**DATA ANALYSIS USING POWER BI**

(*DATA VISUALIZATION AND DATA CLEANING USING power bi)*

DATASET: : <https://www.kaggle.com/rocki37/open-university-learning-analytics-dataset>

Name - Ritesh Prajapati

SRN - R21DG034

Git-hub - https://github.com/Rishhh20

1. **INTRODUCTION TO DATASET:**

This dataset is provided by the learning analytics research group at the Knowledge Media institute, The Open University. The dataset consists of tables with information on student demographics, modules undertaken, time of year the modules start (module presentations), and information on student academic success in terms of grades for assignments and exams, as well as students’ interactions with the university’s Virtual Learning Environment (VLE).

Dataset has so many missing values and there are inconsistencies in the columns. we need to clean the data and inconsistent data should be reported.

This dataset offers two of the elements in the framework: behavior and performance. It contains information about 22 courses, 32,593 students, their assessment results, and logs of their interactions with the VLE represented by daily summaries of student clicks (10,655,280 entries)

**2. DATA OVERVIEW:**

Our data have some missing values, for example

1. Assessment: In this module, in the Date column, there is 5.339806% of missing data

2. Student Assessment: In the Score column, there is 0.099476% of data is missing

3. Student Info: In this dataset, in column imd\_band there is 3.4080707% of data is missing

4. Student Registration: In this, there are 2 columns where the data are missing

a)date\_registration: 0.138066% of missing data

b)data\_unregistration: 69.097659%

5. Vle: week\_from:- 82.3852% missing data

Week\_to:- 82.3853% missing data

**DATA ANALYSIS OF DATASETS :**

1. **ASSESSMENT:-**

Percentage of missing data

|  |  |  |
| --- | --- | --- |
| columns | % of data missing | dtype |
| Code\_module | 0 | object |
| Code\_presentation | 0 | object |
| Id\_assessment | 0 | Int64 |
| Assessment\_type | 0 | object |
| date | 5.339806 | Float 64 |
| weight | 0 | Object 64 |

1. Changed data from int type to object data type
2. Most assessments have 200 total weight

Except for code module CCC -300 and GGG- 100 total weights

100 marks for the Exam.

Except CCC 200 because it has 2 exams

100 marks for CMA+TMA,except for GGG and GGG have 0marks CMA+TMA

|  |  |  |
| --- | --- | --- |
|  | CCC | GGG |
| CMA+TMA | 100 | 0 |
| EXAM | 100 | 100 |
| EXAM | 100 | - |

1. Since CMA is often weighted 0, we will just assign 100 total weight to TMA

CMA=0

TMA=100

Exam=100

4. Check if assessment info is in the Results table

->Compare (assessments, results)

True-188

False-18

18 assignments missing from the results.

Some find exams are missing from the results table

**2.RESULTS:**

|  |  |  |
| --- | --- | --- |
| COLUMNS |  |  |
| Id\_assessment | 0 | Int 64 |
| Id\_student | 0 | Int 64 |
| Date\_submitted | 0 | Int 64 |
| Is\_banked | 0 | Int 64 |
| score | 0.99476 | float 64 |

1. Convert id\_assessment, from int to object dtype

Convert id\_student, from int to object dtype

1. If the result is empty, the assignment is not submitted, so fill the empty cell space with 0.

**3.COURSES:**

|  |  |  |
| --- | --- | --- |
| Code\_module | 0 | object |
| Code\_presentation | 0 | object |
| length | 0 | Int64 |

As no data is missing, here we can move onto next table.

**4.STUDENT REGISTRATION:**

|  |  |  |
| --- | --- | --- |
| code\_module | 0 | Object |
| Code\_presentation | 0 | Object |
| Id\_student | 0 | Int64 |
| Date\_registration | 0.1380 | Float64 |
| Date\_unregistration | 69.09765 | Float64 |

1.convert id\_student to object

Compare if all values in registration table are recorded in the results table.

Compare (org,results)

True:- 26746

False:-5847

There are 5847 students missing from the results table

Are there any student from student\_info table missing from results table

Compare column(info,results)

True:- 26746

False:- 5847

5847students are the same students

Fail:1197

Pass:2

Withdrawn:4648

Only 2 are passed,because of data entry mistake and some data

**5.VLE(Virtual Learning Environment ) resources:**

|  |  |  |
| --- | --- | --- |
| Id\_site | 0 | Int64 |
| code\_module | 0 | object |
| Code\_presentation | 0 | object |
| Activity\_type | 0 | object |
| Week\_from | 82.385292 | Float64 |
| Week\_to | 82.385292 | Float64 |

1.convert id\_site int dtype into object dtype.

**6.VLE INTERATION:**

|  |  |  |
| --- | --- | --- |
| Code\_module | 0 | object |
| Code\_presentation | 0 | object |
| Id\_student | 0 | Int64 |
| Id\_site | 0 | Int64 |
| date | 0 | Int64 |
| Sum\_click | 0 | Int64 |

Id\_student=>object type

Id\_site=>object type

**7.STUDENT INFORMATION:**

|  |  |  |
| --- | --- | --- |
| code\_module | 0 | Object |
| Code\_presentation | 0 | Object |
| id\_student | 0 | Int64 |
| Gender | 0 | Object |
| Region | 0 | object |
| Highest\_education | 0 | object |
| Imd\_band | 3.408707 | object |
| Age\_band | 0 | object |
| No of previous attempt | 0 | Int64 |
| Studied\_credicts | 0 | Int64 |
| Disability | 0 | object |
| Final\_result | 0 | object |

Convert id\_student from int 64 dtype to object dtype.

