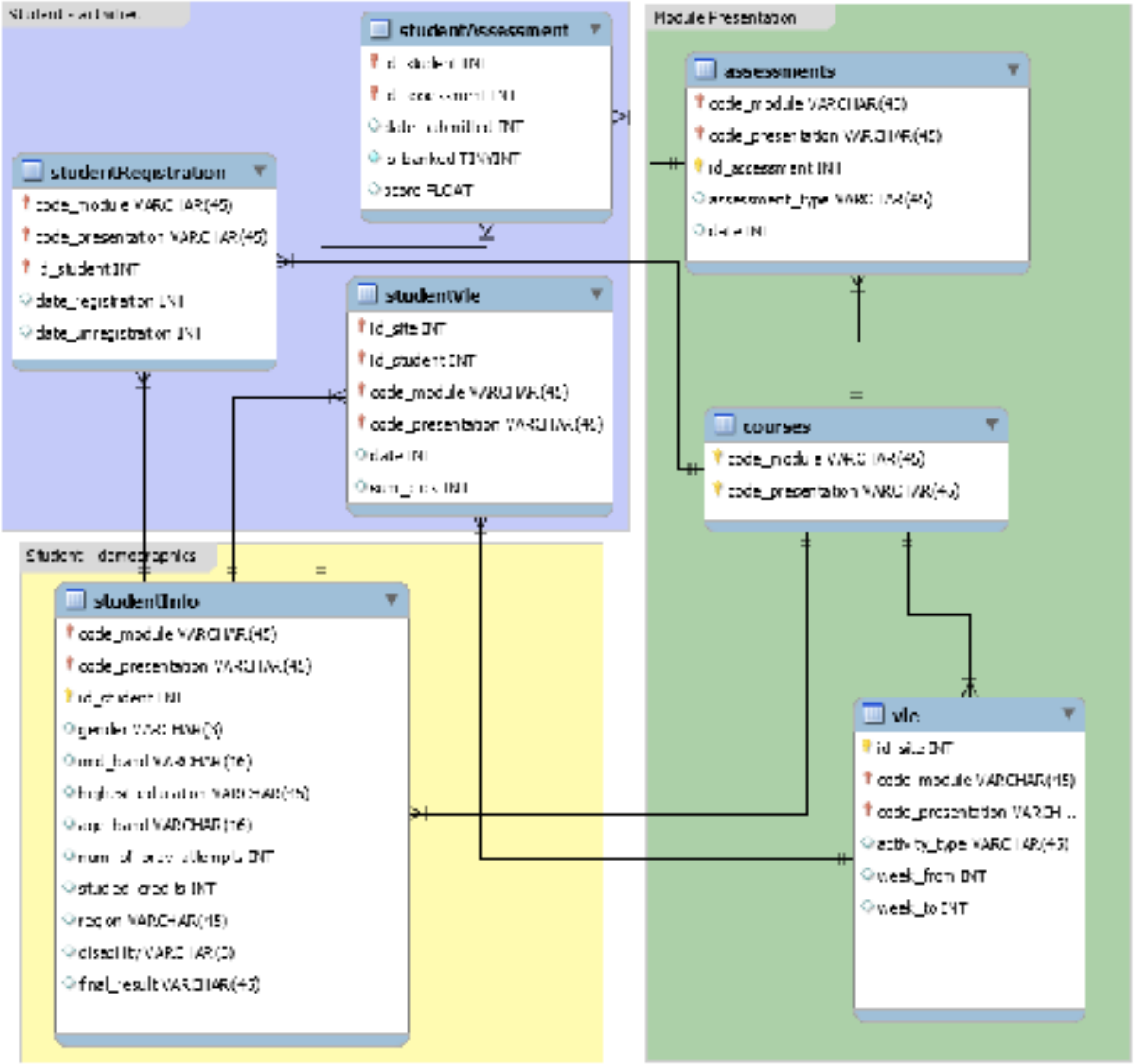
> Such students often have a demoralizing history of educational failure and bring feelings of insecurity and educational and intellectual inferiority to their studies. This problem is compounded by the fact that by joining the Open University a student is not simply put in contact with a body of knowledge; he is forced to accommodate himself to a specialized pattern of interaction and communication. The student does not only have to learn new vocabularies; he must learn to debate and communicate in a manner which is acceptable to the academic community (p. 69).

