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	-

3. 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

2. 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:

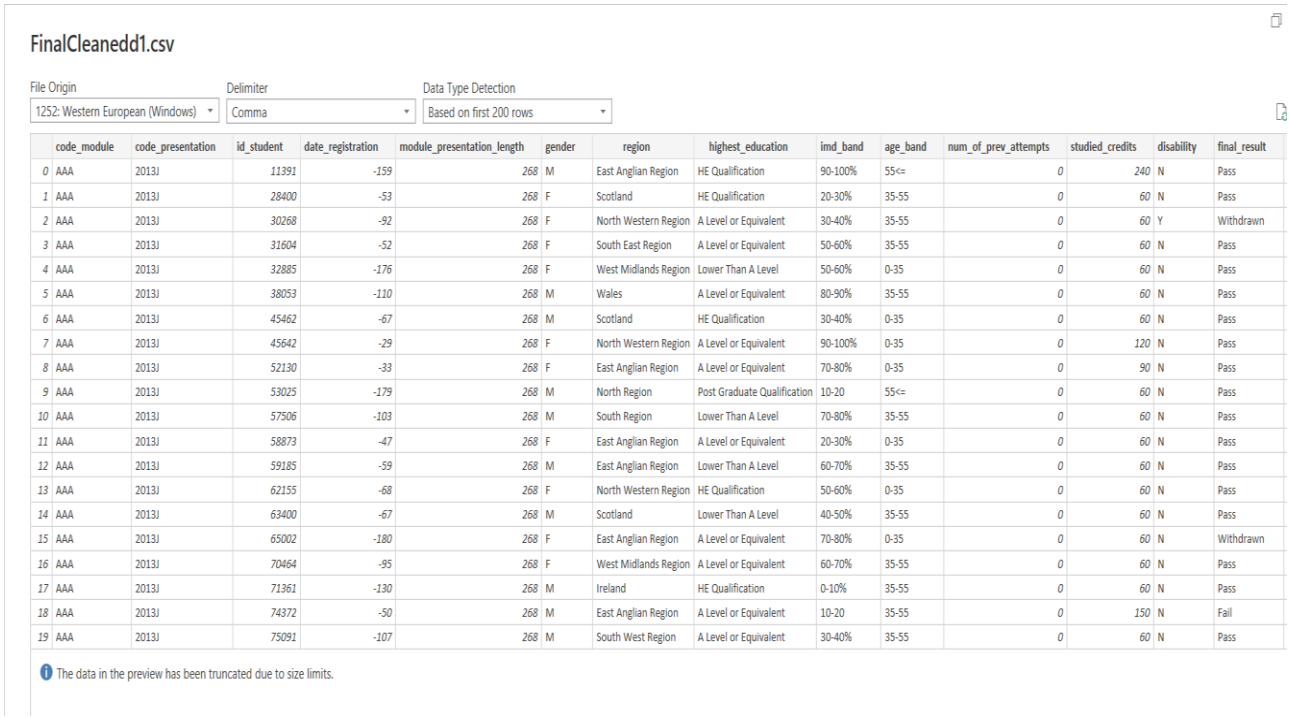
If this columns is empty then student didn’t submit any assignment,it means 100% fail_rate.fill it with 1.

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



FinalCleanedd1.csv

File Origin: 1252: Western European (Windows) | Delimiter: Comma | Data Type Detection: Based on first 200 rows