**MERGE TABLE AND FEATURE ENGINEERING:**

**1.VLE AND MATERIALS**

Vle and material columns are merged and named as vlematerials.

Here dropped columns are week from and week to.Because 82% of the data is empty. These data was not much helpful for our analysis.

Add column=calculated total click per student

Preprocessing is done.

Total\_clicks=no of clicks per student

**2.Reg courses info:**

Student registration and courses are merged together and gives RegCourses. Then REgCourses and Student info is merged and gives RegCoursesInfo.

**3.AssResults:**

Assessment and results are merged together and gives AssResults.

**4.Scores=AssResults**

Sum\_scores=weight\*score

Total\_weight=Total\_weight[‘total weight’]-100

Then, sum\_scores and total\_weight gives score\_weight.

Weighted\_score=score/total\_weight

Dropped Columns:-

Is\_banked

Date\_submitted and

Assessment\_type

**5.late\_rate:**

Late\_rate=total\_late\_submission/total\_assessment

**6.fail\_rate:**

Fail\_rate=total\_fail/total\_assessment

**MERGE ALL THE TABLES**

1.VLE +VLE Materials=total\_click

2. Reg+Courses+Student\_info=RegCoursesInfo

3.Assessment+Results=AssResults

Here total\_click and RegCoursesInfo are merged.

Then AssResults are merged with it.

**FILLING MISSING DATA**

**1.imd\_band:**

Replaced null values with most frequent imd\_bands

**2.date\_registration:**

the median of date\_registration= -57

#date unregistered-median

#fill remaining values with -57

**3.total\_clicks:**

If the value is missing means the student didn’t click any material.so fill missing value with 0.

**4.weighted\_score:**

Weighted score is empty means he didn’t submit any assignment.So replace NaN value with 0.

**5.late\_rate:**

Nan, this column means they have not submitted any assignments,so replace it with 1.then it will be having 100% late\_rate

**6.fail\_rate:**

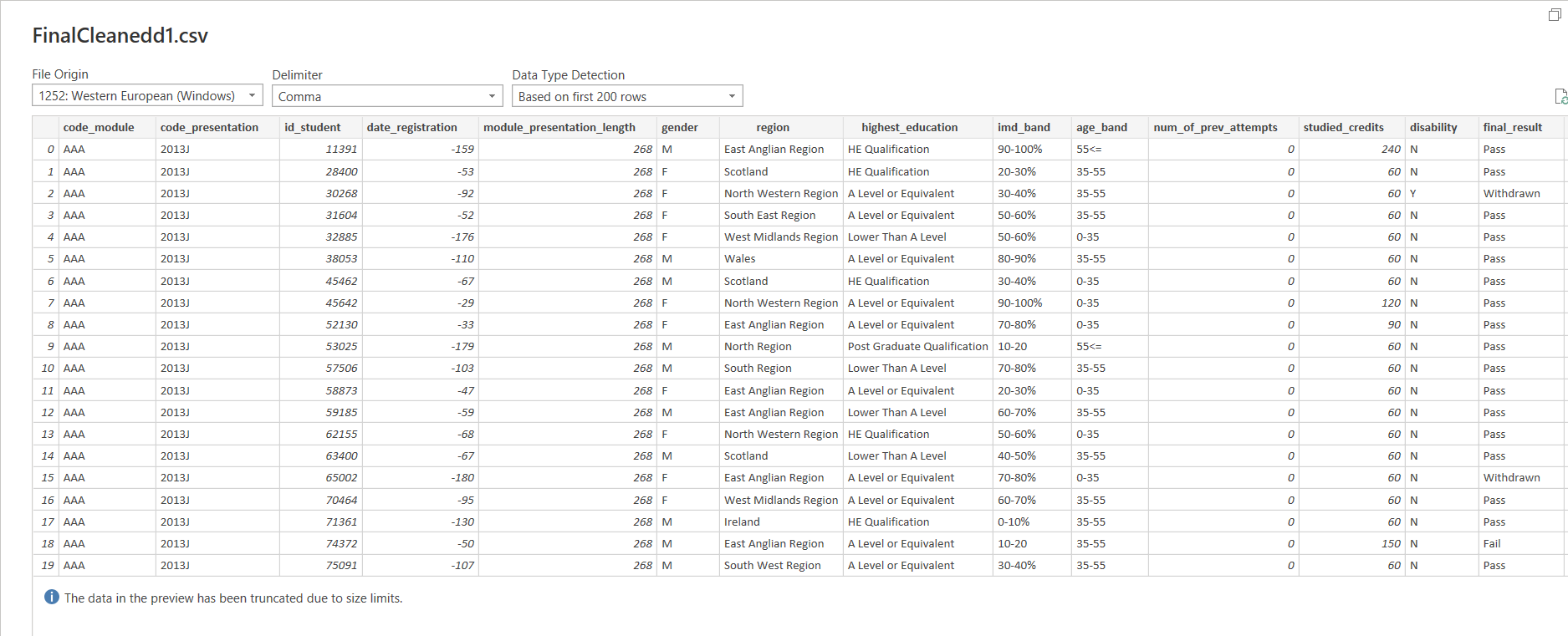
If this columns is empty then student didn’t submit any assignment,it means 100% fail\_rate.fill it with 1.

