FACULTY OF INFORMATION TECHNOLOGY

CITY UNIVERSITY MALAYSIA

CYBERJAYA CAMPUS

Final Project

Project Report: Student Performance Analysis at City University Malaysia

Subject name and ID: BIT2053 Fundamentals of Modern Data.

Lecturer name: Nazmirul Izzad Bin Nasir.

Name	Student ID
Aiman Gamal	202205020021
Chiyangwa Ephraim Panashe	202401010422
Amr Alawadhi	202209010022
Abdelaziz Bashir Hussien	202405010379
Syed Emeirul Habib Bin Syed Ameen	202209020135

Acknowledgment

We, Chiyangwa Ephraim Panashe, Amr Alawadhi, Abdelaziz Bashir, Aiman Gamal, and Syed Emeirul Habib Bin Syed Ameen, would like to thank our lecturer, Nazmirul Izzad Bin Nasir, for his guidance and feedback during this project. We also acknowledge City University Malaysia for providing some resources that supported our work. Finally, we appreciate the collaboration and contributions of all group members in completing this report.

Contents

Title: Enhancing Student Performance and Resource Management at City University Malaysia.

GitHub Repository Link: https://github.com/Emio0000/student-performance-analysis-for-modern-data.git

First: Executive Summary:

City University Malaysia is facing difficulty in efficiently tracking student performance, course performance, and also with allocating resources. These issues affect student achievement and curriculum quality. This project suggests that a Business Intelligence based solution to address the problem will combine the principles of computational thinking with data analytics and visualization software. The solution will be implemented using decomposition, abstraction, and algorithmic thinking to improve performance tracking, course evaluation, and resource optimization. Data preprocessing and analysis will be done with Python; interactive dashboards will be offered on BI platforms like Power BI. The given approach is viable, scalable, and consistent with the institutional objectives, and eventually results in a positive learning outcome and enhanced strategic decision-making.

Second: Problem Statement:

Although student data is available in large amounts, City University Malaysia does not have a well-integrated, data-driven system to track academic performance, to assess course effectiveness, and to optimize resource allocation. Existing systems hamper the ability to restrict struggling students at an early stage, evaluate courses with high rates of failure, and allocate faculty and facilities with high levels of efficiency. The impact of such a gap is the lack of chances to intervene in time, to improve the curriculum, and to develop the institution.

Third : Methodology:

1- Dataset source and data collection:

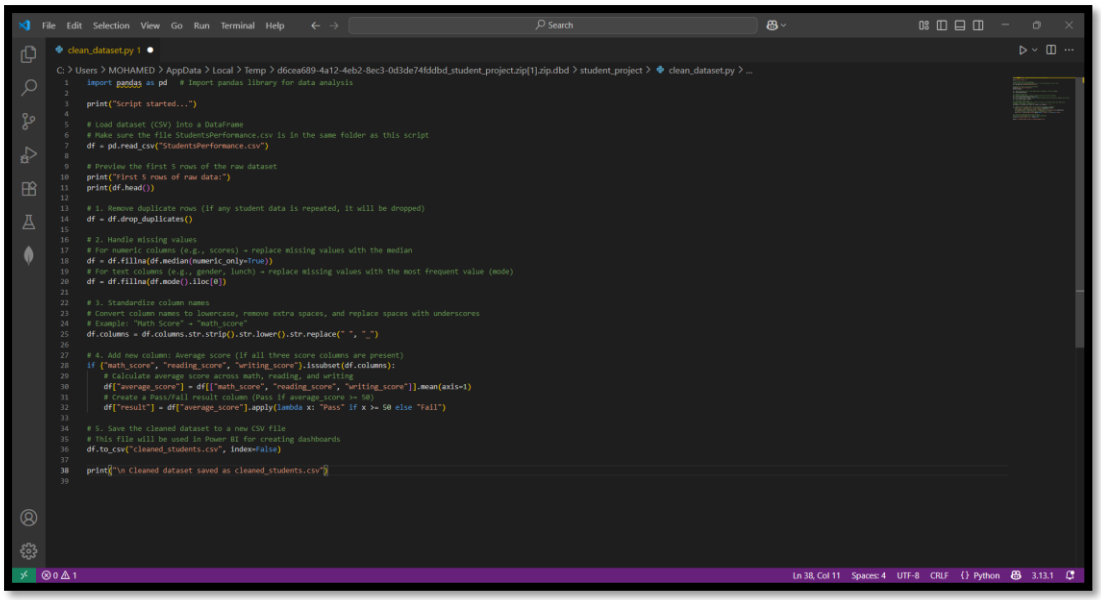
(Students' Performance in Exams Dataset).