	code_module	code_presentation	id_student	date_registration	module_presentation_length	gender	region	highest_education	imd_band	age_band	num_of_prev_attempts	studied_credits	disability	final_result
0	AAA	2013J	11391	-159	268	M	East Anglian Region	HE Qualification	90-100%	55<=	0	240	N	Pass
1	AAA	2013J	28400	-53	268	F	Scotland	HE Qualification	20-30%	35-55	0	60	N	Pass
2	AAA	2013J	30268	-92	268	F	North Western Region	A Level or Equivalent	30-40%	35-55	0	60	Y	Withdrawn
3	AAA	2013J	31604	-52	268	F	South East Region	A Level or Equivalent	50-60%	35-55	0	60	N	Pass
4	AAA	2013J	32885	-176	268	F	West Midlands Region	Lower Than A Level	50-60%	0-35	0	60	N	Pass
5	AAA	2013J	38053	-110	268	M	Wales	A Level or Equivalent	80-90%	35-55	0	60	N	Pass
6	AAA	2013J	45462	-67	268	M	Scotland	HE Qualification	30-40%	0-35	0	60	N	Pass
7	AAA	2013J	45642	-29	268	F	North Western Region	A Level or Equivalent	90-100%	0-35	0	120	N	Pass
8	AAA	2013J	52130	-33	268	F	East Anglian Region	A Level or Equivalent	70-80%	0-35	0	90	N	Pass
9	AAA	2013J	53025	-179	268	M	North Region	Post Graduate Qualification	10-20	55<=	0	60	N	Pass
10	AAA	2013J	57506	-103	268	M	South Region	Lower Than A Level	70-80%	35-55	0	60	N	Pass
11	AAA	2013J	58873	-47	268	F	East Anglian Region	A Level or Equivalent	20-30%	0-35	0	60	N	Pass
12	AAA	2013J	59185	-59	268	M	East Anglian Region	Lower Than A Level	60-70%	35-55	0	60	N	Pass
13	AAA	2013J	62155	-68	268	F	North Western Region	HE Qualification	50-60%	0-35	0	60	N	Pass
14	AAA	2013J	63400	-67	268	M	Scotland	Lower Than A Level	40-50%	35-55	0	60	N	Pass
15	AAA	2013J	65002	-180	268	F	East Anglian Region	A Level or Equivalent	70-80%	0-35	0	60	N	Withdrawn
16	AAA	2013J	70464	-95	268	F	West Midlands Region	A Level or Equivalent	60-70%	35-55	0	60	N	Pass
17	AAA	2013J	71361	-130	268	M	Ireland	HE Qualification	0-10%	35-55	0	60	N	Pass
18	AAA	2013J	74372	-50	268	M	East Anglian Region	A Level or Equivalent	10-20	35-55	0	150	N	Fail
19	AAA	2013J	75091	-107	268	M	South West Region	A Level or Equivalent	30-40%	35-55	0	60	N	Pass

The data in the preview has been truncated due to size limits.

FinalCleanedd1.csv

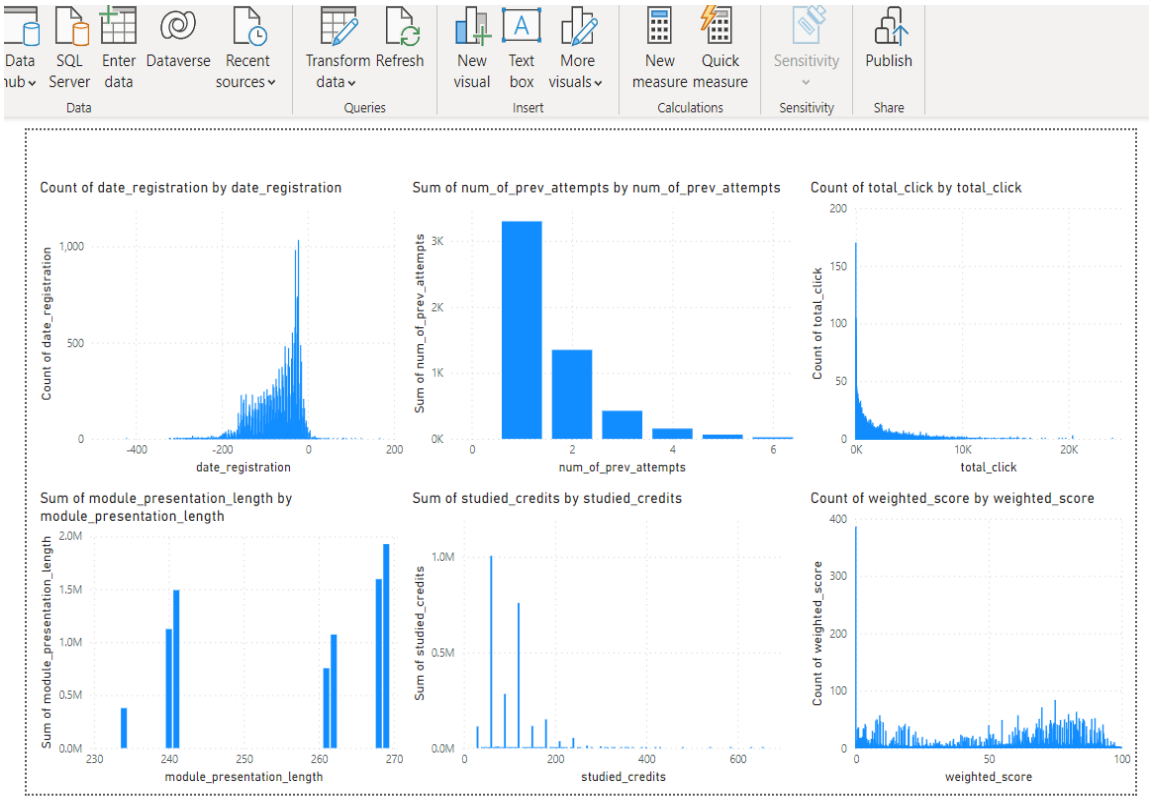
File Origin: 1252: Western European (Windows) | Delimiter: Comma | Data Type Detection: Based on first 200 rows

