

Open University Learning Analytics Dataset

--Inspect table elements

```
SELECT * FROM assessments;  
SELECT * FROM courses;  
SELECT * FROM studentAssessment;  
SELECT * FROM studentInfo;  
SELECT * FROM studentRegistration;  
SELECT * FROM vle;
```

--Change final_result to 0 and 1s

```
UPDATE studentInfo  
SET final_result = REPLACE(final_result, 'Pass', '1');
```

```
UPDATE studentInfo  
SET final_result = REPLACE(final_result, 'Fail', '0');
```

```
UPDATE studentInfo  
SET final_result = REPLACE(final_result, 'Withdraw', '0');
```

--Check if all of final_result is 0s and 1s

```
SELECT * FROM studentInfo WHERE final_result <> '1' AND final_result <> '0';
```

--Remove 0n and Distinction

```
UPDATE studentInfo  
SET final_result = REPLACE(final_result, '0n', '0');
```

```
UPDATE studentInfo  
SET final_result = REPLACE(final_result, 'Distinction', '0');
```

--Check if all of final_result is 0s and 1s (again)

```
SELECT * FROM studentInfo WHERE final_result <> '1' AND final_result <> '0';
```

--Convert final result to int

```
ALTER TABLE studentInfo ALTER COLUMN final_result int;
```

--Find out percentage of presentations passed

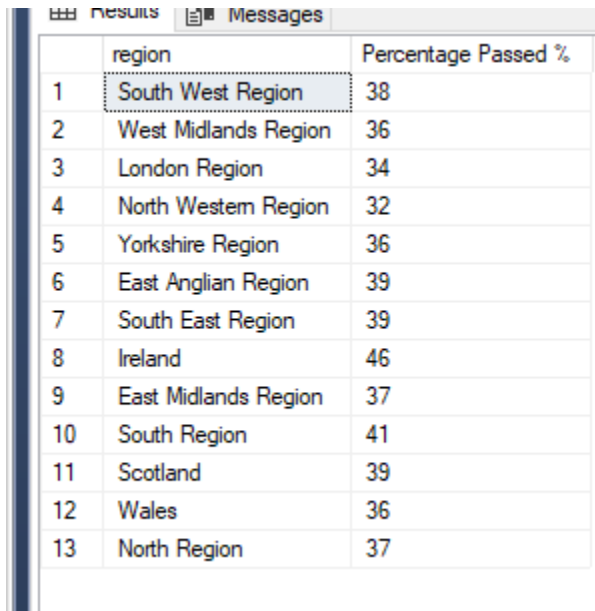
```
SELECT SUM(final_result) * 100 / COUNT(final_result) AS 'Percentage Passed %'  
FROM studentInfo;
```

--(It was 37% overall, how can we increase this?)

QUESTION TO SOLVE: What adjustments can be made to the UK Virtual Learning Environment to increase the percentage of presentations passed to 45% within one year after implementation?

--Find out percentage of presentations passed by region

```
SELECT region, SUM(final_result) * 100 / COUNT(final_result) AS 'Percentage Passed %'  
FROM studentInfo GROUP BY region;
```



	region	Percentage Passed %
1	South West Region	38
2	West Midlands Region	36
3	London Region	34
4	North Western Region	32
5	Yorkshire Region	36
6	East Anglian Region	39
7	South East Region	39
8	Ireland	46
9	East Midlands Region	37
10	South Region	41
11	Scotland	39
12	Wales	36
13	North Region	37

The North Western and London Regions have particularly low percentages. Is there a reason for this?

--Left join the studentInfo table with studentAssessment

```
SELECT *  
FROM studentInfo AS I  
LEFT JOIN studentAssessment AS A  
ON I.id_student = A.id_student;
```

--Compare scores by region

```
SELECT I.region, AVG(A.score) AS AverageScore  
FROM studentInfo AS I  
LEFT JOIN studentAssessment AS A  
ON I.id_student = A.id_student  
GROUP BY region;
```

Scores are pretty evenly spread across regions.

	region	AverageScore
1	West Midlands Region	74
2	East Midlands Region	75
3	Yorkshire Region	74
4	Wales	75
5	Ireland	75
6	South Region	76
7	North Western Region	73
8	North Region	77
9	East Anglian Region	75
10	London Region	74
11	South East Region	76
12	South West Region	75
13	Scotland	76

--Compare scores by region, then highest education only for those two regions

```
SELECT I.region, highest_education, AVG(A.score) AS AverageScore
FROM studentInfo AS I
LEFT JOIN studentAssessment AS A
ON I.id_student = A.id_student
WHERE region = 'London Region' OR region = 'North Western Region'
GROUP BY region, highest_education;
```

Students with no formal qualifications scored significantly lower than those with qualifications.
How many students have no formal qualifications in each region?

	region	highest_education	AverageScore
1	London Region	A Level or Equivalent	74
2	North Western Region	A Level or Equivalent	74
3	North Western Region	No Formal quals	66
4	London Region	Lower Than A Level	72
5	London Region	Post Graduate Qualification	76
6	North Western Region	Lower Than A Level	72
7	London Region	HE Qualification	76
8	North Western Region	HE Qualification	76
9	London Region	No Formal quals	68

