Dataset Overview

Select a dataset

Courses

Analysis

Select an analysis

Final Results Distribution

This is our project of hardware and software for big data.

Open University Learning Analytics Dataset

Loading all the dataset

Courses

	code_module	code_presentation	module_presentation_length
0	AAA	2013J	268
1	AAA	2014J	269
2	BBB	2013J	268
3	BBB	2014J	262
4	BBB	2013B	240

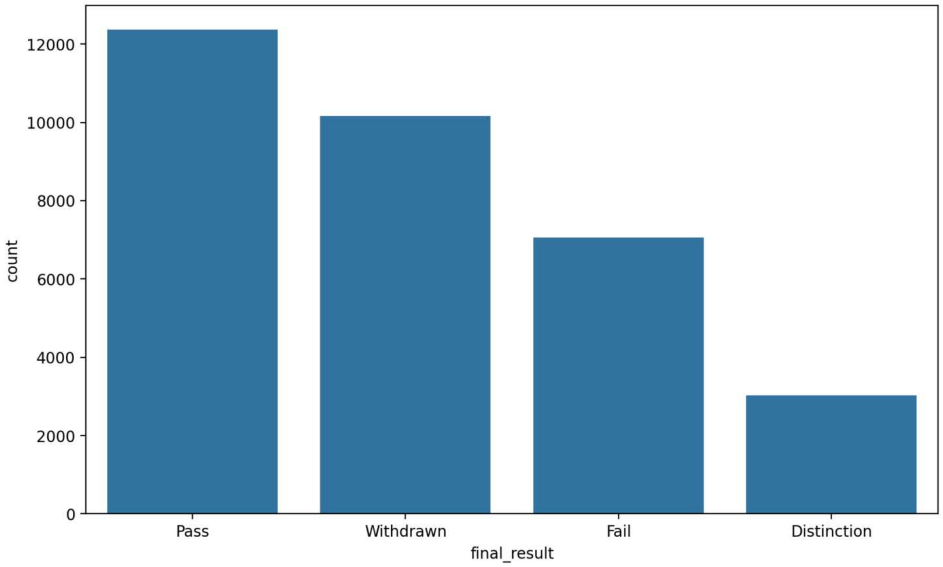
SIZE `('Row, Coloumn', (22, 3))`

Data Types

	0
code_module	object
code_presentation	object
module_presentation_length	int64

	module_presentation_length
count	22
mean	255.5455
std	13.6547
min	234
25%	241
50%	261.5
75%	268
max	269

Distribution of Final Results



NULL VALUES DETECTION

	0
id_site	0
code_module	0
code_presentation	0
activity_type	0
week_from	5,243
week_to	5,243

Find unregistered students according to registration table. Then check whether they are consistent with the final results at StudentInfo table. #If a student is unregistered, final result must be recorded as 'Withdrawn'.

Select unregistered students according to registration table

Unregistered students without a 'Withdrawn' in final result column # Semantic Error -- If a student unregistered, have to have 'Withdrawn' as final result!

	code_module	code_presentation	id_student	gender	region	highest_education
408	BBB	2013B	540,530	F	North Western Region	Lower Than A Level
724	BBB	2013J	362,907	F	South West Region	Lower Than A Level
729	BBB	2013J	365,288	F	South Region	A Level or Equivalent
875	BBB	2013J	554,243	F	South West Region	Lower Than A Level
1,777	BBB	2014J	39,208	F	Wales	A Level or Equivalent

8 of these 9 students unregistered at the first day of the module presentation or earlier (by looking date_unregistration column), this may be a cause for that inconsistency.This also makes a 'Fail' impossible as a final result. The above 9 records' final_result entry will be changed into Withdrawn in the next step.

Correction info_stu table's final_result entries

Consistency of Weights of Assesments

As mentioned under Assessments table explanation, weights of the exams should sum up to %100 and other assessments should sum up to %100. Hence, total weight for each module_presentation should be 200.

