Online Education Engagement Analysis

By: Andreah Cruz, Bindu Madhavi Edara, Harika Eadara, Hemanvitha Katakam, Venkata Siddarth Gullipalli

Dataset Context

- The Dataset we chose was generated in 2017 To analyze the U.K students' performance & engagement based on their learning behavior from different presentations (ie. sessions in American terms) & virtual learning environment (VLE) interactions
  - Dataset contains class years from 2013 (B/J) - 2014 (B/J)
    - "B" refers to the February (Spring) term.
    - "J" refers to the October (Autumn) term.

Abstract

- **Problem Definition:**
  - High dropout rates and low engagement persist in virtual learning environments (VLEs).
  - Inequities in learning outcomes remain a major issue.
  - Current technologies often use a one-size-fits-all approach, failing to meet individual learning needs.
- **Study Overview:**
  - Utilizes the Open University Learning Analytics Dataset (OULAD) with data from 2013–2014, covering U.K. student demographics, behavior, and performance in virtual learning environments.
- **Objectives:**
  - Analyze engagement patterns, predict student success, evaluate dropout factors, and optimize course design.
- **Methodology:**
  - Includes data cleaning, transformation, and preparation for visualization and analysis.
- **Impact:**
  - Findings support personalized learning strategies, improved course materials, and reduced dropout rates for a more effective online education experience.

Problem Understanding and Formulation

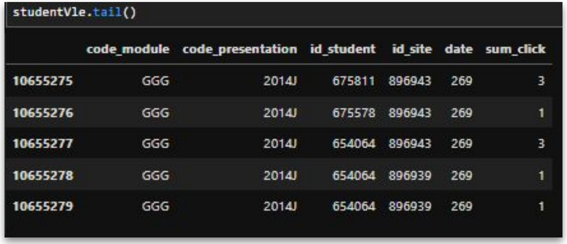
- **Problem:**
  This dataset is key to helping us understand problems that we would want to analyze more about
  - Improving Student Engagement
    - To analyze students' performance & engagement based on their learning behavior from different presentations (ie. sessions in American terms) & virtual learning environment (VLE) interactions
  - Predicting Student Success
    - Understanding and predict what students are at risk of failing or dropping out of a course
  - Understanding Demographic and Socioeconomic Factors
    - Investigating how factors such as age, gender, region, and socioeconomic status affect student performance and engagement.

Motivation For Project

- **Increasing Need:**
  - Post-COVID online education highlights disparities in student engagement and outcomes.
- **Challenges:**
  - Lack of traditional classroom immediacy affects motivation, participation, and retention.
  - Barriers are more pronounced for students from diverse demographics and socioeconomic backgrounds.
- **Data-Driven Approach:**
  - Addressing these issues requires insights from data to optimize the learning experience.
- **OULAD Dataset:**
  - Provides a framework to analyze student behavior, demographics, and academic performance.
  - Helps identify factors that drive or hinder success, fostering equitable and effective learning environments.
- **Impact:**
  - Inform interventions to improve outcomes and create inclusive education systems.

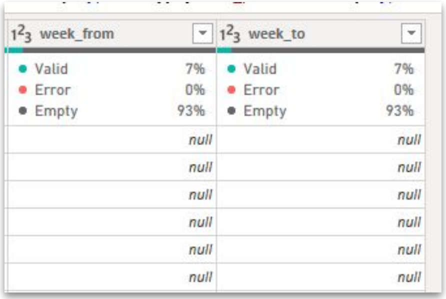
Data Understanding and Exploration

- ❖ Analyze anomalies (ie. missing values & outliers)
  - ➢ Overall most of the files contain <5% null values
  - ➢ Anything more such as 80% specifically must be dropped
- ❖ Analyze structure of data frame
  - ➢ Use .tail() to see max row as well

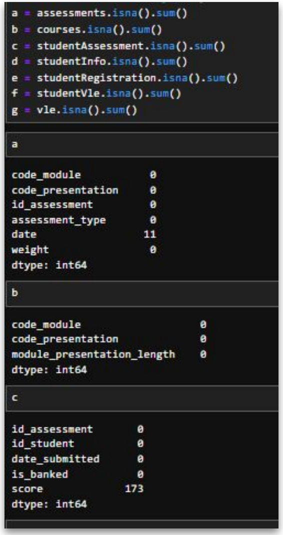


```
studentVle.tail()
```

	code_module	code_presentation	id_student	id_site	date	sum_click
10655275	GGG	2014J	675811	896943	269	3
10655276	GGG	2014J	675578	896943	269	1
10655277	GGG	2014J	654064	896943	269	3
10655278	GGG	2014J	654064	896939	269	1
10655279	GGG	2014J	654064	896939	269	1



week_from		week_to	
● Valid	7%	● Valid	7%
● Error	0%	● Error	0%
● Empty	93%	● Empty	93%
	null		*null*
	null		*null*
	null		*null*
	null		*null*
	null		*null*
	null		*null*