Simpson(2006) has reviewed ways to predict student success and highlighted the significance of statistical methods like logistic regression rather than assumptions and theories related to the profile of these students and influencing factors in their decision to withdraw. It is only possible to create the profile with limited success but nonetheless, the target profile can be used to intervene early on or even offer extra services to the students who are “at risk.” This analysis solution will provide a new insight based not on intuition but on concrete data analysis.

## Data

The dataset contains multiple csv files giving a complete picture of the overall student data. Schema for the data can be found here: https://analyse.kmi.open.ac.uk/open_dataset

The dataset contains information about the student profile(demographics), date of registration and unregistration, courses taken along with assessments, scores and vie engagement.



To start with, we look at the student profile and the possibility of predicting chances of withdrawal based solely on demographics. The data is in the file studentInfo.csv

The data fields(columns) which give an insight into factors affecting withdrawal are:

Gender
Disability
imd_band
highest_education
number of previous attempts

Furthermore, number of previous attempts should be related to assessment which can be found at studentAssessment.csv. The scores may be related to their VLE engagement and it is a possible course to investigate.

## Cleaning and wrangling

Since the data was taken from a well established source, it was very clean with few missing values. The only complication I had was to integrate the data from multiple files to get the complete picture.

## Limitations of data

Student withdrawal is a result of several personal, financial, psychological and other external factors. The information from the dataset is inadequate. In addition, it is a small subset of the overall data. Some missing information can also hinder analysis: Level of entry, course codes.

## Data exploration

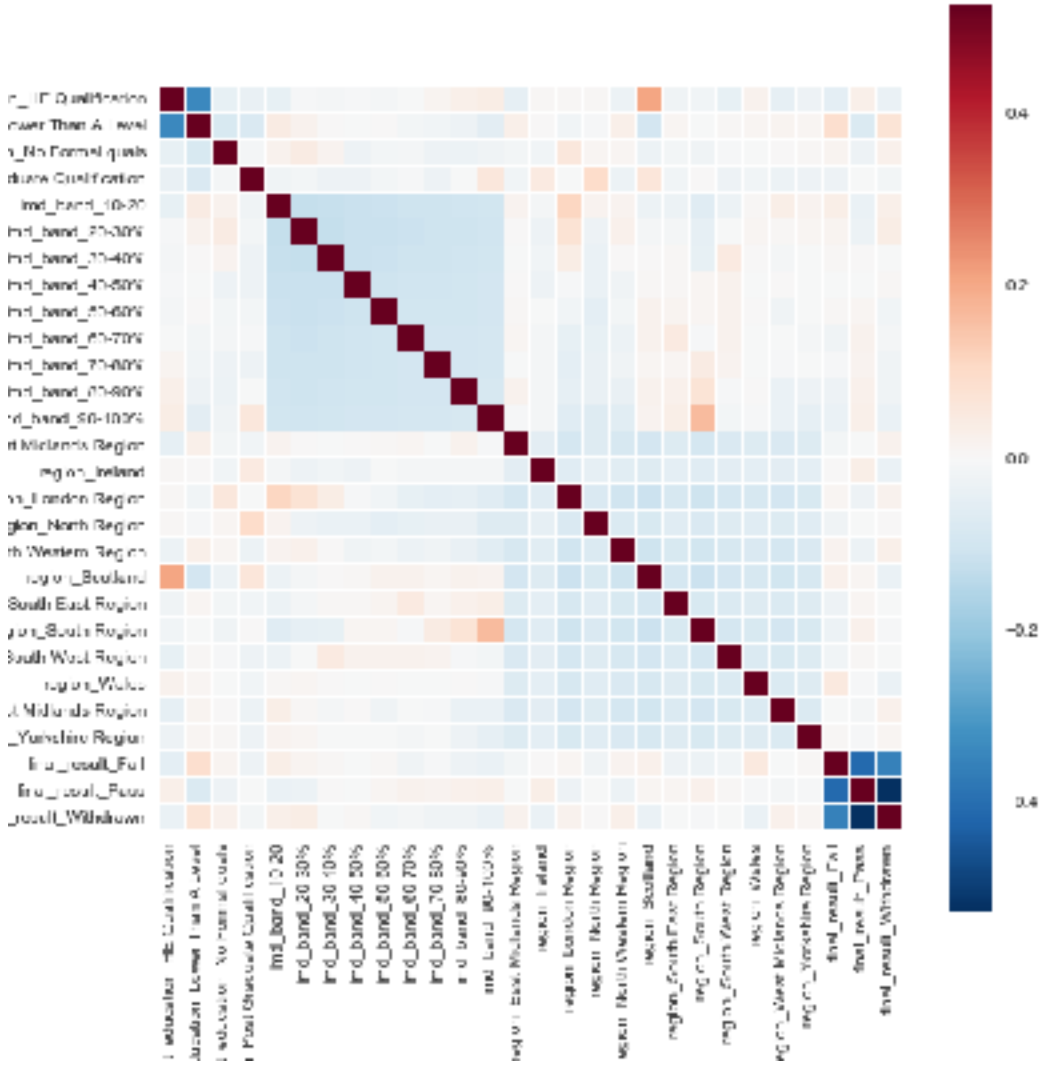
The initial data analysis focussed on getting the overall information on final results of students:
The results are less enthusiatic: Only 39% passes and 10% with distinction.