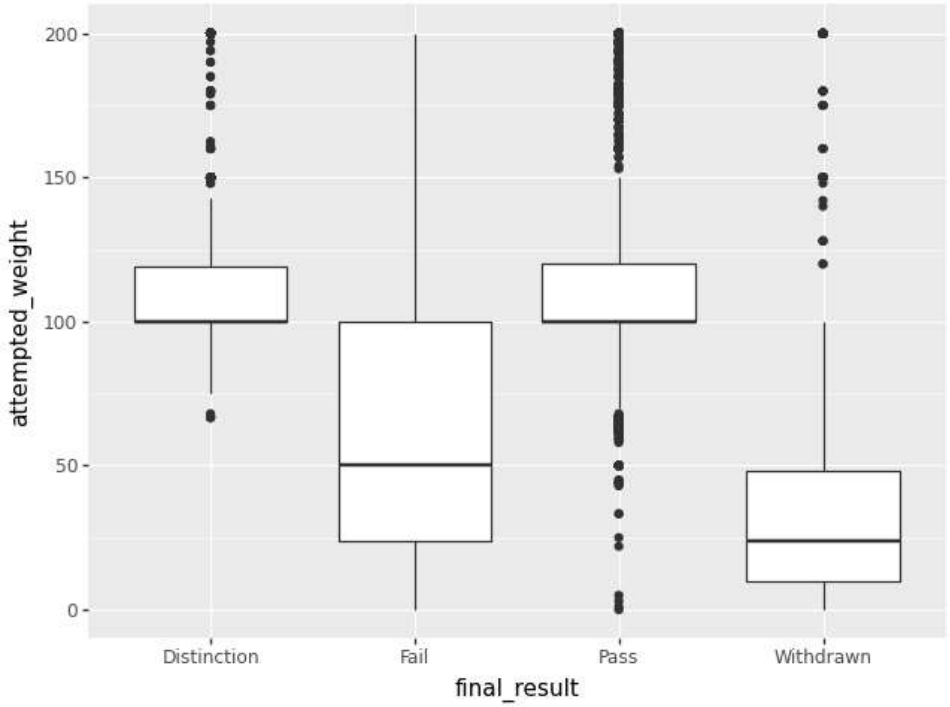
code_module	code_presentation	total_weight
AAA	2013J	200
AAA	2014J	200
BBB	2013B	200
BBB	2013J	200
BBB	2014B	200
BBB	2014J	200
CCC	2014B	300
CCC	2014J	300
DDD	2013B	200
DDD	2013J	200

Module CCC adds up to 300 and module GGG adds up to 100 -- these 2 will be investigated.

code_module	code_presentation	assessment_type	type_weights
CCC	2014B	CMA	25
CCC	2014B	Exam	200
CCC	2014B	TMA	75
CCC	2014J	CMA	25
CCC	2014J	Exam	200
CCC	2014J	TMA	75
GGG	2013J	CMA	0
GGG	2013J	Exam	100
GGG	2013J	TMA	0
GGG	2014B	CMA	0

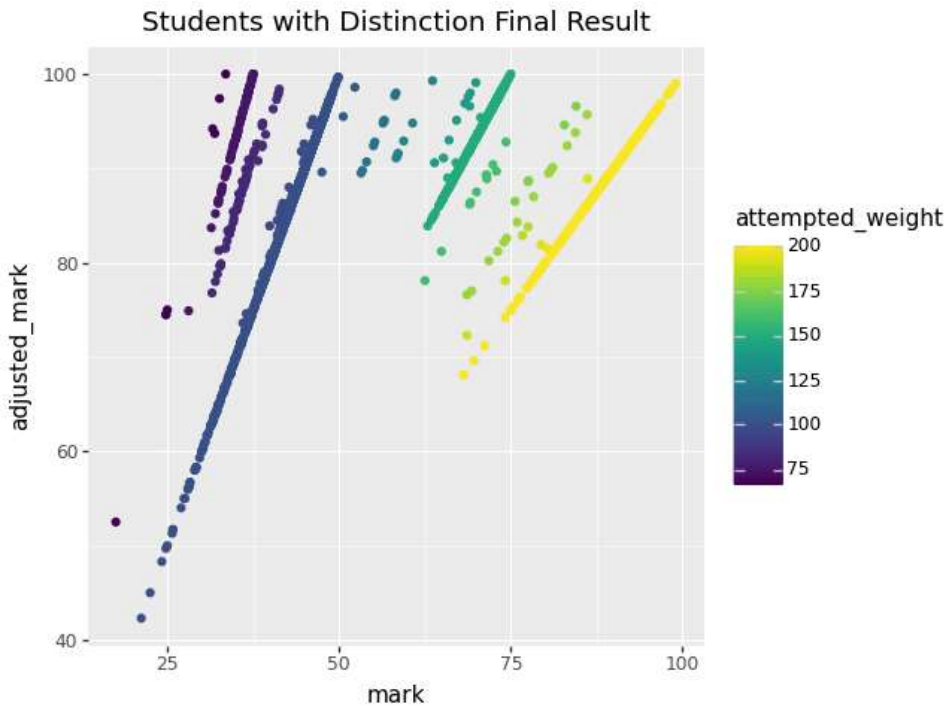
Module CCC has an inconsistency. Total exam weight is 200%. Module CCC has 2 exams in every presentation so It is very likely to be entered as 2 times 100"%" instead of 50"%" and 50%. This will be fixed in the upcoming step."Also Module GGG is incompatible. There is no weight assigned neither to TMA type nor to CMA type. For convenience TMA type assignments" weight will be imputed as to add up to 100.