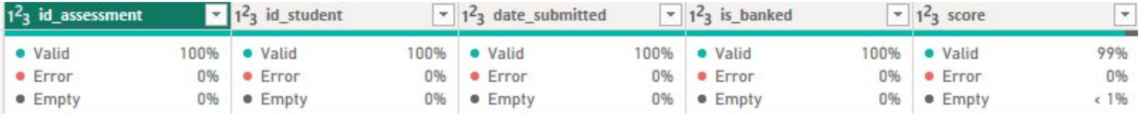


```
a = assessments.isna().sum()
b = courses.isna().sum()
c = studentAssessment.isna().sum()
d = studentInfo.isna().sum()
e = studentRegistration.isna().sum()
f = studentVle.isna().sum()
g = vle.isna().sum()
```

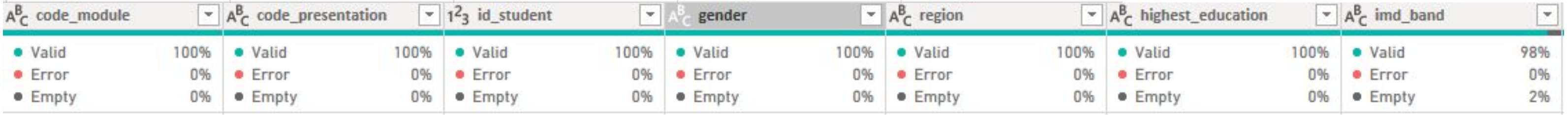
```
a

code_module          0
code_presentation    0
id_assessment        0
assessment_type      0
date                11
weight               0
dtype: int64
```

```
b

code_module                   0
code_presentation             0
module_presentation_length    0
dtype: int64
```

```
c

id_assessment       0
id_student          0
date_submitted      0
is_banked           0
score             173
dtype: int64
```



id_assessment		id_student		date_submitted		is_banked		score	
● Valid	100%	● Valid	100%	● Valid	100%	● Valid	100%	● Valid	99%
● Error	0%	● Error	0%	● Error	0%	● Error	0%	● Error	0%
● Empty	0%	● Empty	0%	● Empty	0%	● Empty	0%	● Empty	< 1%

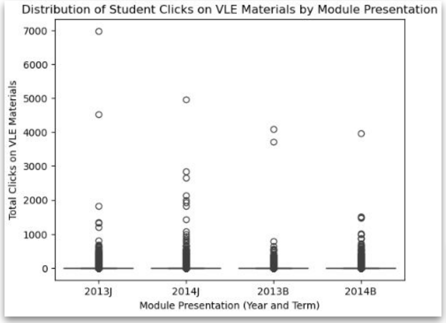


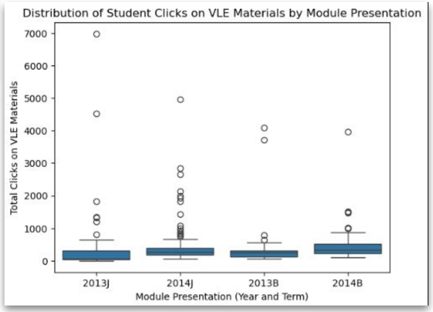
code_module		code_presentation		id_student		gender		region		highest_education		imd_band	
● Valid	100%	● Valid	100%	● Valid	100%	● Valid	100%	● Valid	100%	● Valid	100%	● Valid	98%
● Error	0%	● Error	0%	● Error	0%	● Error	0%	● Error	0%	● Error	0%	● Error	0%
● Empty	0%	● Empty	0%	● Empty	0%	● Empty	0%	● Empty	0%	● Empty	0%	● Empty	2%

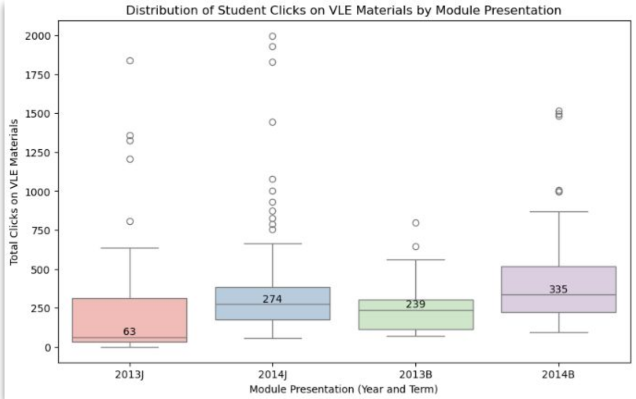
Data Preparation Steps:

Data preparation is the crucial stage that transforms raw data into a clean, structured, and usable format for analysis or modeling. It involves several steps that ensure our data is in the best possible state before feeding it into machine learning models or analytical tools.