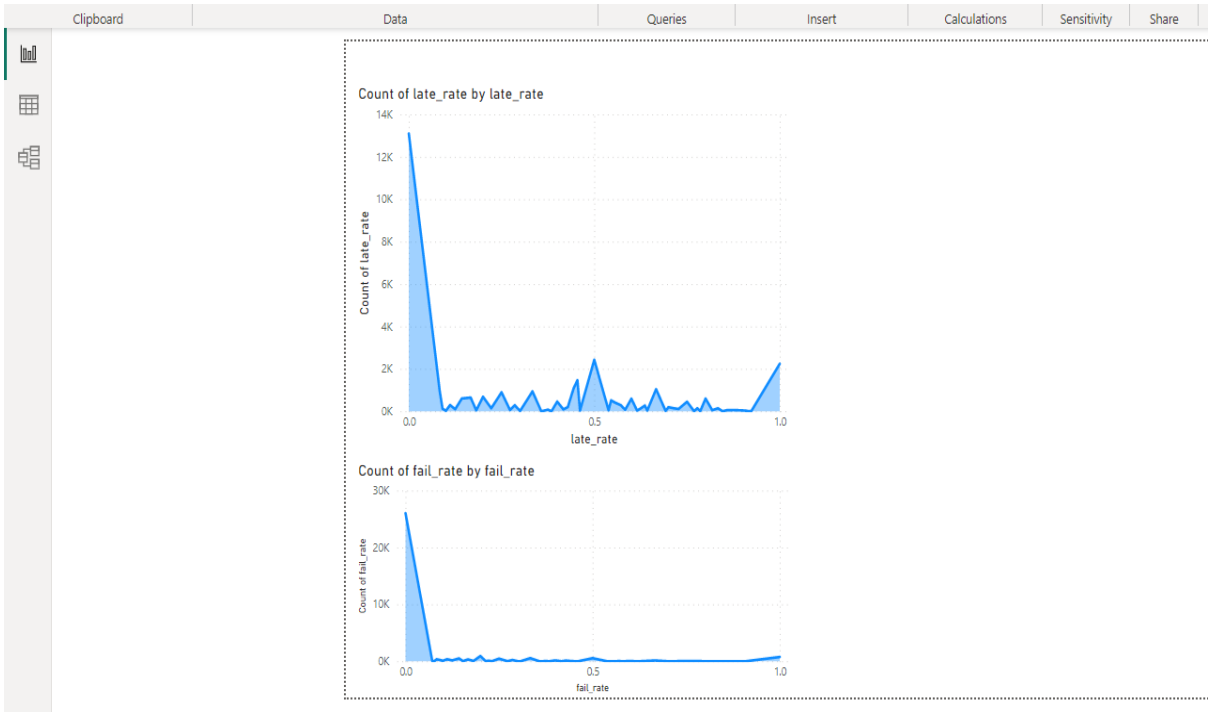
date_registration	module_presentation_length	gender	region	highest_education	imd_band	age_band	num_of_prev_attempts	studied_credits	disability	final_result	total_click	weighted_score	late_rate	fail_rate
-159	268	M	East Anglian Region	HE Qualification	90-100%	55<=	0	240	N	Pass	934	82.4	0	0
-53	268	F	Scotland	HE Qualification	20-30%	35-55	0	60	N	Pass	1435	65.4	0.4	0
-92	268	F	North Western Region	A Level or Equivalent	30-40%	35-55	0	60	Y	Withdrawn	281	65.4	0.4	0
-52	268	F	South East Region	A Level or Equivalent	50-60%	35-55	0	60	N	Pass	2158	76.3	0	0
-176	268	F	West Midlands Region	Lower Than A Level	50-60%	0-35	0	60	N	Pass	1034	55	1	0.4
-110	268	M	Wales	A Level or Equivalent	80-90%	35-55	0	60	N	Pass	2445	66.9	0.2	0
-67	268	M	Scotland	HE Qualification	30-40%	0-35	0	60	N	Pass	1492	67.8	0.4	0
-29	268	F	North Western Region	A Level or Equivalent	90-100%	0-35	0	120	N	Pass	1428	72.5	0.4	0
-33	268	F	East Anglian Region	A Level or Equivalent	70-80%	0-35	0	90	N	Pass	1894	71.2	0.2	0
-179	268	M	North Region	Post Graduate Qualification	10-20	55<=	0	60	N	Pass	3158	79	0	0
-103	268	M	South Region	Lower Than A Level	70-80%	35-55	0	60	N	Pass	1319	74	0	0
-47	268	F	East Anglian Region	A Level or Equivalent	20-30%	0-35	0	60	N	Pass	1732	73.7	0.2	0
-59	268	M	East Anglian Region	Lower Than A Level	60-70%	35-55	0	60	N	Pass	1102	80.8	0	0
-68	268	F	North Western Region	HE Qualification	50-60%	0-35	0	60	N	Pass	3388	76.4	0	0
-67	268	M	Scotland	Lower Than A Level	40-50%	35-55	0	60	N	Pass	2737	71.2	0	0
-180	268	F	East Anglian Region	A Level or Equivalent	70-80%	0-35	0	60	N	Withdrawn	171	20.2	0	0
-95	268	F	West Midlands Region	A Level or Equivalent	60-70%	35-55	0	60	N	Pass	1053	53.5	0	0.2
-130	268	M	Ireland	HE Qualification	0-10%	35-55	0	60	N	Pass	2327	79.7	0	0
-50	268	M	East Anglian Region	A Level or Equivalent	10-20	35-55	0	150	N	Fail	116	32.8	0.5	0.25
-107	268	M	South West Region	A Level or Equivalent	30-40%	35-55	0	60	N	Pass	2992	65.8	0.2	0

