**1. Data used in the article:**

- File data: studentinfo

- File data: courses

**2. Tools used:**

- Excel (data source)

- SQL (EDA, merge and label data)

- Python (Jupyter Notebook) (Multiple Regression)

- Tableau (Visualization)

**3. Purpose of the lesson:**

Find the correlation between students' final results with age, disability, educational level, and length of the presentation module.

- General statistics on the number of students, and the distribution of information according to student information.

**4. Implementation method:**

- Quantitative method: multivariate regression

**5. Assumptions:**

- Each student has a different and unique student id.

**6. Initial comments:**

Looking at the data, I noticed great variability in final outcomes among students.

When I look at factors such as age, and disability.., I realize there seems to be a tendency to repeat like being able to write an equation whose dependent variable is final\_result.

Therefore, I proceed with the quantitative processing steps: multivariable regression with 1 dependent variable and 4 independent variables including age\_band, disability, module\_presentation\_length, and highest\_education.

My expectation is to be able to come up with a regression equation with a high R^2, based on which the course builder can think of building a profile of potential students for their courses. surname.

**7. Implementation steps:**

Step 1: I proceed to review the existing dataset. From there, look for files I can link by primary key to reach the expected conclusion. (Via ETL system image)

Step 2: After I have determined the exact dataset I want to use, I proceed to import the data into SQL to conduct data association and labeling.

(Note: Labeling data here, I want to convert variables with qualitative values ​​to quantification to serve step 4)

To serve the step of running the quantitative model, I have labeled the data in the SQL file in the UPDATE statements. Specifically, I would like to quote the corresponding commands:

Set age\_band = Case when age\_band = '0-35' Then '0'

when age\_band = '35-55' Then '1'

Else '2'

End

Update [EDA].[dbo].[studentInfo]

Set disability = Case when disability = 'N' Then '0'

Else '1'

End

Update [EDA].[dbo].[studentInfo]

Set highest\_education = Case when highest\_education = 'HE Qualification' Then '4'

when highest\_education = 'A Level or Equivalent' Then '2'

when highest\_education = 'Post Graduate Qualification' Then '3'

when highest\_education = 'Lower Than A Level' Then '1'

Else '0'

End

Update [EDA].[dbo].[studentInfo]

Set final\_result = Case when final\_result = 'Pass' Then '2'

when final\_result = 'Fail' Then '1'

Else '0'

Step 3: Export processed data to Excel

Step 4: Import data into Tableau to conduct data visualization

Step 5: Import data into Jupyter to perform multivariable linear regression

**8. New discoveries:**

The regression model results are not good because R^2 is very small, and RMSE is large. However, it shows the suitable diversity of the course for many different audiences.

Through data visualization, it can be seen that up to the time of implementation, students aged 0-35, without disabilities, with a level of education or equivalent are the students who make up the majority of students. students take courses. And the CCC - 2014J course has the highest number of students.

**9. Conclusion:**

The courses are diverse and suitable for many different types of students. However, students with a high level of education have not yet been attracted to the courses.

The courses have a good degree of grading, which is demonstrated by a course completion rate that is not overwhelming compared to the failure rate. Another problem raised here is that the rate of students dropping out of school is quite high, 27.97% of the total number of students participating. As a result, student retention rates for courses are low. It is advisable to build learning attraction in courses or reduce difficulty in courses to increase student engagement.

**10. The desired idea is realized:**

- Analyze the factors affecting the students' abandonment of the course.

- Portrait of subjects who tend to give up the course

(However, there was a problem during the test, I found out that the number of student ids in student info is higher than the number of student ids in student assessment. This is also the reason why I need more time to do it. consider when choosing the above ideas as key insights)