``



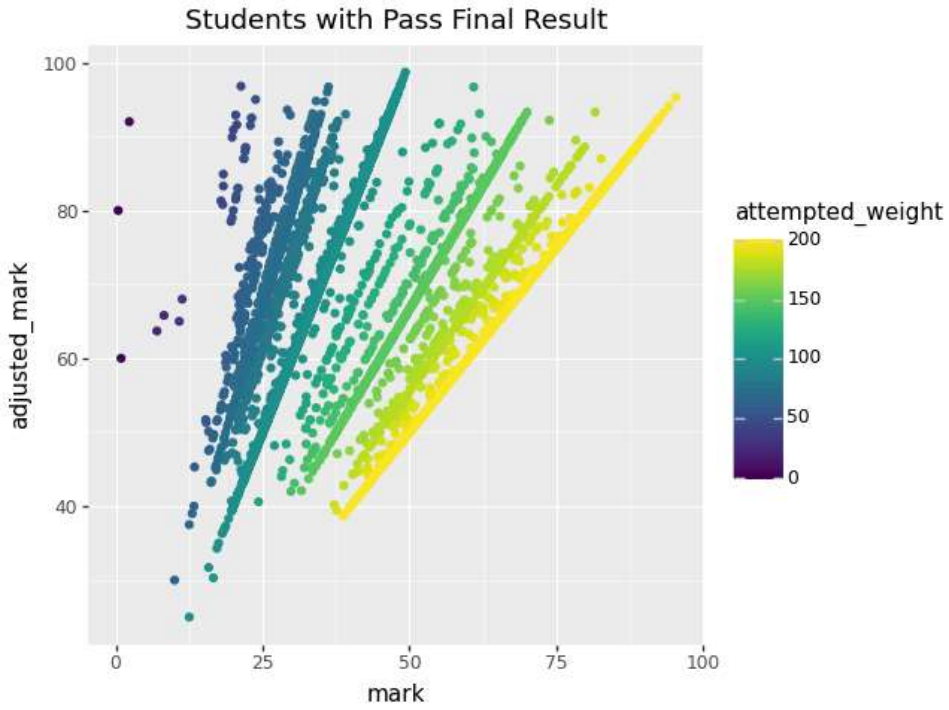
boxplot and final score

``



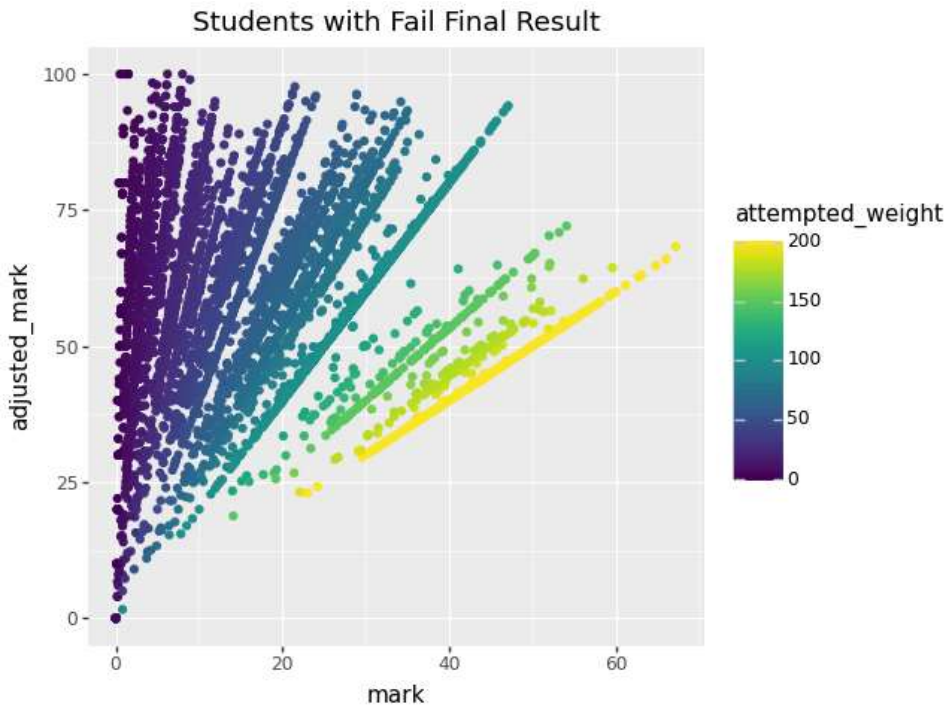
students Final result with distinction

``



students pass with final result

``



Students with Fail Final Result

By comparing the above 3 graphs-

Difference between adjusted_mark and mark is much higher for "Fail" students.Suggesting that these students probably didnt attempted these missing assessments indeed. Issue is not about being included in the dataset.For the attempted weight of 200 (attempting all assessments);Almost all Distinction results are higher than 70.Almost all Pass marks are higher than 40Therefore, 0-40 band will be accapted as "Fail", 40-70 band will be accapted as "Pass" and 70-100 as "Distinction".

Most of the entries of 'week_from' and 'week_to' attributes are missing so the analysis will not be focusing on the datesIn order to get ride of the extra load on memory, these columns will be dropped in the next step.

Null values in student registration data

	0
code_module	0
code_presentation	0
id_student	0
date_registration	45
date_unregistration	22,521

70'%' of the rows are missing date_unregistration. This means that 70'%' of the students don't withdraw the modules.