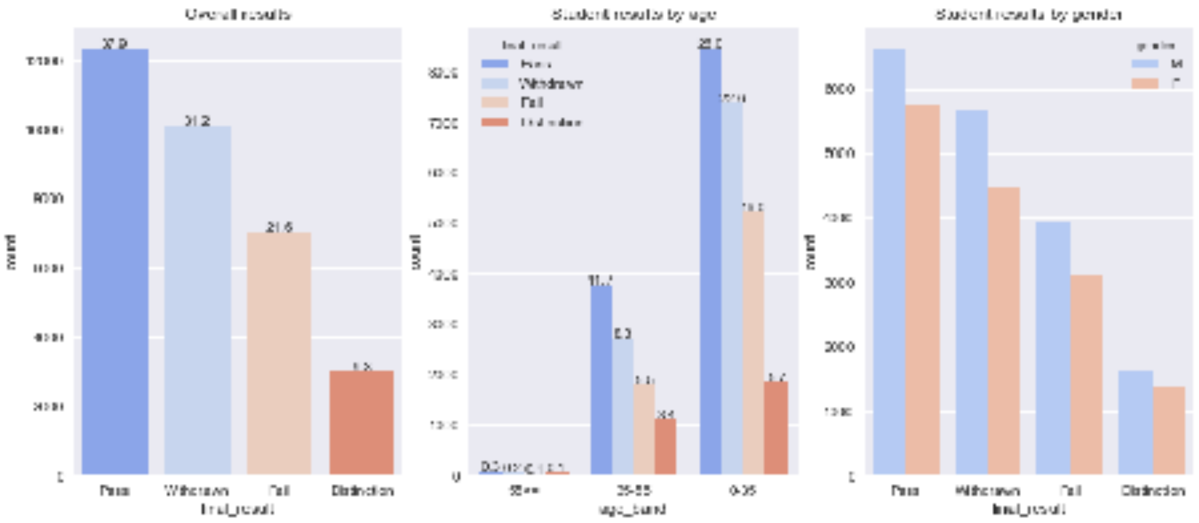
The percentage of students who withdraw is quite high as is those who Fail.
Hence further analysis will be given to these groups.