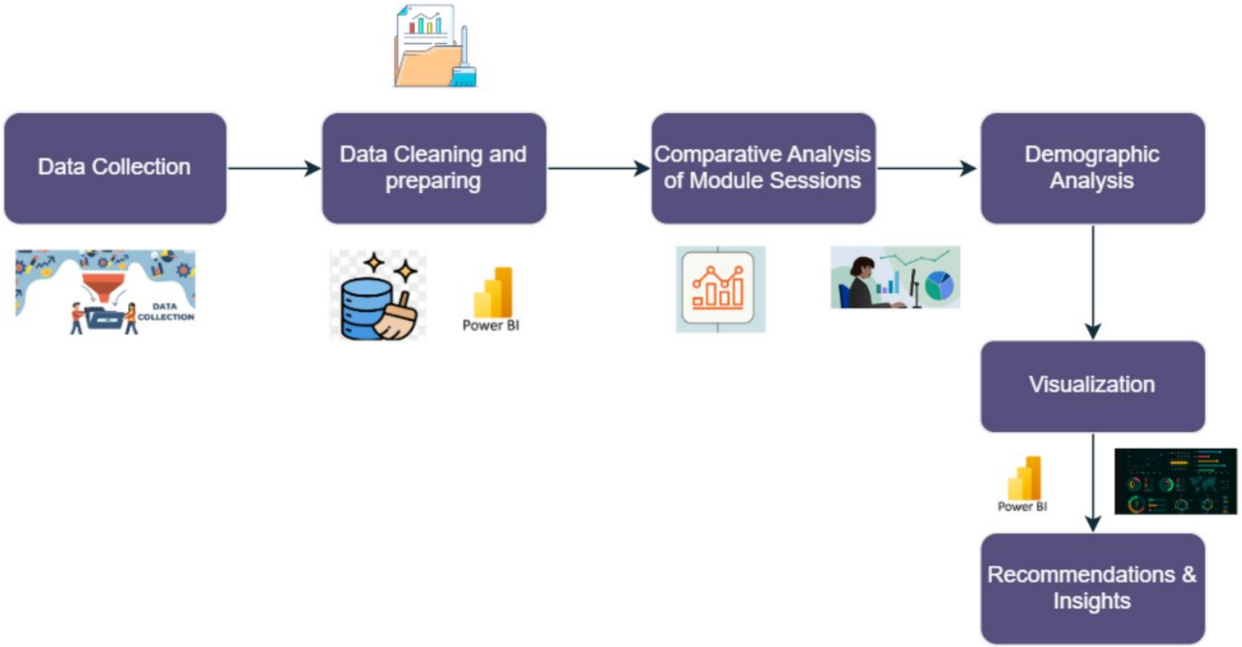
1. **Handling Missing Data**:
   - **Impute** missing values with mean, median, mode, or use more advanced techniques like multiple imputation.
   - Alternatively, **remove** rows or columns with too much missing data (typically >30%).
2. **Handling Duplicates**:
   - Remove **duplicate rows** using .drop_duplicates() to avoid bias.
   - Remove unnecessary columns
3. **Handling Outliers**:
   - **Remove** or **cap** extreme outliers based on business logic.







Data Modeling



Data Collection

Data Cleaning and preparing

Comparative Analysis of Module Sessions

Demographic Analysis

Visualization

Recommendations & Insights

Data Modeling

Purpose:

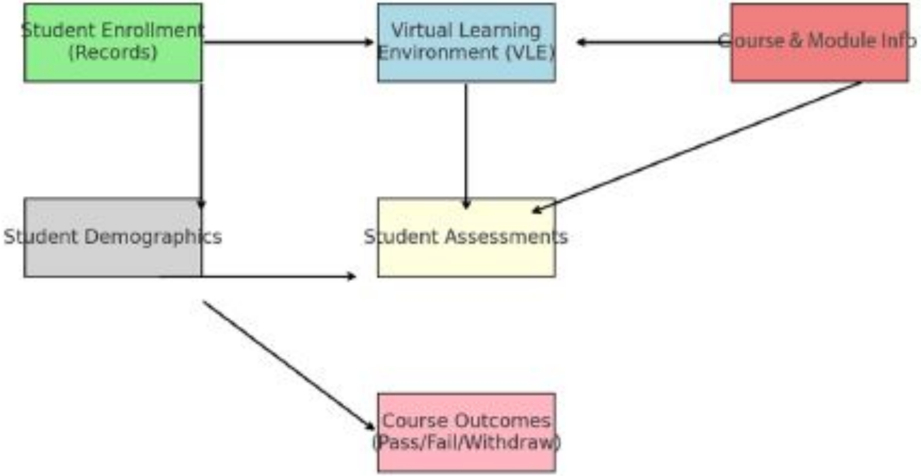
- Designed to integrate, process, and analyze student engagement data from the Virtual Learning Environment (VLE).

Key Components:

- Includes student demographics, module details, assessments, and interaction data.