	0
code_module	0
code_presentation	0
id_student	0
date_registration	0
date_unregistration	0

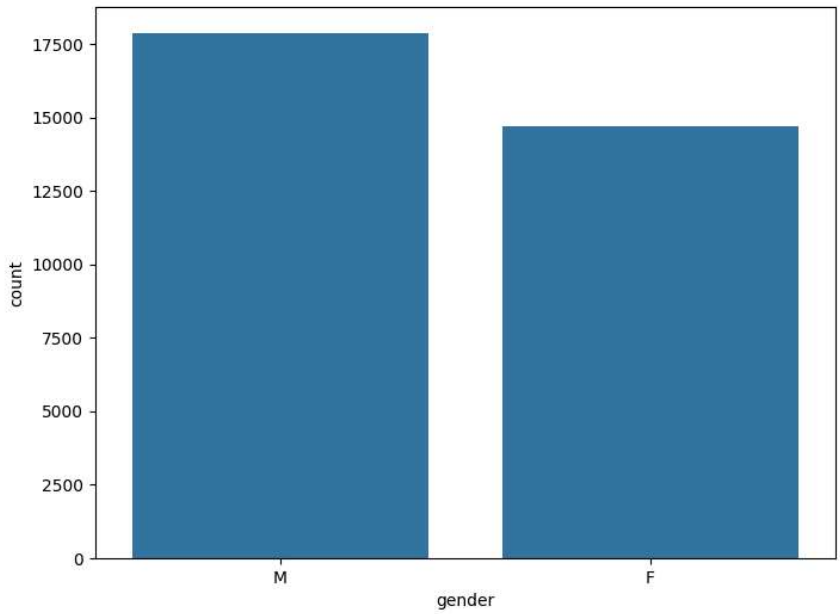
	0
code_module	0
code_presentation	0
id_student	0
gender	0
region	0
highest_education	0
imd_band	0
age_band	0
num_of_prev_attempts	0
studied_credits	0

```
(32593, 12)
```

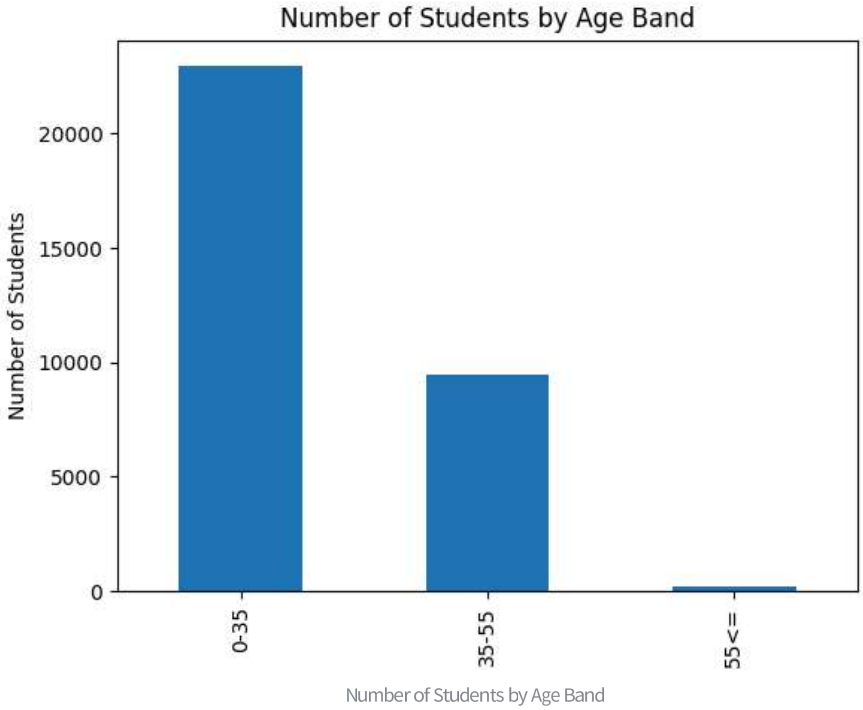
	0
id_assessment	0
id_student	0
date_submitted	0
is_banked	0
score	0

	0
code_module	0
code_presentation	0
id_assessment	0
assessment_type	0
date	0
weight	0