The next step is to find correlation between students who withdraw and:

* region
* imd_score
* highest education
* previous attempts
* number of credits

To do so, data was converted (in the original dataset) from categorical to numerical as it is easier to analyze. This allowed us to visually represent the correlation


HE Qualification
Lower Than A Level
No Formal quals
Graduate Qualification
imd_band_10-20
imd_band_20-30%
imd_band_30-40%
imd_band_40-50%
imd_band_50-60%
imd_band_60-70%
imd_band_70-80%
imd_band_80-90%
imd_band_90-100%
East Midlands Region
region_Ireland
London Region
region_North Region
North Western Region
region_Scotland
South East Region
region_South Region
South West Region
region_Wales
West Midlands Region
Yorkshire Region
final_result_Fail
final_result_Pass
final_result_Withdrawn
0.4
0.2
0.0
−0.2
−0.4


Overall results
count
12000
10000
8000
6000
4000
2000
0
37.9
31.2
21.6
Pass
Withdrawn
Fail
Distinction
final_result
Student results by age
final_result
Pass
Withdrawn
Fail
Distinction
8000
7000
6000
5000
4000
3000
2000
1000
0
28.0
11.7
8.3
55<=
35-55
0-35
age_band
Student results by gender
gender
M
F
6000
5000
4000
3000
2000
1000
0
Pass
Withdrawn
Fail
Distinction
final_result

An insight into profile of students who withdraw¶
As expected, the number of students who withdraw is higher for

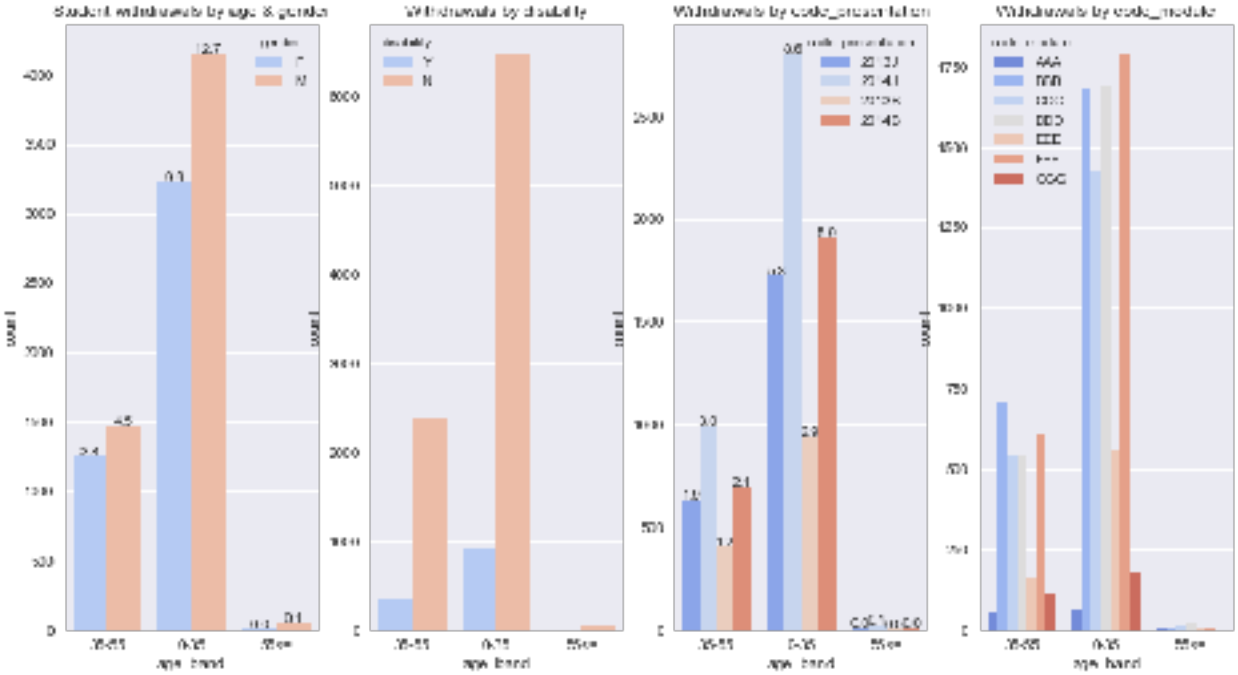
- 0-35 age group,
- not disabled students
- male students.