e limits.

DATA VISUALIZATION

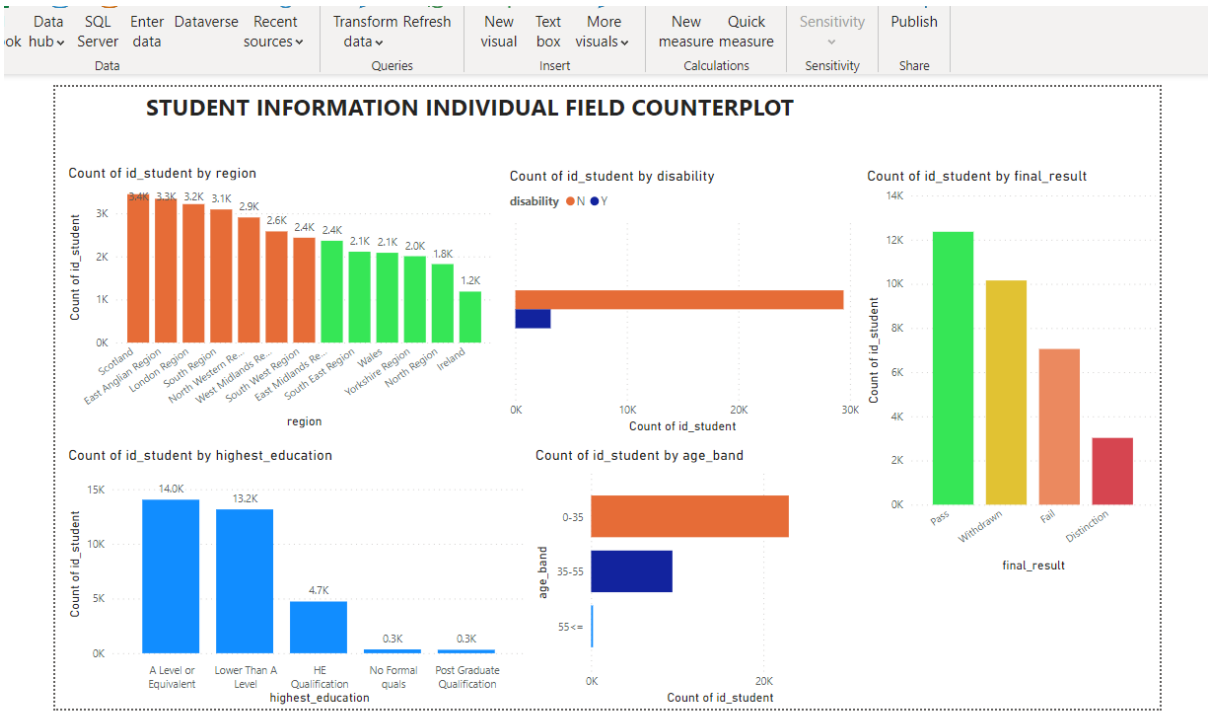
1.DESCRIPTIVE ANALYSIS:it uses data aggregation and data mining to provide insight into the past and Answers :"What has happened?"The descriptive analytics does exactly what name implies they "describe" or summarize raw data and make it interpretable .