Exploratory data analysis

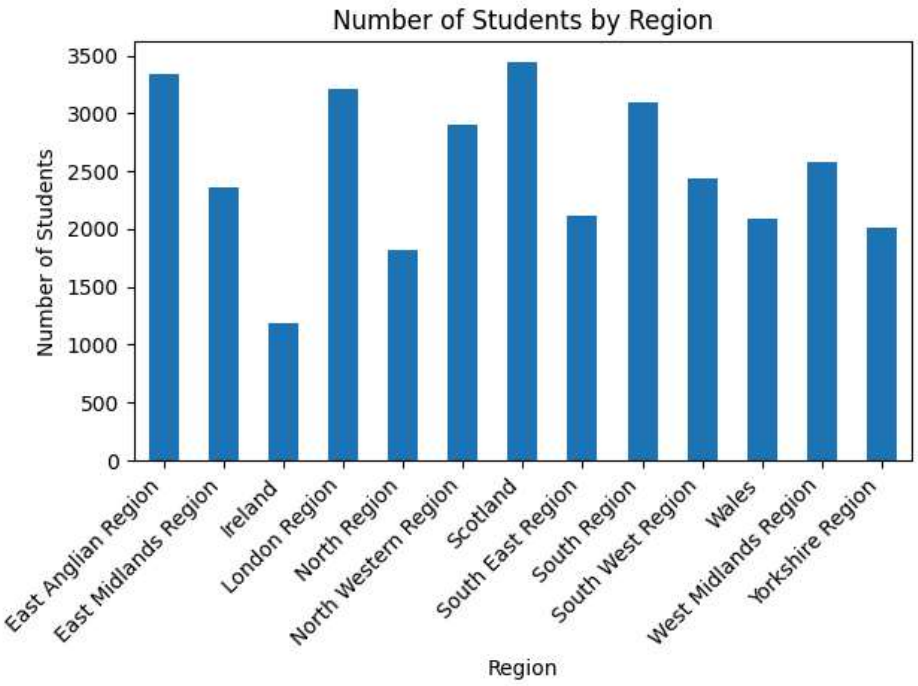


Gender Distribution

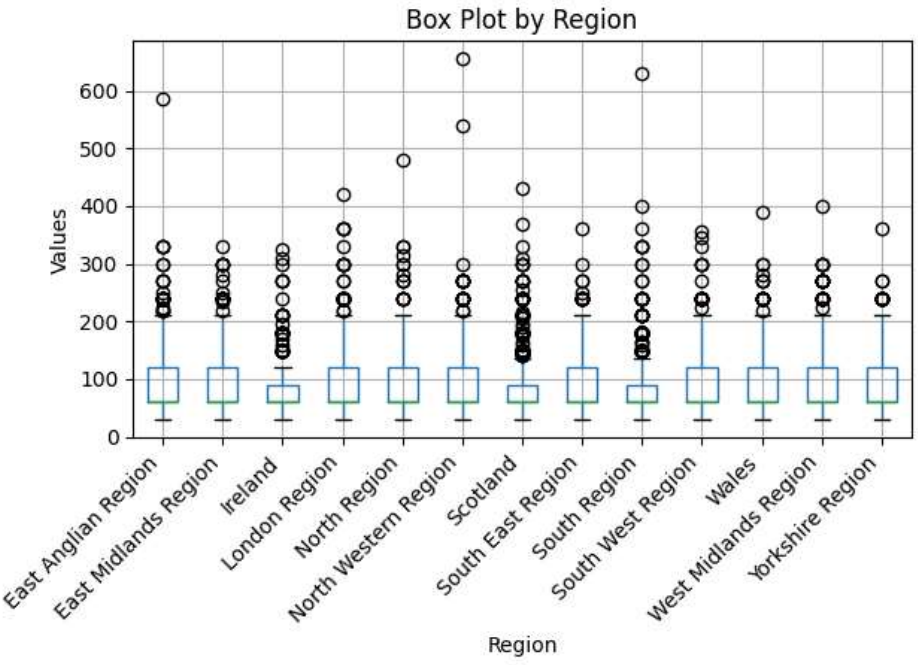


Number of Students by Age Band

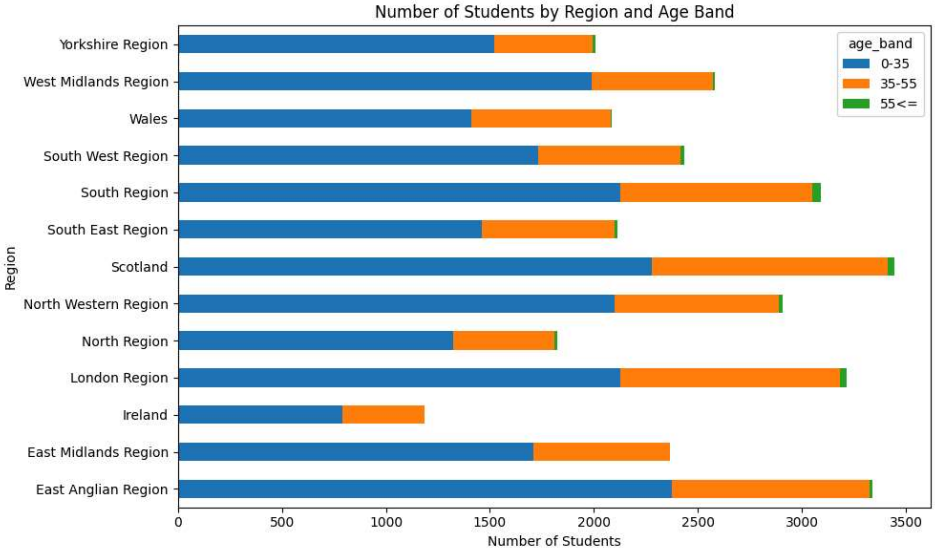
Students by regions



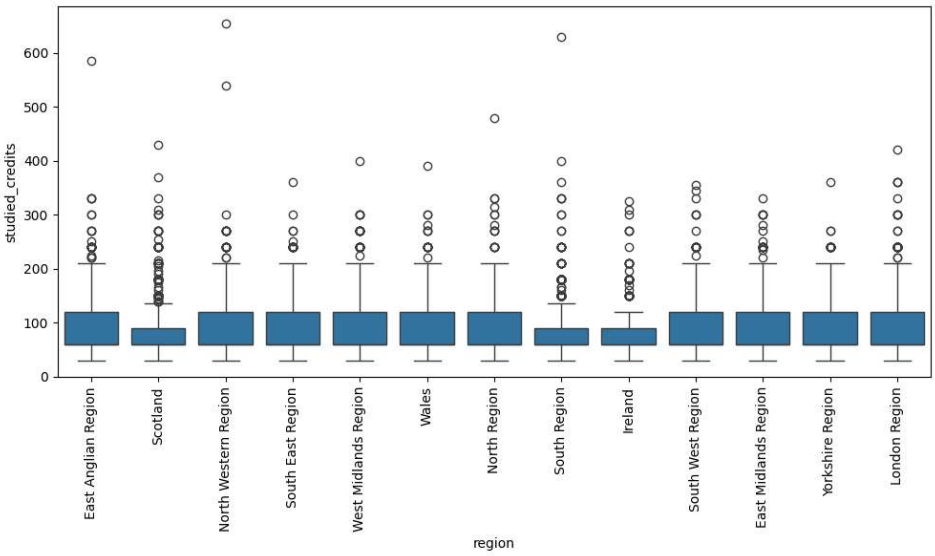
Number of Students by Region



Box Plot by Region



Number of Students by Region and Age Band



students credit by region

selecting a subset of cols which are of importance to us and grouping them by student id and aggregating them using median

	id_student	num_of_prev_attempts	studied_credits
0	3,733	0	60
1	6,516	0	60
2	8,462	0.5	75
3	11,391	0	240
4	23,629	2	60

new dataframe for students information with his education, region and age deatils. dropped duplicate values

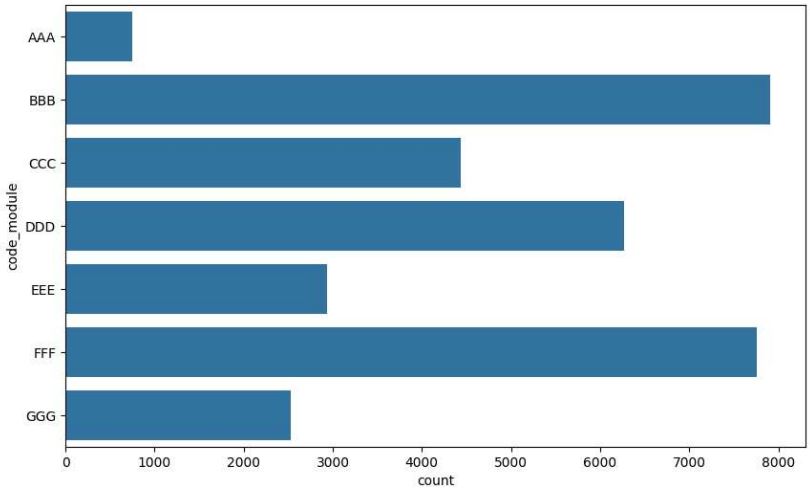