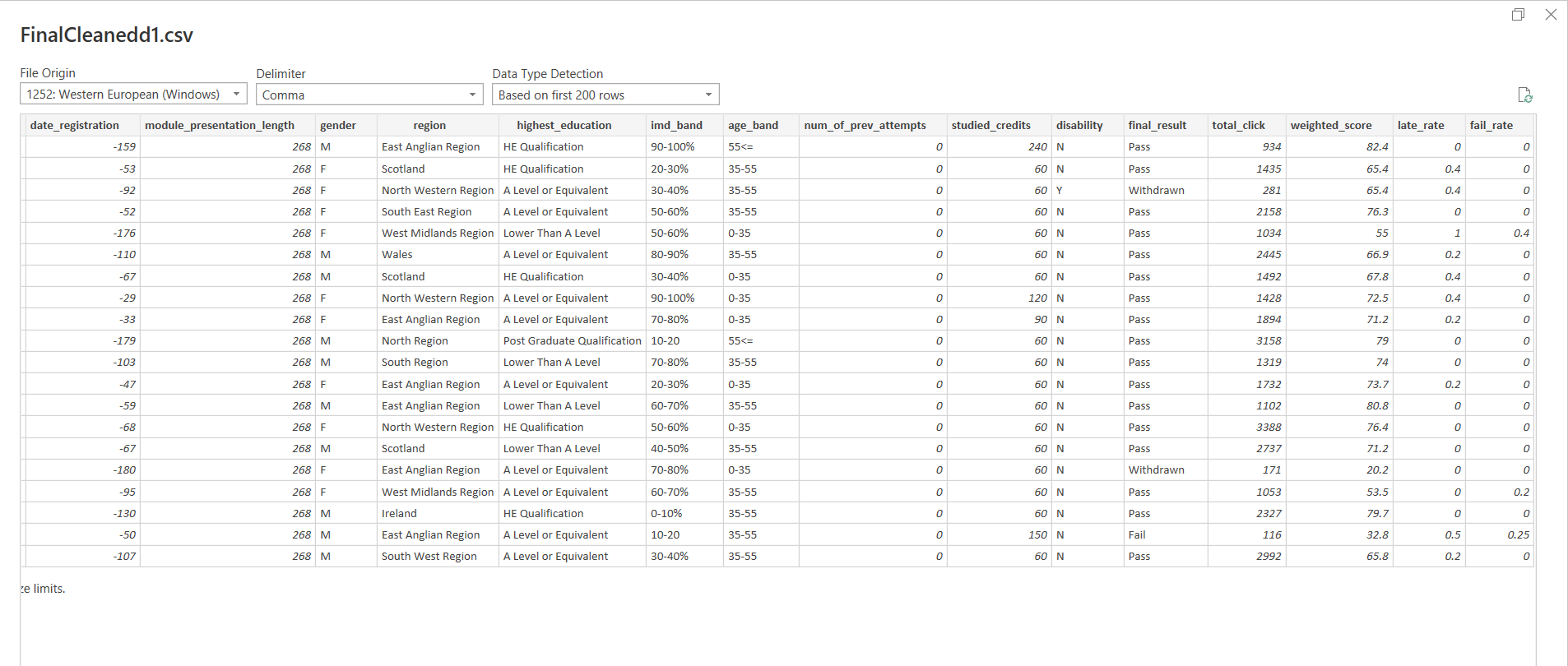
Drop date\_unregistered column.

Statistical analytics tools provide various ways for reorganizing raw data to see new patterns by calculating characteristics such as averages, frequencies, variations, rankings, ranges and deviations.

Descriptive statistics consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution. They include both numerical (e.g. central tendency measures such as mean, mode, median or measures of variability) and graphical tools (e.g. histogram, box plot, scatter plot…) which give a summary of the dataset and extract important information such as central tendencies and variability.