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

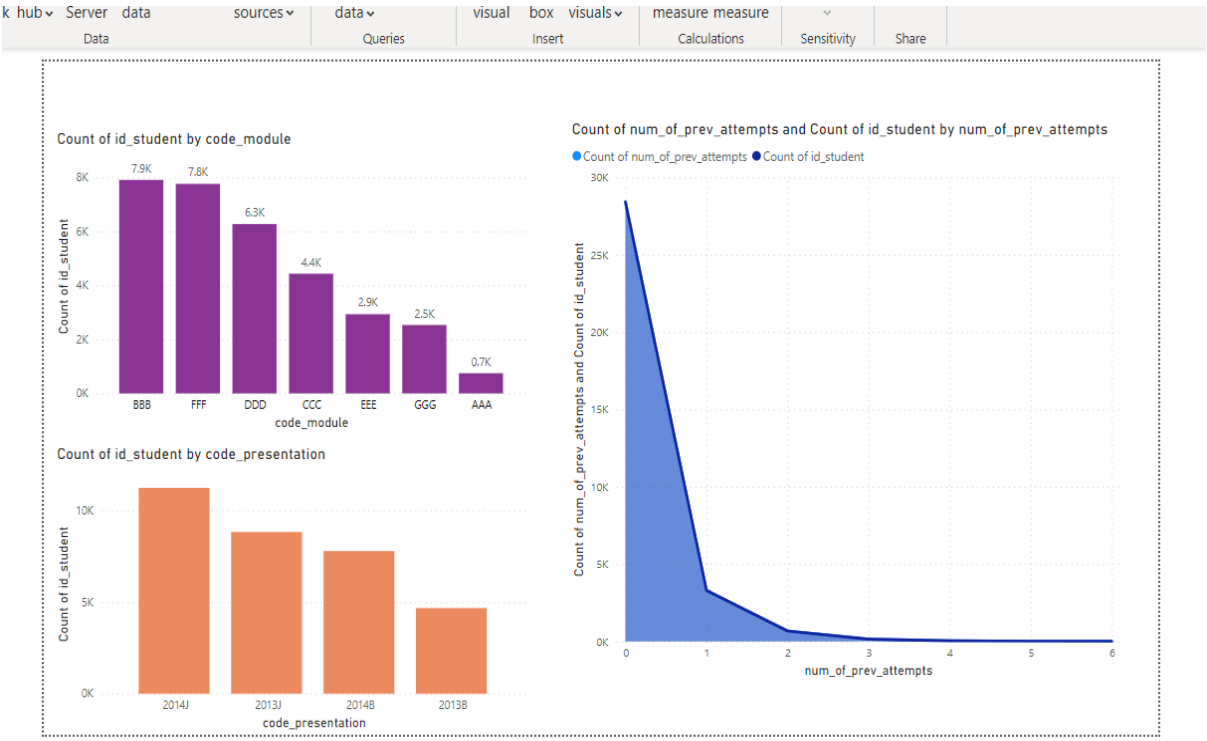
STUDENT INFORMATION INDIVIDUAL FILED COUNTERPLOT