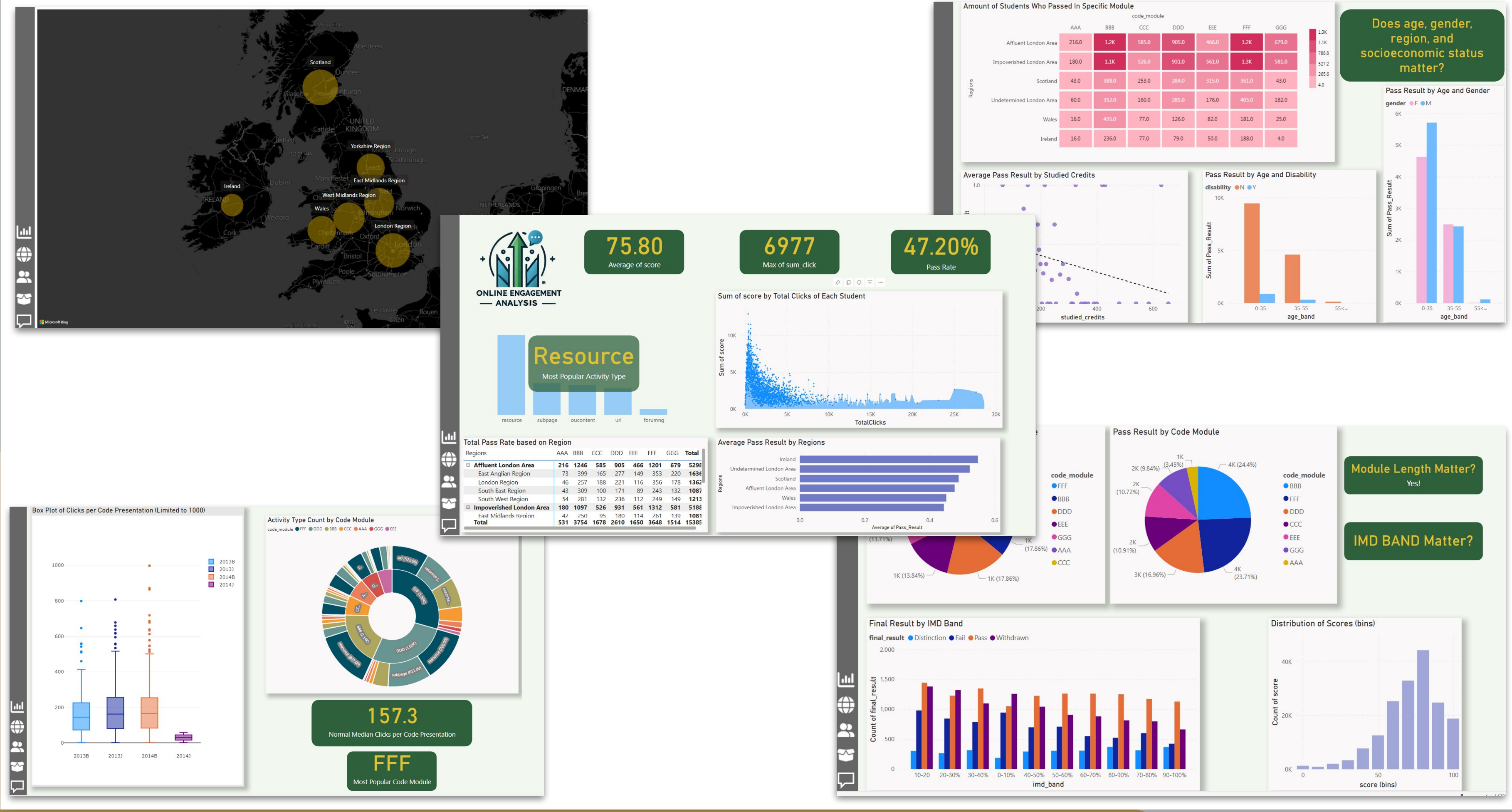
Structure:

- Organized into a relational model to support data exploration, visualization, and actionable insights.


Student Enrollment (Records)
Virtual Learning Environment (VLE)
Course & Module Info
Student Demographics
Student Assessments
Course Outcomes (Pass/Fail/Withdraw)

Data Model Overview (As reference to 2 previous slide figures)

- **Role of the Data Model:**
  - Backbone of analysis, supporting insights and informed decision-making in online education systems.
- **Core Design Components:**
  - Student Demographics
  - Student Enrollment
  - Course and Module Information
  - Assessments
  - VLE Interactions
- **Implementation and Tools:**
  - Relational schema (e.g., star schema) integrates data dimensions with fact tables for scalability and efficiency.
  - Power BI was used for data cleaning, relationship building, and pipeline creation.
- **Challenges and Scalability:**
  - Addressed inconsistent data formats, missing values, and normalization issues.
  - Scalable to handle large datasets and future modules or metrics.



ONLINE ENGAGEMENT ANALYSIS

75.80
Average of score

6977
Max of sum_click

47.20%
Pass Rate

Resource
Most Popular Activity Type

157.3
Normal Median Clicks per Code Presentation

FFF
Most Popular Code Module

Does age, gender, region, and socioeconomic status matter?

Module Length Matter?
Yes!

IMD BAND Matter?