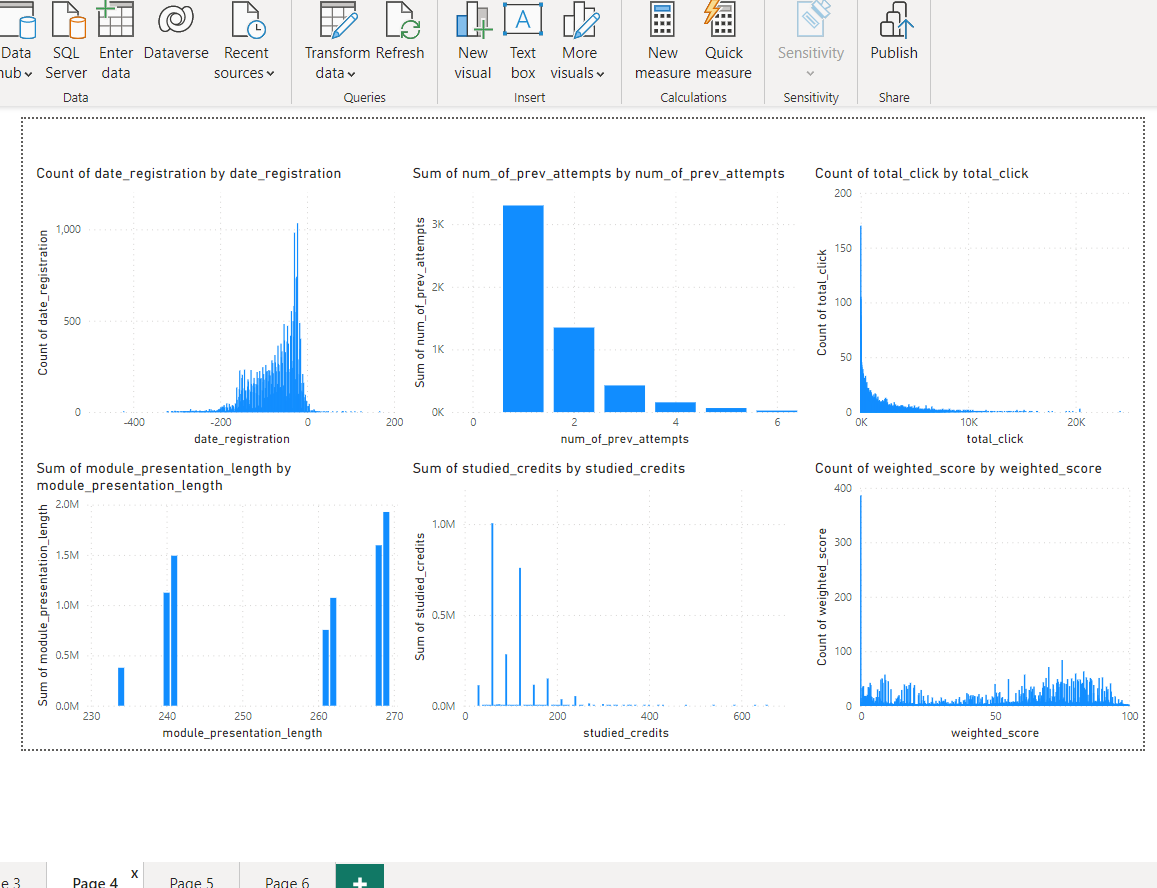
**FINAL CLEANING OF DATASET AND ALL ARE MERGED**