This is not significant as more students enroll in this group.

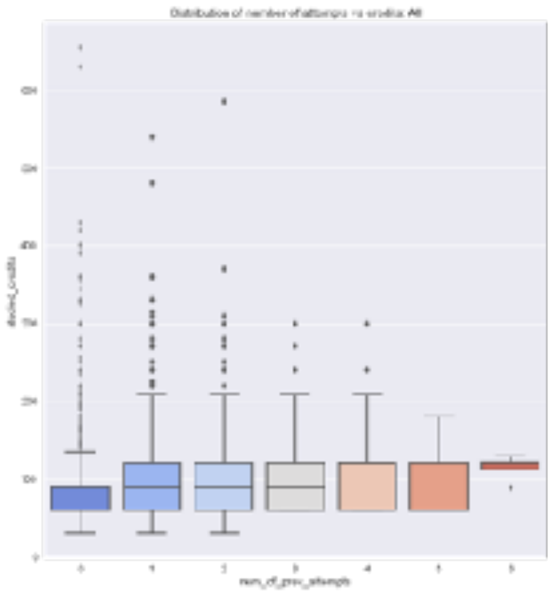
For the same reasons, more students withdraw from BBB, DDD and FFF code modules
Clearly, fewer students from Ireland, Scotland and Wales (negative correlation)
More students from Midlands, London and the North-West

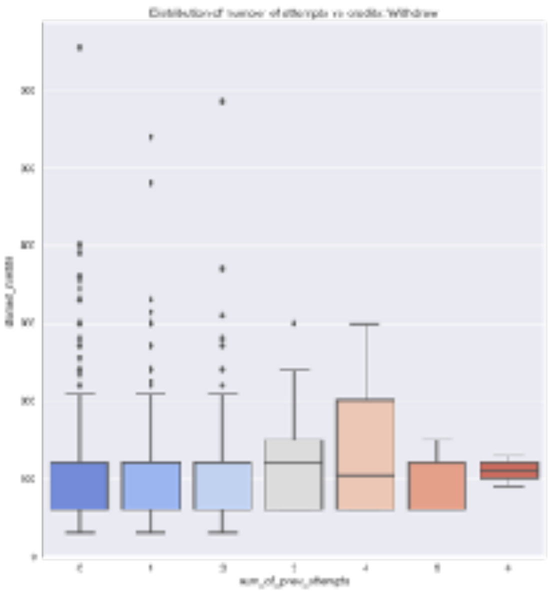
More interesting information can be gleaned from the heatmaps

higher_education is correlated to imd_band: those from 10-30% having A level Quals or less whereas 90-100% having a Post Graduate Qual. This is a presumed trend and it can be observed here. It can also be seen that the students who already have a degree are less likely to withdraw



Increase from 2013 to 2014.
More students who enroll in October withdraw compared to January but overall, more student enroll in October. So this is not very significant.





While it is not directly relevant, number of previous attempts is more for students who have not attended university before. Hence more the number of studied credits, higher the chance of withdrawal

## Approach to final solution

After the initial data exploration, the data is going to be separated into a data frame of “at risk” students, or disadvantaged students. A correlation is done with assessment scores or VLE engagement.
Using the result, we can decide the features to classify the data using Logistic Regression to determine whether a student will withdraw or not.