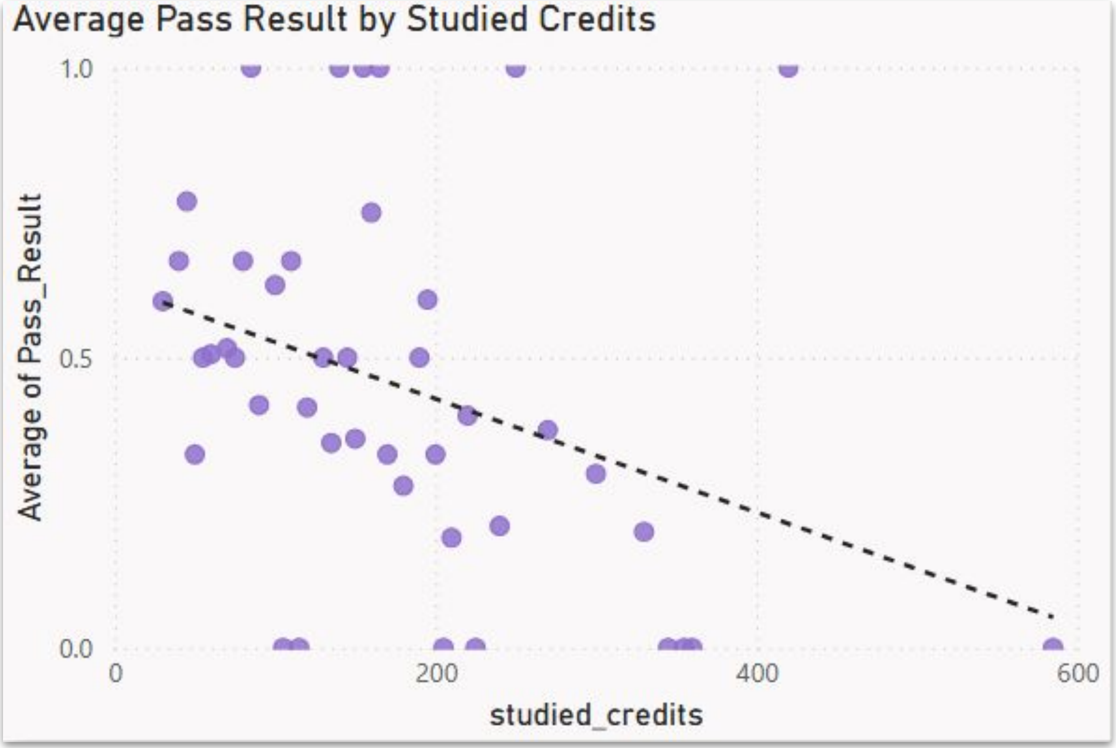
Amount of Students Who Passed In Specific Module

Regions	AAA	BBB	CCC	DDD	EEE	FFF	GGG
Affluent London Area	216.0	1.2K	585.0	905.0	466.0	1.2K	679.0
Impoverished London Area	180.0	1.1K	526.0	931.0	561.0	1.3K	581.0
Scotland	43.0	388.0	253.0	284.0	315.0	361.0	43.0
Undetermined London Area	60.0	352.0	160.0	285.0	176.0	405.0	182.0
Wales	16.0	435.0	77.0	126.0	82.0	181.0	25.0
Ireland	16.0	236.0	77.0	79.0	50.0	188.0	4.0

Total Pass Rate based on Region

Regions	AAA	BBB	CCC	DDD	EEE	FFF	GGG	Total
⊟ Affluent London Area	**216**	**1246**	**585**	**905**	**466**	**1201**	**679**	**5298**
East Anglian Region	73	399	165	277	149	353	220	**1636**
London Region	46	257	188	221	116	356	178	**1362**
South East Region	43	309	100	171	89	243	132	**1087**
South West Region	54	281	132	236	112	249	149	**1213**
⊟ Impoverished London Area	**180**	**1097**	**526**	**931**	**561**	**1312**	**581**	**5188**
East Midlands Region	42	250	95	180	114	261	139	**1081**
Total	**531**	**3754**	**1678**	**2610**	**1650**	**3648**	**1514**	**15385**

VISUALIZATIONS



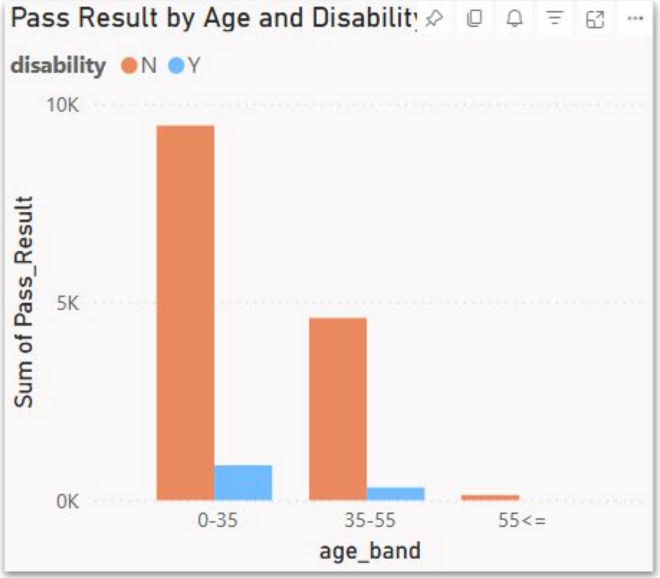
The scatter plot demonstrates a **negative correlation** between the number of studied credits and the average pass result.