	id_student	gender	region	highest_education	imd_band	age_band
0	11,391	M	East Anglian Region	HE Qualification	90-100%	55<=
1	28,400	F	Scotland	HE Qualification	20-30%	35-55
2	30,268	F	North Western Region	A Level or Equivalent	30-40%	35-55
3	31,604	F	South East Region	A Level or Equivalent	50-60%	35-55
4	32,885	F	West Midlands Region	Lower Than A Level	50-60%	0-35

```
(28857, 6)
```

	0
id_student	int64
gender	object
region	object
highest_education	object
imd_band	object
age_band	object

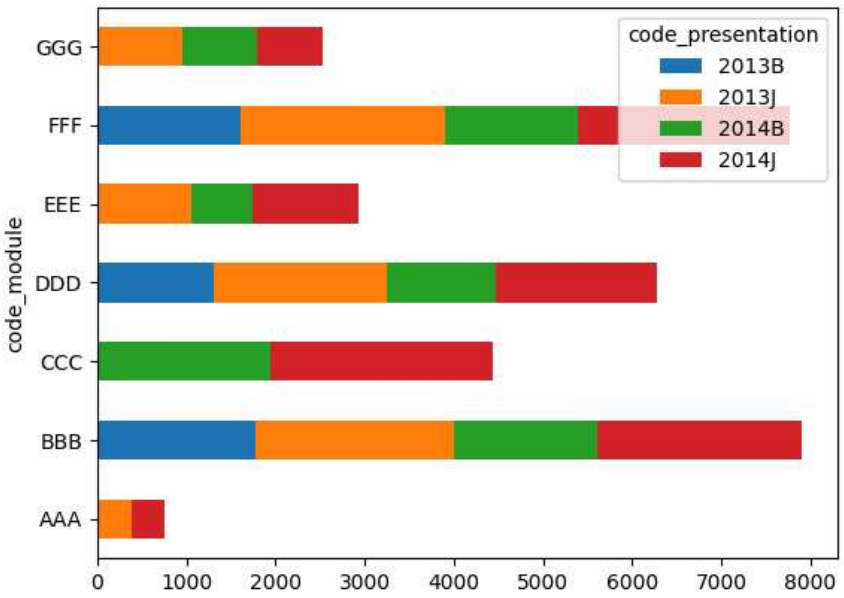
	id_student
count	28,857
mean	707,834.1328
std	550,960.5738
min	3,733
25%	508,100
50%	590,555
75%	646,481
max	2,716,795

students enrolled in different courese wise



students in different courses

BBB and FFF is most popular courses in between students.