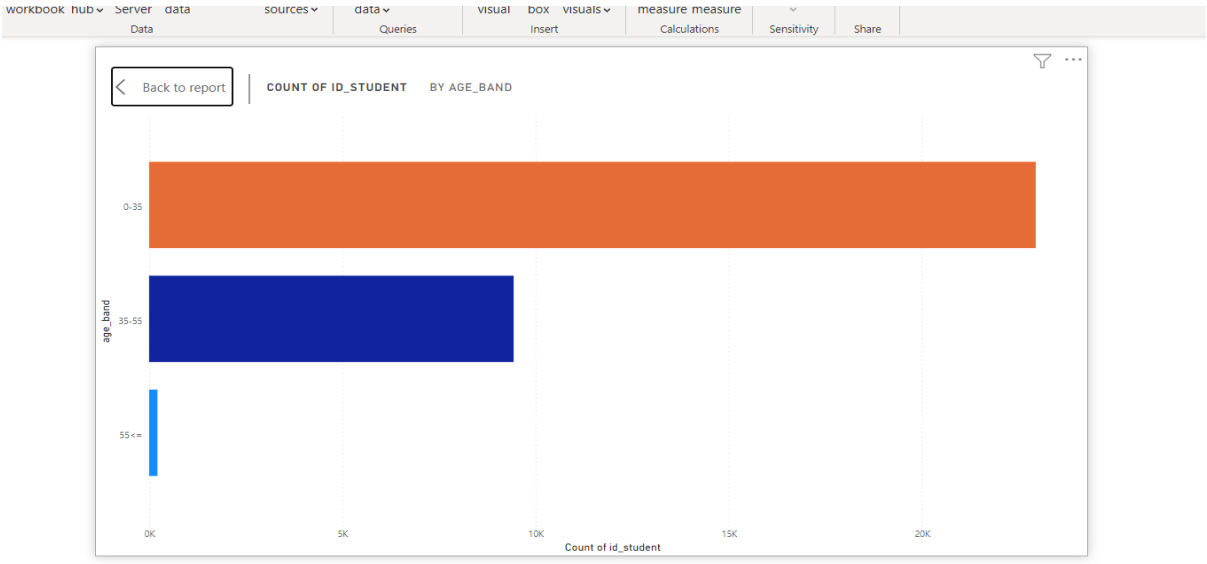


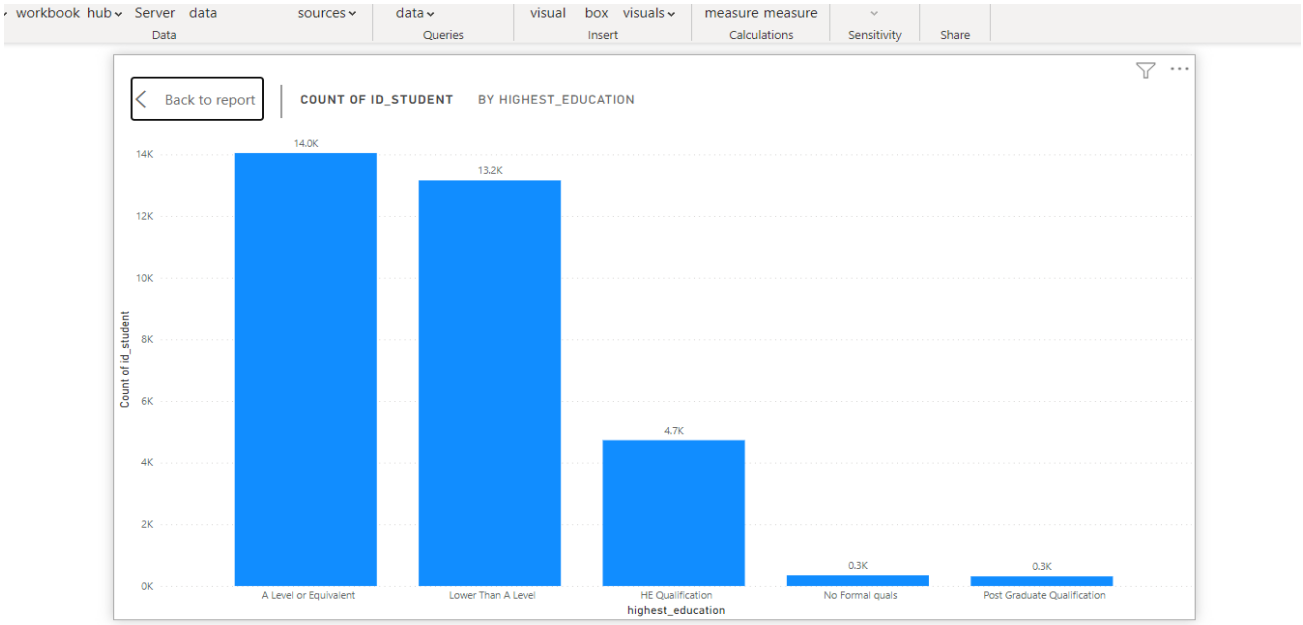
3.PREDICTIVE ANALYSIS:

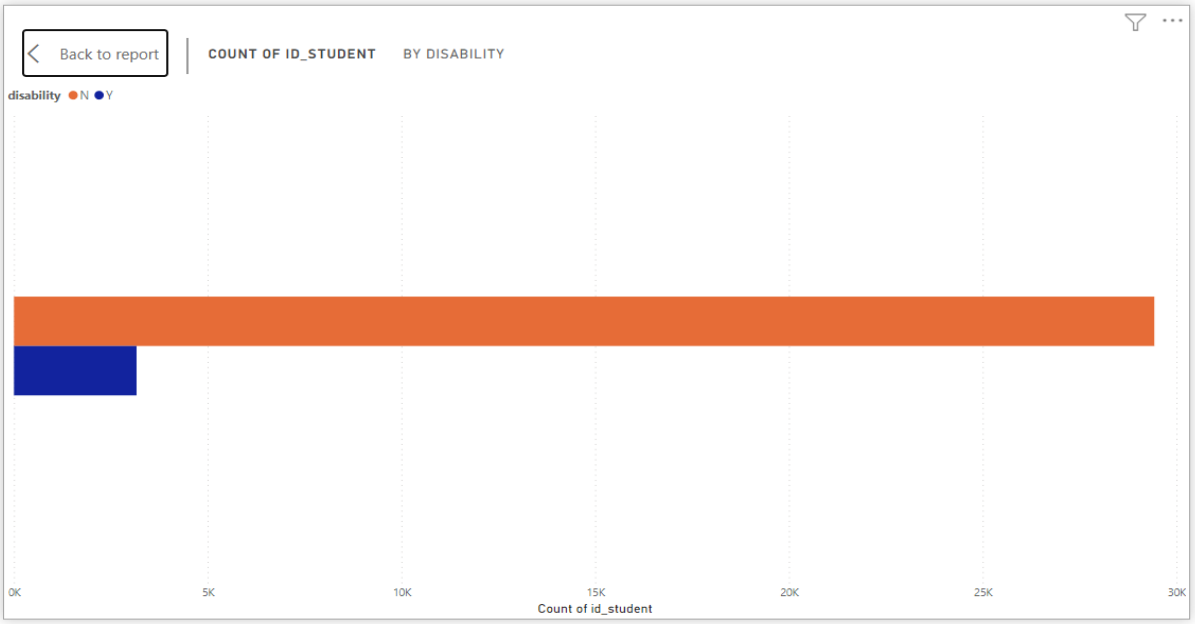
It uses statistical models and forecasts techniques to understand the future and Answers “what could happen?”