Interpretation:

- As the number of studied credits increases, the likelihood of achieving a passing result (including both "Pass" and "Distinction") seems to decrease.
- Students with fewer credits tend to have higher average pass rates, while those with more credits are less likely to pass or achieve distinction.

Possible Explanations:

- **Workload Pressure**
- **Student Challenges**
- **Study Quality**

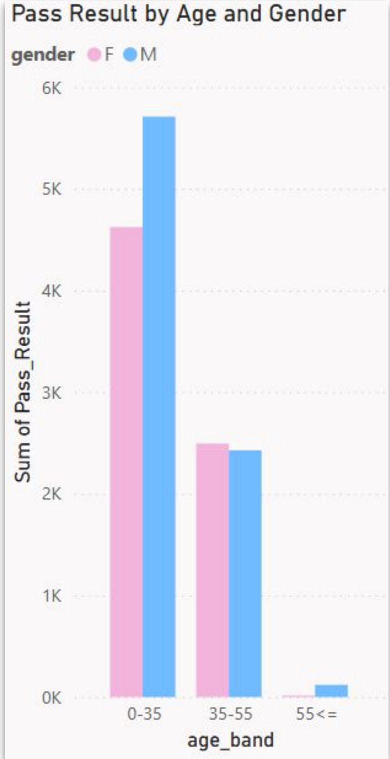


Pass Result by Age and Disability (Left):

- The bar chart shows that individuals without disabilities (orange) significantly outperform those with disabilities (blue) across all age bands.
- The 0-35 age group has the highest number of pass results for both disabled and non-disabled individuals.
- There is a sharp decline in pass results as age increases, especially for individuals with disabilities.

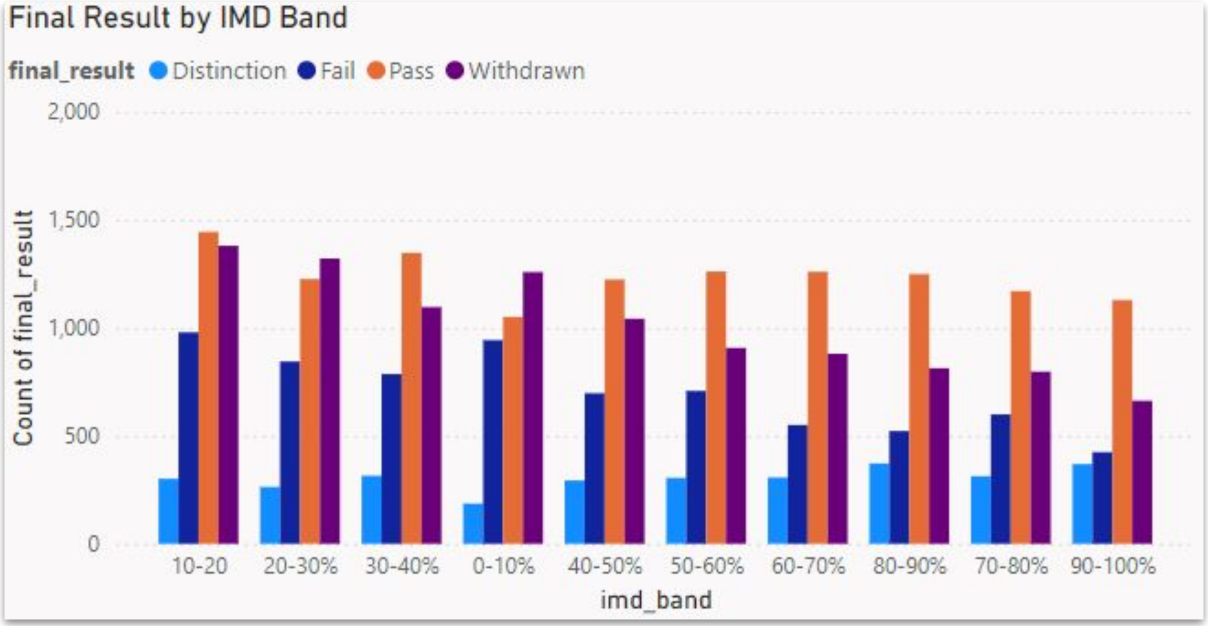
Pass Result by Age and Gender (Right):

- This chart highlights differences in pass results by gender across age bands.
- In the 0-35 age group, males (blue) outperform females (pink), with a notable gap in results.
- For the 35-55 age band, the pass results for males and females are more balanced.
- The 55+ age group has very few pass results, showing a steep decline in success rates regardless of gender.



Key Observations:

- Higher IMD (less deprivation) correlates with fewer failures (orange) and withdrawals (purple).
- Lower IMD (more deprivation) shows higher failure and withdrawal rates, highlighting challenges faced by these students.