courses offer in different intakes and enrolled number of students

'B' is for courses offered in Feb and 'J' is for courses offered in Oct.

course 'CCC' is something introduced in 2014 only.

course 'AAA' has a very low student count as compared to other courses

	id_student	activity_type	sum_click
0	6,516	dataplus	5.25
1	6,516	forumng	2.5771
2	6,516	homepage	3.1456
3	6,516	oucontent	8.1793
4	6,516	resource	1.6316

Merging Data

Basic info of the dataset:

```
(213166, 22)
```

	id_student	num_of_prev_attempts	studied_credits	date_registration	date_unregistration	id_ass
count	213,166	213,166	213,166	213,166	213,166	
mean	704,426.372	0.1574	78.4474	-67.7414	22.8787	26,
std	549,752.5243	0.4623	38.3606	48.6592	42.9405	8,
min	3,733	0	30	-322	-365	
25%	506,930	0	60	-96	12	
50%	585,662	0	60	-54	12	
75%	633,609	0	90	-29	12	
max	2,716,795	6	655	167	444	

```
RandomForestClassifier(random_state=42)
```

Confusion Matrix:

0	1	2	3
3,220	41	2,748	1
68	3,155	3,370	32
555	570	22,338	8
15	79	334	6,100

Classification Report:

precision recall f1-score support

```
0       0.83      0.54      0.65      6010
       1       0.82      0.48      0.60      6625
       2       0.78      0.95      0.85     23471
       3       0.99      0.93      0.96      6528

accuracy                           0.82     42634
```

macro avg 0.86 0.72 0.77 42634 weighted avg 0.82 0.82 0.80 42634

None

None

First few rows of the dataset after preparation:

date_registration	code_module	code_presentation	id_student	gender	region	highest_education	im
1970-01-01 00:00:00	1	2	571,732	0	7	2	
1970-01-01 00:00:00	1	2	570,721	0	1	1	
1970-01-01 00:00:00	1	2	320,661	0	4	0	
1970-01-01 00:00:00	1	2	534,428	0	4	2	
1970-01-01 00:00:00	1	2	534,428	0	4	2	

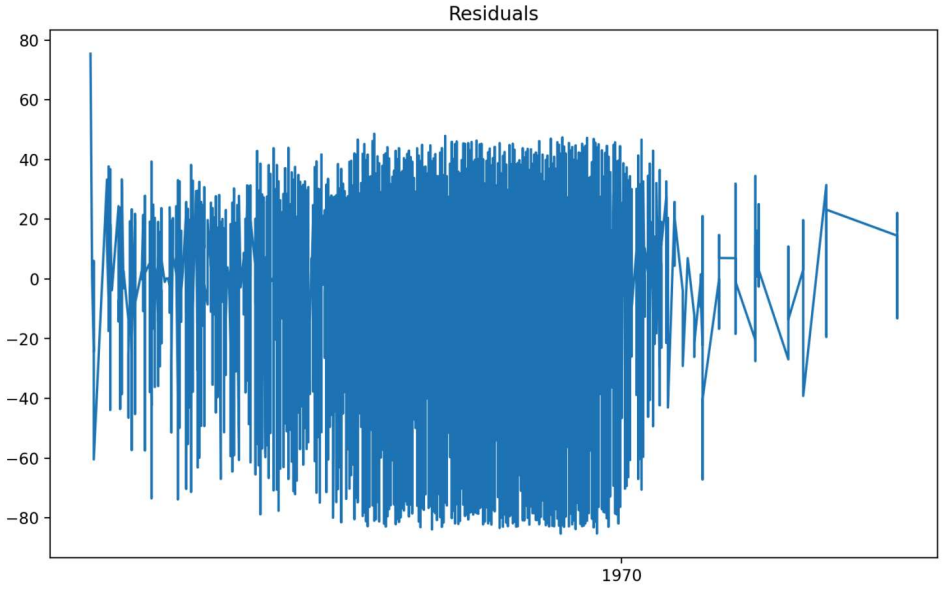
ARIMA Model Summary:

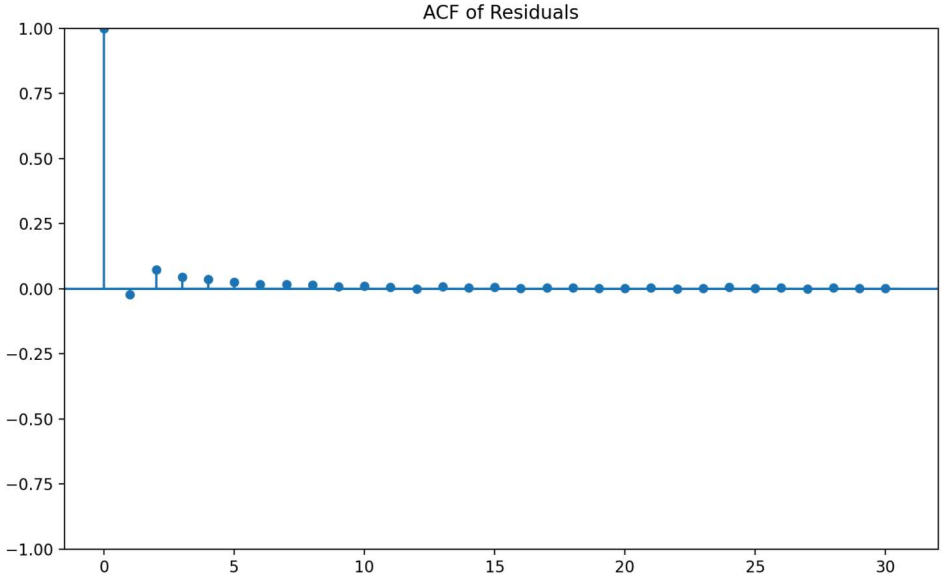
```
SARIMAX Results
==========================================================================================
Dep. Variable:                    score   No. Observations:                213166
Model:                 ARIMA(1, 1, 1)   Log Likelihood             -920062.599
Date:                Thu, 20 Jun 2024   AIC                        1840131.199
Time:                        15:56:57   BIC                        1840162.008
Sample:                             0   HQIC                       1840140.241
                                - 213166
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2524      0.002    137.837      0.000       0.249       0.256
ma.L1         -0.9933      0.000  -3961.094      0.000      -0.994      -0.993
sigma2       328.4802      0.736    446.402      0.000     327.038     329.922
===================================================================================
Ljung-Box (L1) (Q):                 100.86   Jarque-Bera (JB):             65870.0
Prob(Q):                              0.00   Prob(JB):                         0.0
Heteroskedasticity (H):               1.00   Skew:                            -1.0
Prob(H) (two-sided):                  0.87   Kurtosis:                         4.7
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-ste
```

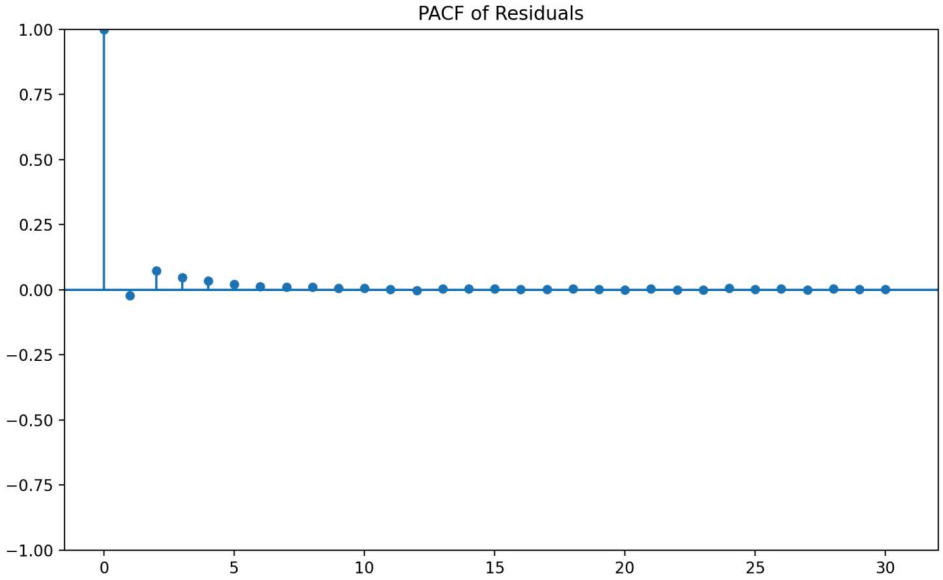
Potential date columns:

```
[
  0 : "date_unregistration"
  1 : "date_submitted"
  2 : "date"
]
```


Error: The column 'date_registration' does not exist in the dataset.







ValueError: Dataframe must have columns "ds" and "y" with the dates and values respectively.

Traceback:

```
File "C:\Users\kk928\AppData\Local\Programs\Python\Python312\Lib\site-packages
    exec(code, module.__dict__)

File "C:\Users\kk928\streamlit_oulad_app.py", line 714, in <module>
    model.fit(train)

File "C:\Users\kk928\AppData\Local\Programs\Python\Python312\Lib\site-packages
    model_inputs = self.preprocess(df, **kwargs)
                   ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

File "C:\Users\kk928\AppData\Local\Programs\Python\Python312\Lib\site-packages
    raise ValueError(
```