contains information on the marks gained by students in multiple subjects. It provides insights into factors that may influence academic achievement in the future, such as parental background and tests preparations. The dataset is particularly useful for analysing patterns in student performance and identifying the correlations between demographic factors and exam outcomes. By using this dataset, the university can explore key drivers of academic success and design targeted interventions to support students productivity.(The Dataset link: Students Performance in Exams).

2- Data Preprocessing and Cleaning

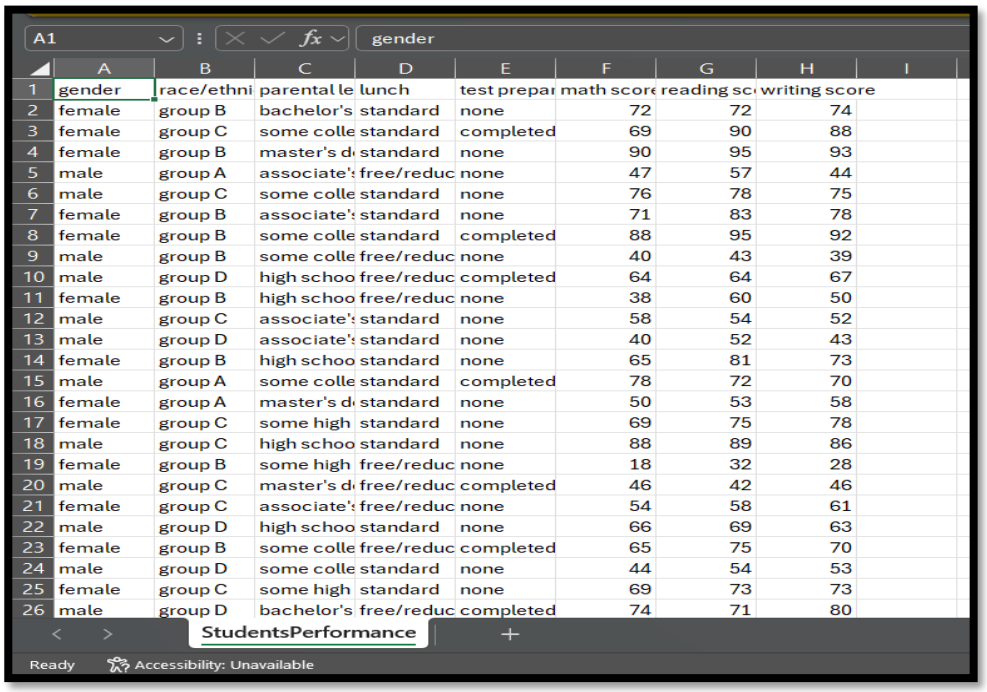
A systematic approach was used to process the data in Python and Pandas library to prepare the dataset to conduct the analysis. These steps are as follows:

1) **Data Importation**: The raw information (StudentsPerformance.csv) was loaded into Pandas Data Frame, which would be investigated and examined.
2) **Duplicate Removiing**: To eradicate errors in the data and biases in further analysis, duplicate Records were removed.
3) **Handling Missing Values**: In numeric attributes (e.g., test scores), the value has been replaced with the median to maintain the central tendency, and in categorical attributes (gender, lunch type), it has been replaced with the mode to ensure logical consistency.
4) **Naming of Columns**: The column names were standardized by changing them to lower case, removing excess spaces, and substituting them with underscores. This improved the usability of the data provided and the ability to analyse it.
5) **Feature Engineering**: The Average scores variable was changed by dividing math points, reading points, and writing points by 3, a categorical outcome variable was computed, and students were graded Pass or Fail by checking whether their average score was 50.
6) **Export Data**: The clean and rich data were exported to cleanedstudents.csv. It is based on this file that additional analysis and dashboard development is performed Power BI for visualization.