--Find count of students with no qualifications by region

```
SELECT I.region, COUNT(*) AS 'Number of students'
FROM studentInfo AS I
LEFT JOIN studentAssessment AS A
```

```

ON I.id_student = A.id_student
WHERE highest_education = 'No formal quals'
GROUP BY region;

```

The chart confirms that there are more students in this category for London and North Western regions. The North Region also has a notable number of students with no formal qualifications.

	region	Number of students
1	West Midlands Region	166
2	East Midlands Region	48
3	Yorkshire Region	124
4	Wales	90
5	South Region	108
6	North Western Region	209
7	North Region	209
8	East Anglian Region	194
9	London Region	389
10	South West Region	90
11	South East Region	78
12	Scotland	25

There may be a correlation between students with no formal qualifications and failing this presentation. How many students are in each region?

--Find distinct number of students per region

```

SELECT DISTINCT I.region, COUNT(*) AS 'Number of students'
FROM studentInfo AS I
LEFT JOIN studentAssessment AS A
  ON I.id_student = A.id_student
GROUP BY I.region;

```

--Find percentage of students per region with no formal qualifications in London

```

SELECT region, highest_education,
(SELECT COUNT(*)
FROM studentInfo
  WHERE region = 'London Region' AND highest_education = 'No formal quals'
GROUP BY region
)
* 100 /
(SELECT COUNT(*)
FROM studentInfo
WHERE region = 'London Region'
) AS 'Percentage %'

```

```
FROM studentInfo
WHERE region = 'London Region' AND highest_education = 'No formal quals'
GROUP BY region, highest_education
```

(2%)

--Now for the North Western region

```
SELECT region, highest_education,
(SELECT COUNT(*)
FROM studentInfo
WHERE region = 'North Western Region' AND highest_education = 'No formal quals'
GROUP BY region
)
* 100 /
(SELECT COUNT(*)
FROM studentInfo
WHERE region = 'North Western Region'
) AS 'Percentage %'
FROM studentInfo
WHERE region = 'North Western Region' AND highest_education = 'No formal quals'
GROUP BY region, highest_education
```

(1%)

Such a small percentage of low scores from students with no formal qualifications! Higher education doesn't seem to be the problem. How many students had a retake by region?

--Count students if they retook the presentation by region

```
SELECT region,
COUNT(*) AS 'Number of students'
FROM studentInfo
WHERE num_of_prev_attempts > 0
GROUP BY region
```

	region	Number of students
1	South West Region	297
2	West Midlands Region	382
3	London Region	416
4	North Western Region	366
5	Yorkshire Region	242
6	East Anglian Region	393
7	South East Region	254
8	Ireland	163
9	East Midlands Region	297
10	South Region	346
11	Scotland	448
12	Wales	331
13	North Region	237

Nothing unusual here. Compared to student population numbers in each region, these numbers are relatively small. Maybe the module or the presentation itself are the problem?

--Look at one code module at a time for London, this is the general procedure

```
SELECT region, code_module,
COUNT(*) AS 'Number of students',
(SELECT SUM(final_result) * 100 / COUNT(final_result) FROM studentInfo WHERE region =
'London Region' AND code_module = 'AAA' GROUP BY region, code_module) AS 'Percent
Passed'
FROM studentInfo
WHERE region = 'London Region' AND code_module = 'AAA'
GROUP BY region, code_module
```

Rather than type several pages for this problem, the results of this procedure are below:

London

AAA: 65 students, 64% pass rate
BBB: 708 students, 31% pass rate
CCC: 469 students, 26% pass rate
DDD: 519 students, 33% pass rate
EEE: 241 students, 41% pass rate
FFF: 868 students, 34% pass rate
GGG: 346 students, 40% pass rate

North Western

AAA: 55 students, 65% pass rate
BBB: 598 students, 21% pass rate
CCC: 381 students, 22% pass rate

DDD: 629 students, 30% pass rate
EEE: 259 students, 44% pass rate
FFF: 718 students, 35% pass rate
GGG: 266 students, 39% pass rate

It seems that BBB-GGG are more difficult exams, but fewer students have taken AAA.. How do the pass rates compare to a different region?

South Region

AAA: 88 students, 62% pass rate
BBB: 652 students, 41% pass rate
CCC: 395 students, 29% pass rate
DDD: 628 students, 38% pass rate
EEE: 265 students, 52% pass rate
FFF: 783 students, 41% pass rate
GGG: 281 students, 49% pass rate

This leads to the first implementation I would recommend for the VLE, but this assumes such a resource is either not available or under-utilized. I would be asking the board of education questions about their supplemental resources for students before implying direct causation. However, I do not have them on speed dial, so...:

IMPLEMENTATION 1: Open a national tutorial service online for students with a fast track to getting matched with a tutor online for students in lower scoring districts. This can be done using IP address tracking.

How do the presentations look?