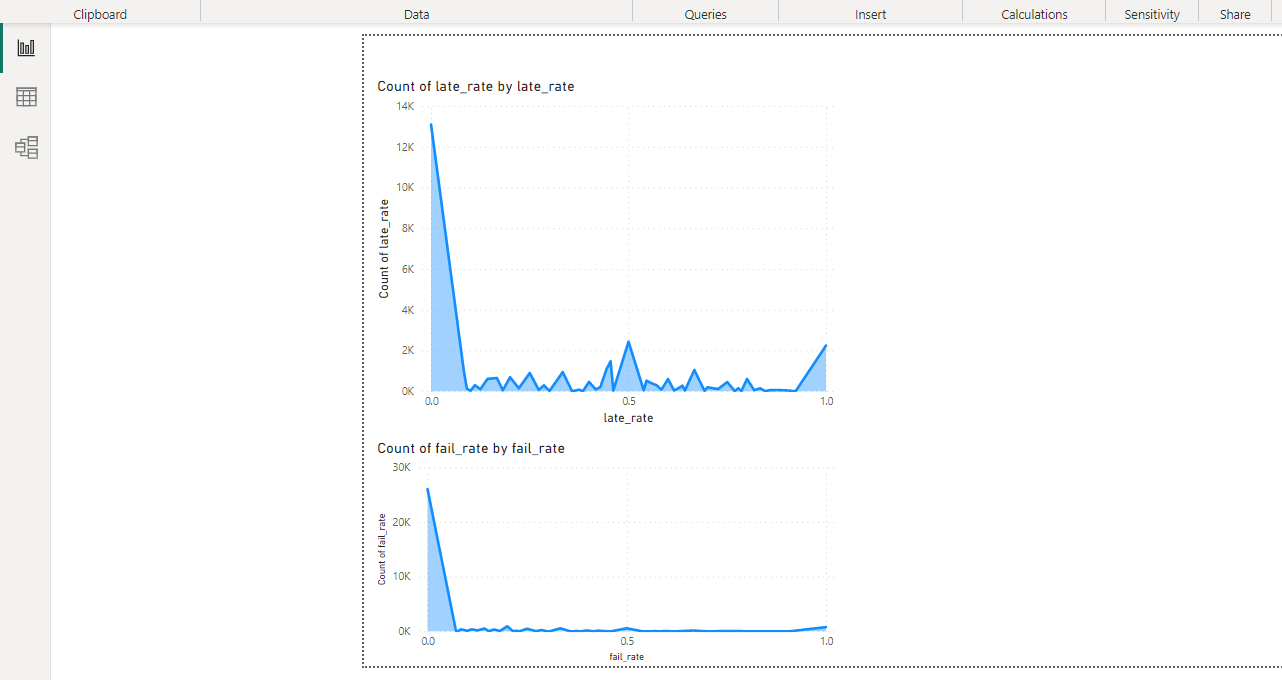




**DATA VISUALIZATION**

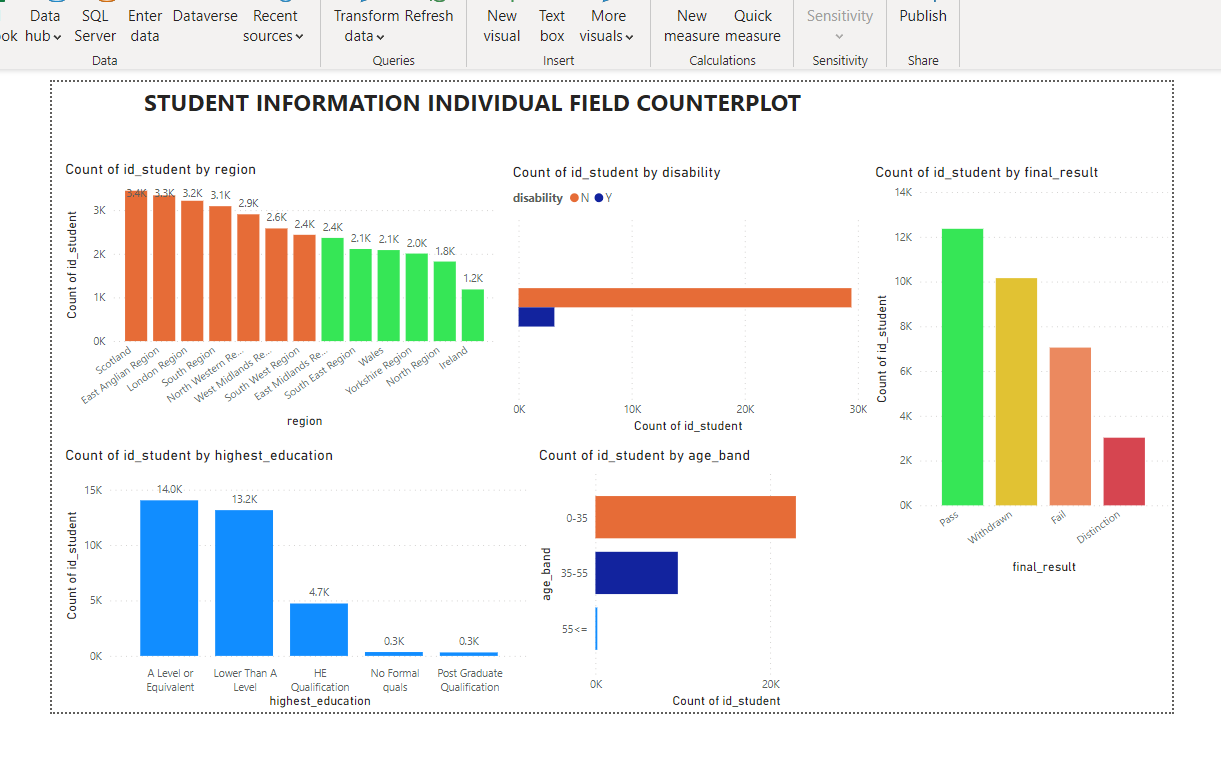
**1.DESCRIPTIVE ANALYSIS:**it uses data aggregation and data mining to provide insight into the past and Answers :”What has happened?”The descriptive analytics does exactly what name implies they “describe” or summarize raw data and make it interpretable .

****



2.DIAGNOSTIC ANALYTICS:determines why something happened in past.takes deeper look at data to understand the root cause of the event.