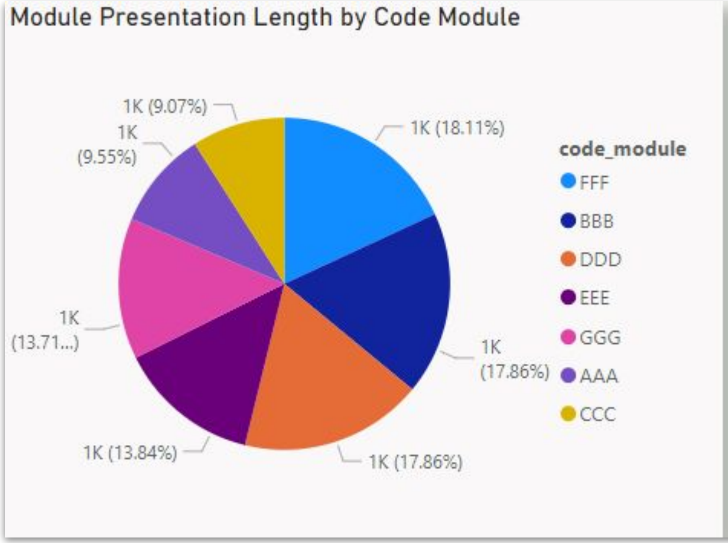


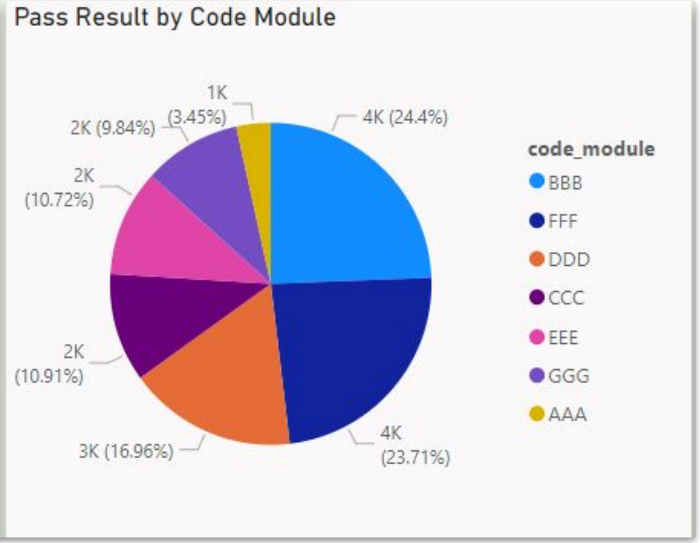
Trends:

- Students in higher IMD bands are more likely to achieve distinctions (blue) and passes (pink).
- Socioeconomic factors significantly impact academic outcomes.

Takeaway:

- Reducing deprivation-related barriers could improve success rates for students in lower IMD bands.





1. **Key Observations:**
   - **Module Presentation Length (Left):**
     - Longer presentation lengths (e.g., FFF, BBB) are more common for modules like FFF (18.11%) and BBB (17.86%).
     - Modules with shorter lengths have smaller proportions in the chart.
   - **Pass Results (Right):**
     - Modules with longer presentation lengths (e.g., BBB, FFF) correspond to higher pass rates, as shown in their larger share of the pass result distribution (24.4% for BBB and 23.71% for FFF).
2. **Insight:**
   - Longer module presentations may contribute to higher student success, as modules with extended presentation durations show a strong correlation with higher pass rates.
3. **Takeaway:**
   - Investing in longer, well-structured module presentations could improve student outcomes and pass rates.

EVALUATION

- Focused on the effectiveness of the Power BI dashboard in communicating insights and addressing key questions in online education.
- Assessed each visualization contribution to the project objectives and its alignment with the overall narrative.
- Demonstrated the dashboard’s role in improving student outcomes by identifying at-risk students through engagement and performance metrics.
- Highlighted actionable insights, such as targeting modules with low pass rates to guide interventions and resource allocation.

Evaluation of Performance & Impact of Visuals

Interactivity Improvements:

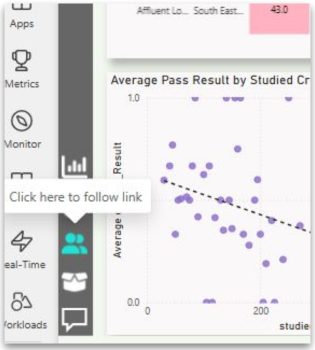
- Enhanced visuals with interactivity, allowing users to drill down or up for detailed exploration of data .
- Created hierarchies within the regions column to categorize data and focus on specific regions.

User Navigation:

- Added navigation buttons on the bottom left for easy maneuvering between boards.

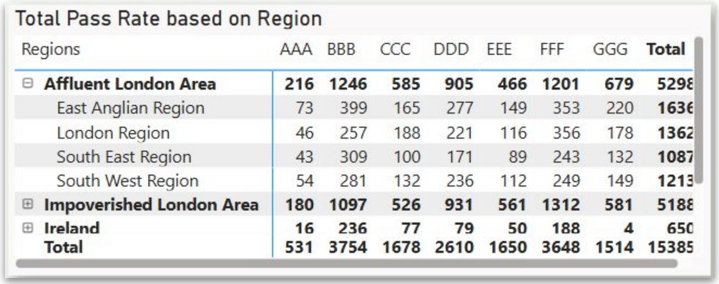
Highlighting Key Points:

- Included bold-text shapes to emphasize critical details and direct user attention to important insights.