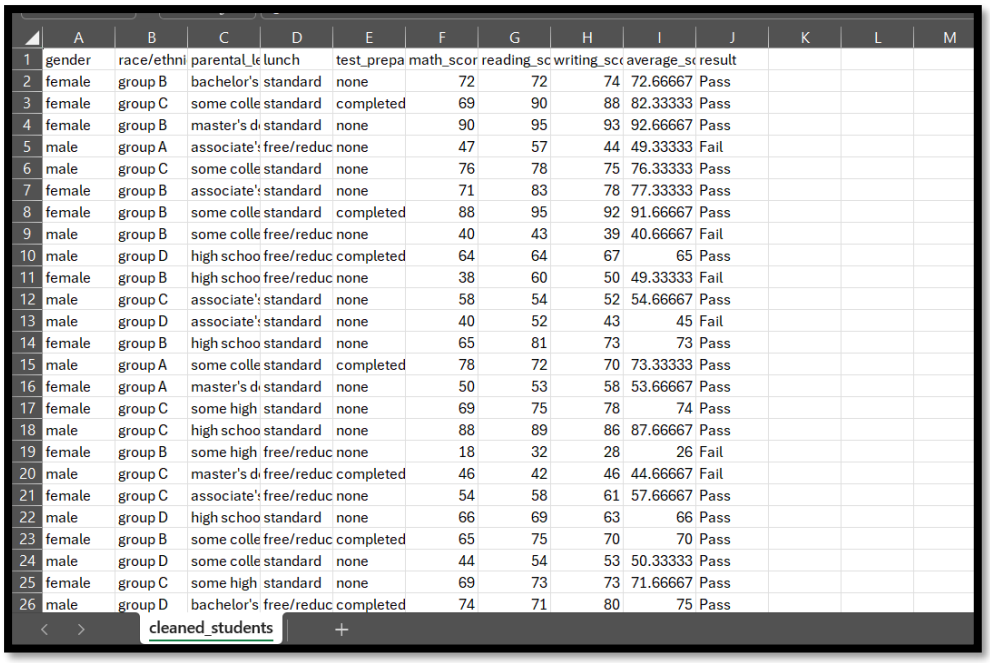
```python
import pandas as pd  # Import pandas library for data analysis

print("Script started...")

# Load dataset (CSV) into a DataFrame
# Make sure the file StudentsPerformance.csv is in the same folder as this script
df = pd.read_csv("StudentsPerformance.csv")

# Preview the first 5 rows of the raw dataset
print("First 5 rows of raw data:")
print(df.head())

# 1. Remove duplicate rows (if any student data is repeated, it will be dropped)
df = df.drop_duplicates()

# 2. Handle missing values
# For numeric columns (e.g., scores) → replace missing values with the median
df = df.fillna(df.median(numeric_only=True))
# For text columns (e.g., gender, lunch) → replace missing values with the most frequent value (mode)
df = df.fillna(df.mode().iloc[0])

# 3. Standardize column names
# Convert column names to lowercase, remove extra spaces, and replace spaces with underscores
# Example: "Math Score" → "math_score"
df.columns = df.columns.str.strip().str.lower().str.replace(" ", "_")

# 4. Add new column: Average score (if all three score columns are present)
if {"math_score", "reading_score", "writing_score"}.issubset(df.columns):
    # Calculate average score across math, reading, and writing
    df["average_score"] = df[["math_score", "reading_score", "writing_score"]].mean(axis=1)
    # Create a Pass/Fail result column (Pass if average_score >= 50)
    df["result"] = df["average_score"].apply(lambda x: "Pass" if x >= 50 else "Fail")

# 5. Save the cleaned dataset to a new CSV file
# This file will be used in Power BI for creating dashboards
df.to_csv("cleaned_students.csv", index=False)

print("\n Cleaned dataset saved as cleaned_students.csv")
```

(Figure-1, that shows the cleaning and preprocessing of the dataset)