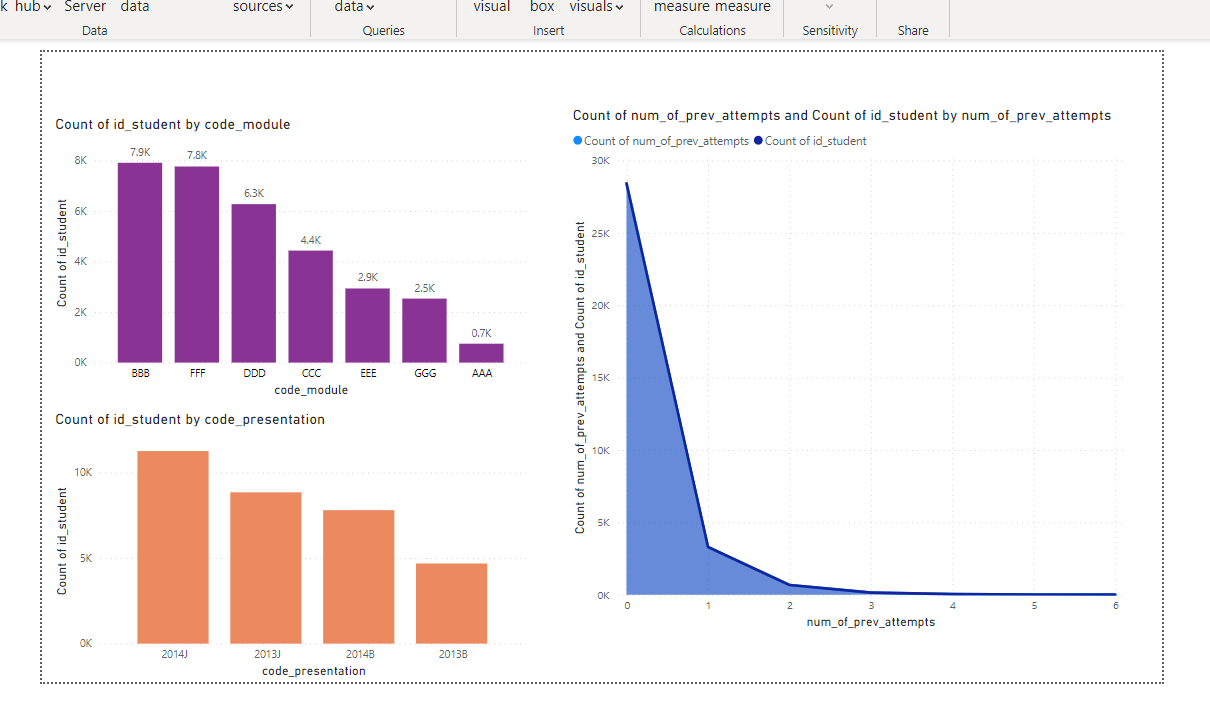
**STUDENT INFORMATION INDIVIDUAL FILED COUNTERPLOT**