Apps
Metrics
Monitor
eal-Time
Vorkloads
Click here to follow link
Affluent Lo... South East...
43.0
Average Pass Result by Studied Cr
1.0
0.0
Average _Result
0
200
studie

Benefits of Enhancements:

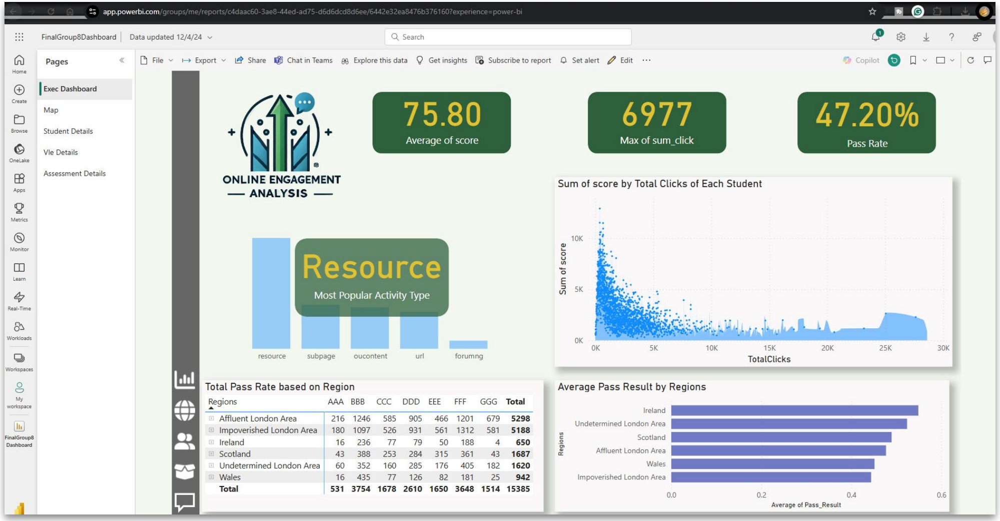
- Improved accessibility and customization of the dashboard.
- Enabled users to gain deeper insights and make more informed decisions effectively.



Total Pass Rate based on Region

Regions	AAA	BBB	CCC	DDD	EEE	FFF	GGG	Total
⊟ **Affluent London Area**	**216**	**1246**	**585**	**905**	**466**	**1201**	**679**	**5298**
East Anglian Region	73	399	165	277	149	353	220	**1636**
London Region	46	257	188	221	116	356	178	**1362**
South East Region	43	309	100	171	89	243	132	**1087**
South West Region	54	281	132	236	112	249	149	**1213**
⊞ **Impoverished London Area**	**180**	**1097**	**526**	**931**	**561**	**1312**	**581**	**5188**
⊞ **Ireland**	**16**	**236**	**77**	**79**	**50**	**188**	**4**	**650**
Total	**531**	**3754**	**1678**	**2610**	**1650**	**3648**	**1514**	**15385**

DEPLOYMENT



Total Pass Rate based on Region

Regions	AAA	BBB	CCC	DDD	EEE	FFF	GGG	Total
Affluent London Area	216	1246	585	905	466	1201	679	**5298**
Impoverished London Area	180	1097	526	931	561	1312	581	**5188**
Ireland	16	236	77	79	50	188	4	**650**
Scotland	43	388	253	284	315	361	43	**1687**
Undetermined London Area	60	352	160	285	176	405	182	**1620**
Wales	16	435	77	126	82	181	25	**942**
Total	**531**	**3754**	**1678**	**2610**	**1650**	**3648**	**1514**	**15385**

Accessibility and Scalability:

- Hosted on Power BI and integrates seamlessly with institutional systems, such as university LMS, for timely insights.

Real-Time Updates:

- Dynamically incorporates new data to ensure relevance and accuracy.

High Performance:

- Handles large datasets (e.g., 10-million-row studentVLE) while maintaining optimal performance in multi-user scenarios.

Impact:

- Identifies at-risk students early and highlights demographic trends to inform targeted interventions and policy decisions.

CONCLUSION AND FUTURE WORK

- Higher interaction levels with VLE materials correlate with improved academic performance.
- Quality of engagement is equally important as quantity, as outliers reveal variability.
- Disparities in pass rates are observed across regions, age groups, and disability statuses, emphasizing the need for targeted support.
- An interactive Power BI dashboard enables educators to monitor engagement and identify at-risk students.
- The dashboard facilitates timely interventions and implementation of personalized learning strategies.
- The research highlights the effectiveness of learning analytics in improving online education.
- Future work includes real-time analytics, predictive modeling, and incorporating variables like student feedback.

THANK YOU