London

2013B: 521 students, 30%
2014B: 812 students, 32%
2013J: 803 students, 39%
2014J: 1,080 students, 32%

North Western

2013B: 402 students, 31%
2014B: 738 students, 27%
2013J: 808 students, 35%
2014J: 958 students, 35%

South Region

2013B: 437 students, 40%
2014B: 772 students, 37%

2013J: 847 students, 46%
2014J: 1,036 students, 41%

No particular presentation seems to be the issue. Are students taking too many credits?

```
SELECT AVG(studied_credits) AS 'Average credits'
FROM studentInfo;
```

On average, students take 79 credits. Who is taking more?

```
SELECT region, COUNT(*) AS 'Number of students'
FROM studentInfo
WHERE studied_credits > 79
GROUP BY region
```

	region	Number of students
1	South West Region	890
2	West Midlands Region	961
3	London Region	1329
4	North Western Region	1108
5	Yorkshire Region	724
6	East Anglian Region	1142
7	South East Region	751
8	Ireland	378
9	East Midlands Region	885
10	South Region	1062
11	Scotland	1145
12	Wales	751
13	North Region	694

That's quite a few for both London and North Western. How do the percentages look?

```
SELECT region,
(SELECT COUNT(*)
FROM studentInfo
WHERE studied_credits > 79 AND region = 'London Region'
GROUP BY region) * 100 /
(SELECT COUNT(*)
FROM studentInfo
WHERE region = 'London Region'
GROUP BY region) AS 'Percent'
FROM studentInfo
WHERE region = 'London Region'
GROUP BY region;
```


(London 41%, North Western 38%, South 34%)

The percentage of students taking more than average credit loads are a bit higher where the students performed poorly. One of the many reasons a student may take more credits in a semester is to save money on tuition. Another reason is that they are students preparing for/or are in medical school. Medical school (at least in the states) requires students to take larger than average course loads. If the case is that these two regions have more medical students, then implementation 1 should be adjusted to include more tutors for students in medicine. Otherwise...

IMPLEMENTATION 2: Lower the number of required credit hours required to meet federal full time student requirements. Full time students usually qualify for more financial aid from the government than part time students.

Speaking of finances, the imd_band is essentially a measure of poverty for this study. However, IMD is usually reported as a number from 1 to 32,844 where 1 is the most deprived area and 32,844 is the least deprived area. I attempted to research what the percentages mean in this context, but failed to understand if 0% or 100% was more/less deprived. I'm sure if I lived in the UK this would make more sense, and this highlights the importance of asking questions before and during analysis. Imd_band would have been most likely a key factor, but it is unusable as is without a proper explanation in the data description.

Next, I looked at binary categories to make sure something simple didn't get missed. Gender didn't reveal any differences. Neither did disability. How about module length of time?

--Find the mean number of days in the module

```
SELECT code_module,  
(SELECT AVG(module_presentation_length) FROM courses WHERE code_module = 'AAA') AS  
'AVG module length'  
FROM courses  
WHERE code_module = 'AAA'  
GROUP BY code_module
```

AAA: 268
BBB: 251
CCC: 255
DDD: 251
EEE: 259
FFF: 254
GGG: 257

All the modules are approximately the same length. How about the final result by module percentage?

--Find percentage passed per module

```
SELECT code_module,  
(SELECT SUM(final_result)  
FROM studentInfo  
WHERE code_module = 'AAA'  
GROUP BY code_module) * 100 /  
(SELECT COUNT(final_result)  
FROM studentInfo  
WHERE code_module = 'AAA'  
GROUP BY code_module) AS 'Percentage'  
FROM studentInfo  
WHERE code_module = 'AAA'  
GROUP BY code_module
```

AAA: 65%

BBB: 38%

CCC: 26%

DDD: 35%

EEE: 44%

FFF: 38%

GGG: 44%

Overall, CCC is the module with the lowest score for all the UK. It doesn't seem to be a regional issue here, but more of a nationwide issue.

IMPLEMENTATION 3: Improve upon the CCC module nationwide by reviewing materials related to the course. Since the low scores in CCC are not isolated, it is most likely the fault of the curriculum rather than instructor or student.

To review: The board of education in the UK is advised to make 3 changes to bring about an increase in passing test scores.

1. Open or improve upon a nationwide tutoring service
2. Lower the number of required credits to be a full time student
3. Review CCC module for curriculum related issues

If the UK were to implement this and test scores increase to 45% or higher 1 year after implementation, then this project would be considered a success. The purpose of this project was to show that the author, Allie Wicklund, is capable of using SQL to make data-driven decisions and can manipulate complex databases. Also, if this data were up to date, the

implementations could be used to improve the educational process for the United Kingdom and would affect all university students there. This project may be done, but there's always much to learn. Would the conclusions be the same now in 2021? How will they change in the future? How does history affect the curriculum? Data isn't perfect, but it is sometimes useful.

Allie Wicklund is a data analyst from Michigan who has a strong background in applied mathematics. She is fluent in Python and SQL, and has experience with R, SAS, Tableau, and more. She has a blog about data at alliewicklund.com and can be reached at contact@alliewicklund.com. Feel free to say hi!