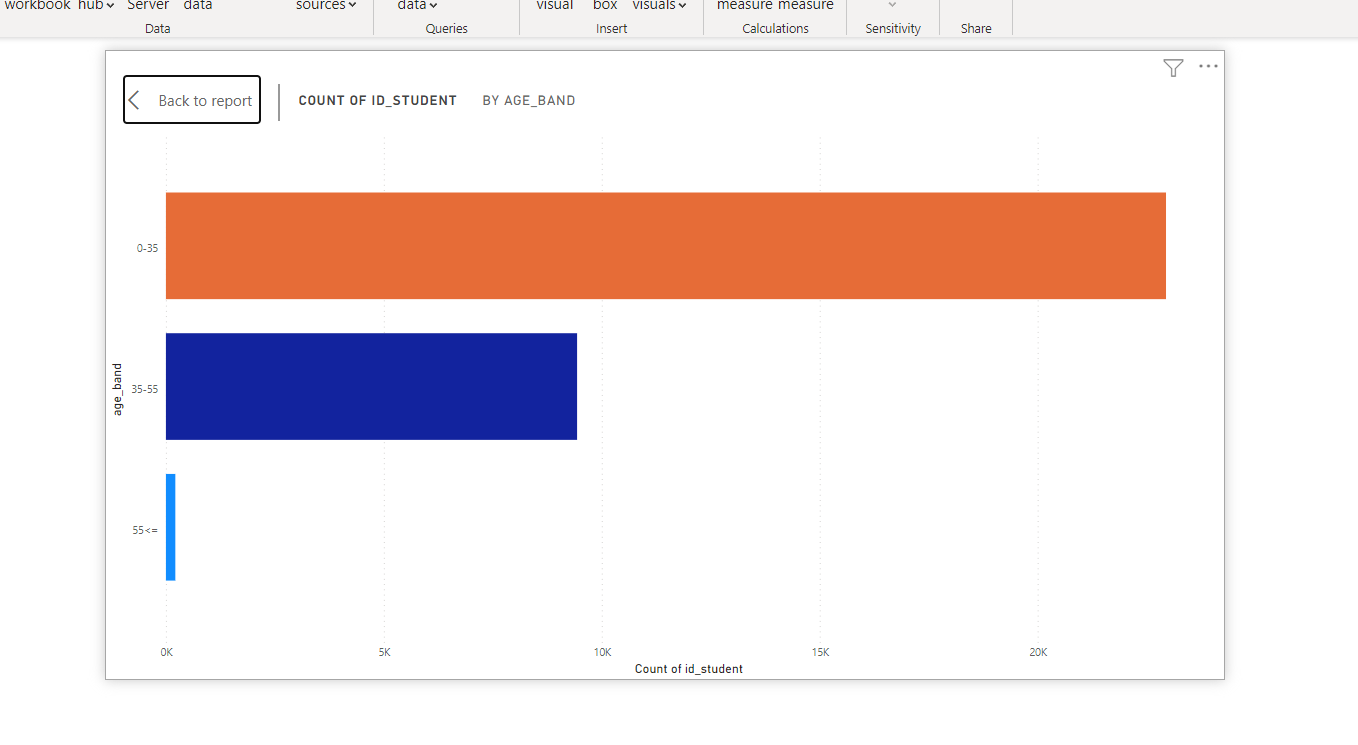


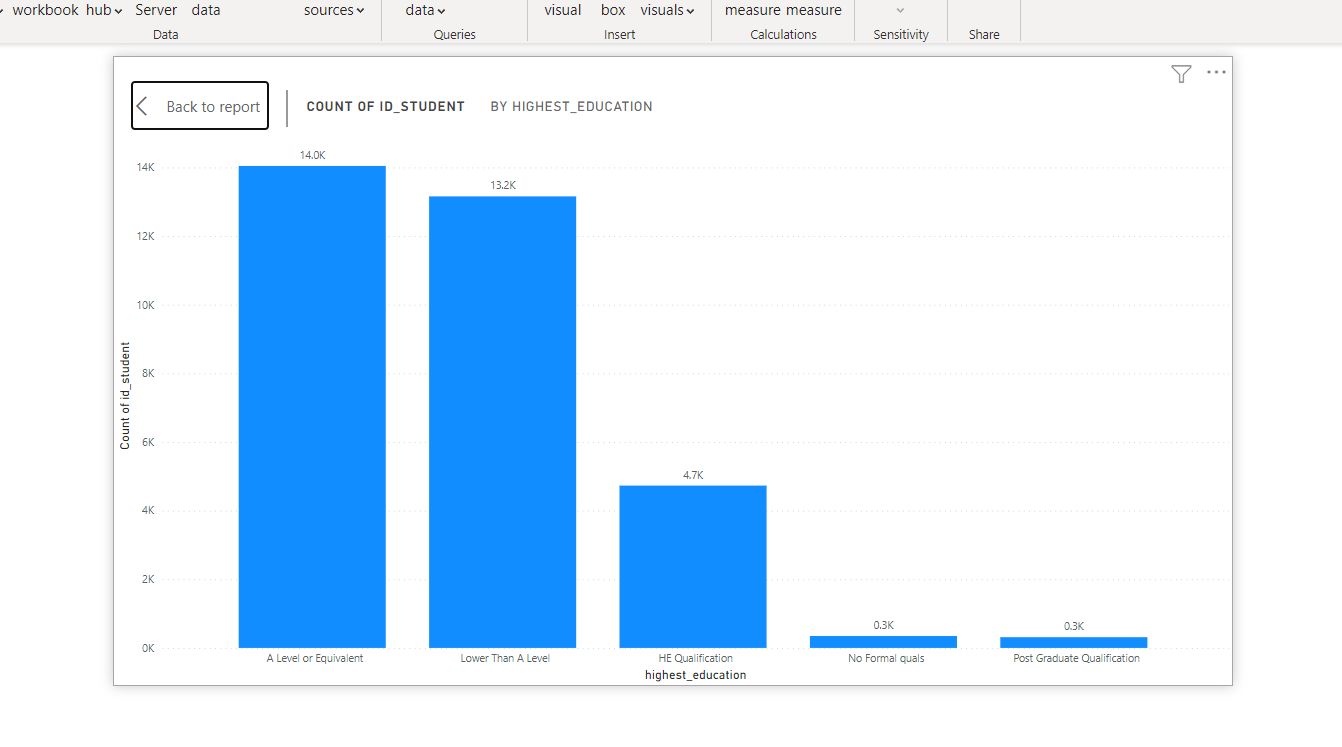
3.PREDICTIVE ANALYSIS:

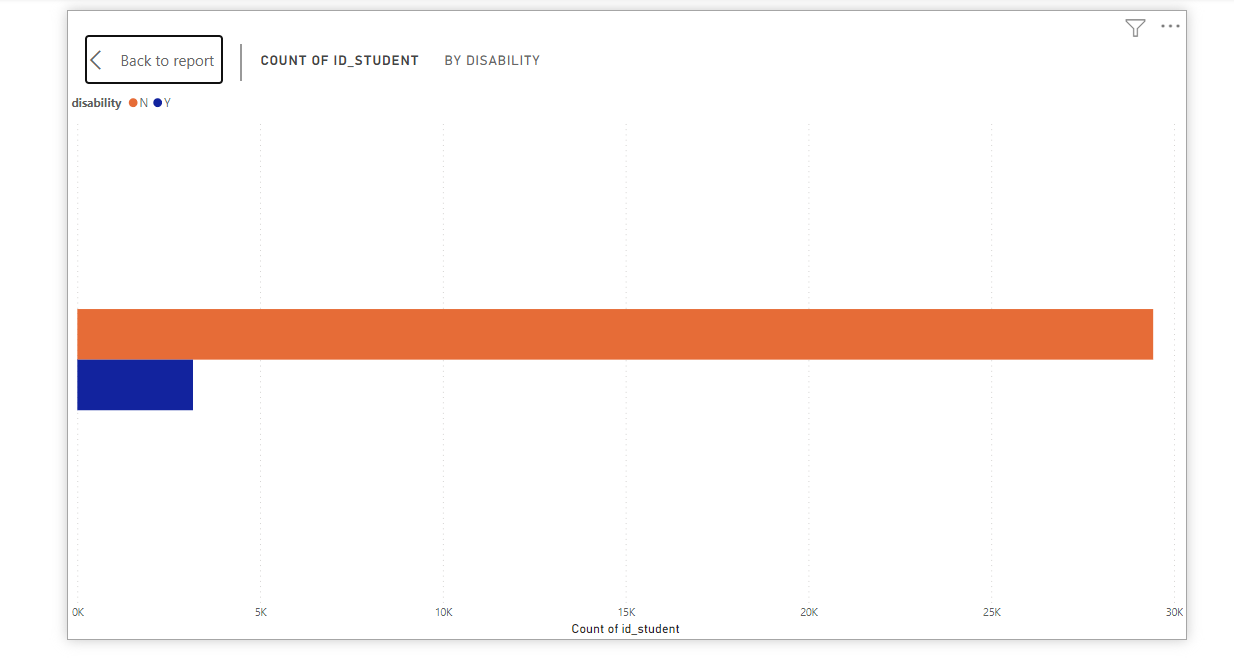
It uses statistical models and forecasts techniques to understand the future and Answers “what could happen?”