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

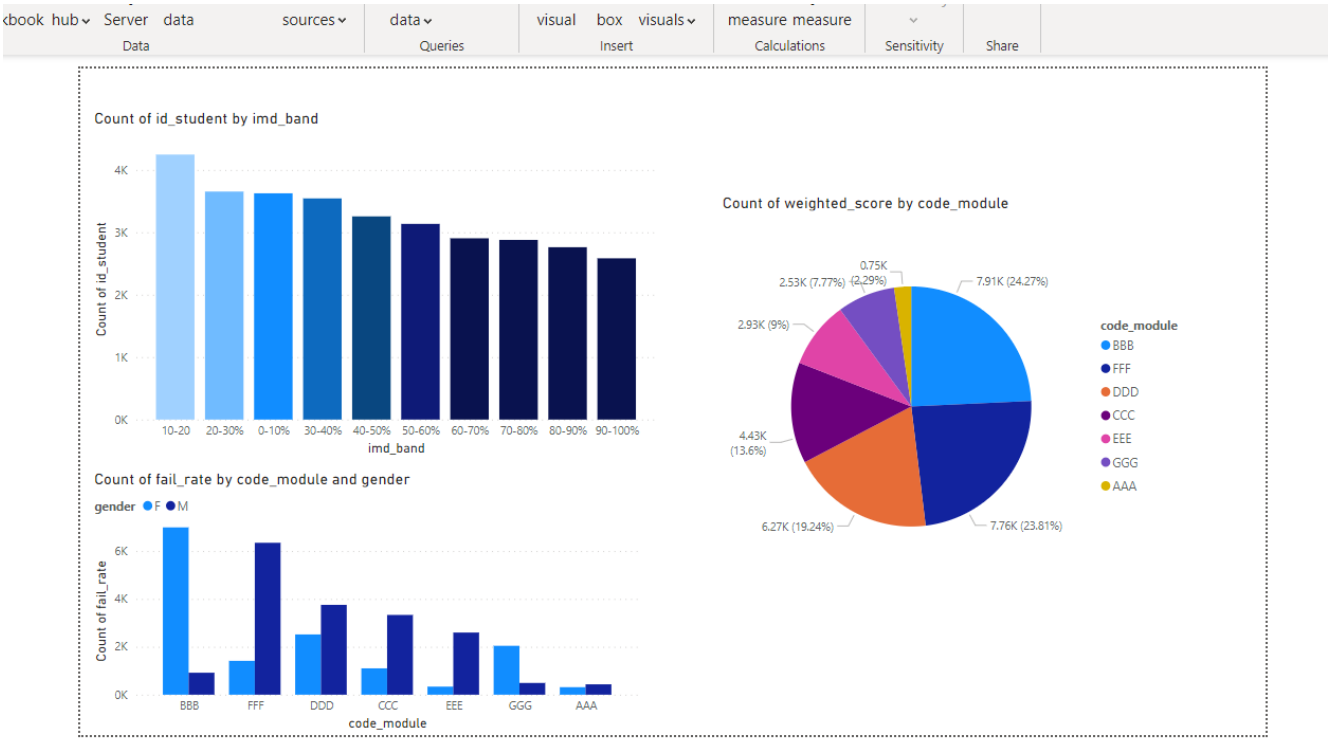








Overall conclusion so far We have seen that number there is a relationship between number of clicks and final result, and particularly that it might be possible at the lower end of engagement to predict withdrawal or failure from VLE engagement. However our independent variables by themselves don't seem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