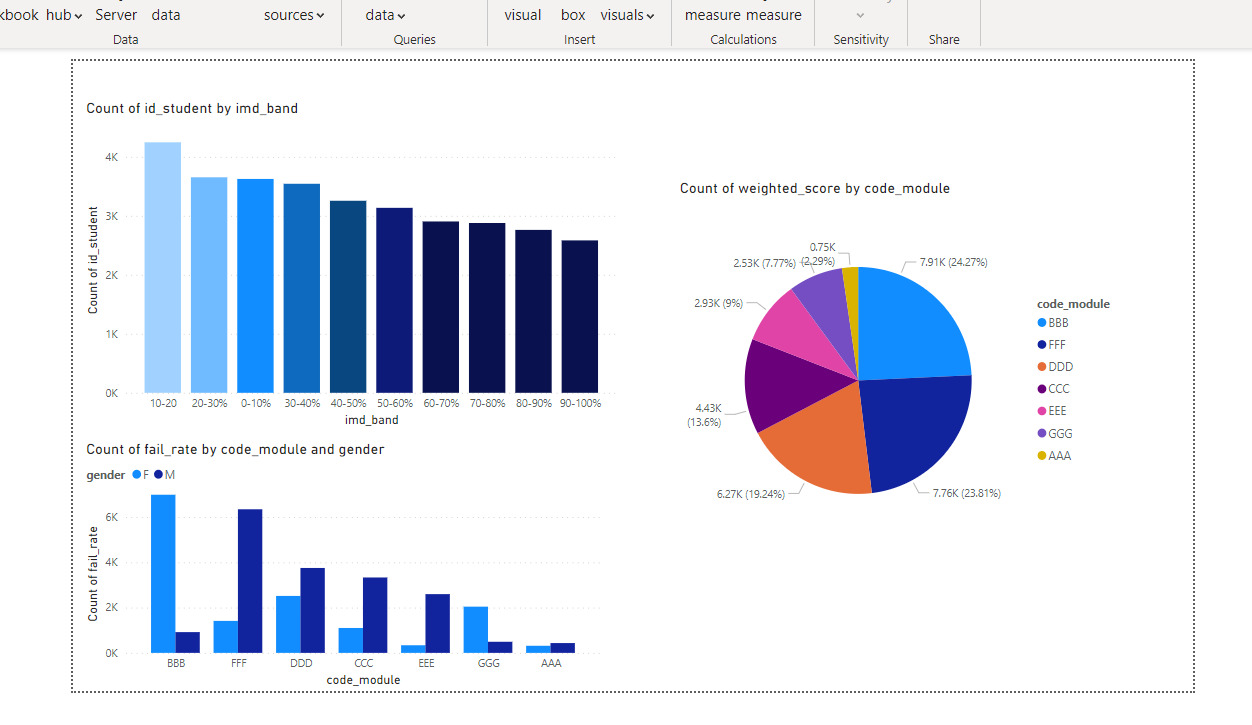
Provides actionable insights based on data. And it provides estimates about the likelihood of a future outcomes.







Overall conclusion so far We have seen that number there is a relationship between number of clicks and final result, and particularly that it might be possible at the lower end of engagement to predict withdrawal or failure from VLE engagement. However our independent variables by themselves don't seem to be predictive of mean assessment score.



And final it has prescriptive analytics uses optimization and simulation algorithms to advice on possible outcomes and answers:what should we do?

It allows users to prescribe a number of different possible actions and guide towards a solution.

In a nutshell, this analytics is all providing advice.

Statistical analysis is the process of collecting and analyzing data in order to discern patterns and trends. It is a method for removing bias from evaluating data by employing numerical analysis. This technique is useful for collecting the interpretations of research, developing statistical models, and planning surveys and studies.

Statistical analysis is a scientific tool that helps collect and analyze large amounts of data to identify common patterns and trends to convert them into meaningful information. In simple words, statistical analysis is a data analysis tool that helps draw meaningful conclusions from raw and unstructured data.

The conclusions are drawn using statistical analysis facilitating decisionmaking and helping businesses make future predictions on the basis of past trends. It can be defined as the science of collecting and analyzing data to identify trends and patterns and presenting them. Statistical analysis involves working with numbers and is used by businesses and other institutions to make use of data to derive meaningful information.