	A	B	C	D	E	F	G	H	I
1	gender	race/ethni	parental le	lunch	test prepar	math score	reading sc	writing score	
2	female	group B	bachelor's	standard	none	72	72	74	
3	female	group C	some colle	standard	completed	69	90	88	
4	female	group B	master's d	standard	none	90	95	93	
5	male	group A	associate's	free/reduc	none	47	57	44	
6	male	group C	some colle	standard	none	76	78	75	
7	female	group B	associate's	standard	none	71	83	78	
8	female	group B	some colle	standard	completed	88	95	92	
9	male	group B	some colle	free/reduc	none	40	43	39	
10	male	group D	high schoo	free/reduc	completed	64	64	67	
11	female	group B	high schoo	free/reduc	none	38	60	50	
12	male	group C	associate's	standard	none	58	54	52	
13	male	group D	associate's	standard	none	40	52	43	
14	female	group B	high schoo	standard	none	65	81	73	
15	male	group A	some colle	standard	completed	78	72	70	
16	female	group A	master's d	standard	none	50	53	58	
17	female	group C	some high	standard	none	69	75	78	
18	male	group C	high schoo	standard	none	88	89	86	
19	female	group B	some high	free/reduc	none	18	32	28	
20	male	group C	master's d	free/reduc	completed	46	42	46	
21	female	group C	associate's	free/reduc	none	54	58	61	
22	male	group D	high schoo	standard	none	66	69	63	
23	female	group B	some colle	free/reduc	completed	65	75	70	
24	male	group D	some colle	standard	none	44	54	53	
25	female	group C	some high	standard	none	69	73	73	
26	male	group D	bachelor's	free/reduc	completed	74	71	80	

(Fig-2 the dataset after cleaning & processing)



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	gender	race/ethni	parental_le	lunch	test_prepa	math_scor	reading_sc	writing_sc	average_s	result			
2	female	group B	bachelor's	standard	none	72	72	74	72.66667	Pass			
3	female	group C	some colle	standard	completed	69	90	88	82.33333	Pass			
4	female	group B	master's d	standard	none	90	95	93	92.66667	Pass			
5	male	group A	associate'	free/reduc	none	47	57	44	49.33333	Fail			
6	male	group C	some colle	standard	none	76	78	75	76.33333	Pass			
7	female	group B	associate'	standard	none	71	83	78	77.33333	Pass			
8	female	group B	some colle	standard	completed	88	95	92	91.66667	Pass			
9	male	group B	some colle	free/reduc	none	40	43	39	40.66667	Fail			
10	male	group D	high schoo	free/reduc	completed	64	64	67	65	Pass			
11	female	group B	high schoo	free/reduc	none	38	60	50	49.33333	Fail			
12	male	group C	associate'	standard	none	58	54	52	54.66667	Pass			
13	male	group D	associate'	standard	none	40	52	43	45	Fail			
14	female	group B	high schoo	standard	none	65	81	73	73	Pass			
15	male	group A	some colle	standard	completed	78	72	70	73.33333	Pass			
16	female	group A	master's d	standard	none	50	53	58	53.66667	Pass			
17	female	group C	some high	standard	none	69	75	78	74	Pass			
18	male	group C	high schoo	standard	none	88	89	86	87.66667	Pass			
19	female	group B	some high	free/reduc	none	18	32	28	26	Fail			
20	male	group C	master's d	free/reduc	completed	46	42	46	44.66667	Fail			
21	female	group C	associate'	free/reduc	none	54	58	61	57.66667	Pass			
22	male	group D	high schoo	standard	none	66	69	63	66	Pass			
23	female	group B	some colle	free/reduc	completed	65	75	70	70	Pass			
24	male	group D	some colle	standard	none	44	54	53	50.33333	Pass			
25	female	group C	some high	standard	none	69	73	73	71.66667	Pass			
26	male	group D	bachelor's	free/reduc	completed	74	71	80	75	Pass			

cleaned_students

(Fig-3 the dataset after performing feature engineering, and adding the Average score of students, and their grade(pass/fail)).

3- Data Analysis

1) Descriptive Analysis (What is happening now?)

Resource Utilization: Track how many students access cloud resources like online libraries, learning platforms, and when their usage peaks.

Access Patterns: Analyse which learning resources are most frequently accessed and which are underutilized.

Performance Correlation: Compare students' cloud usage time with their average academic performance to see if higher usage aligns with better outcomes.

Example case: On average, students who spend more than 3 hours weekly on cloud-based test prep materials score 15% higher in exams (that is an observation/descriptive analysis).

2) Predictive Analysis (What could happen in the future?)

Student Performance Prediction in the future: Use machine learning models to predict which students risk failing based on their cloud usage behaviour like log-in frequency, and time spent.

Personalized Recommendations: Build recommendation systems that suggest additional resources to students for example; extra practice in weak subjects, based on their usage history and performance trends.

Scalability Forecasting: Predict future demand for cloud storage and processing power as enrolment grows, ensuring the university can scale resources efficiently.

4- Visualization And Power BI Dashboards

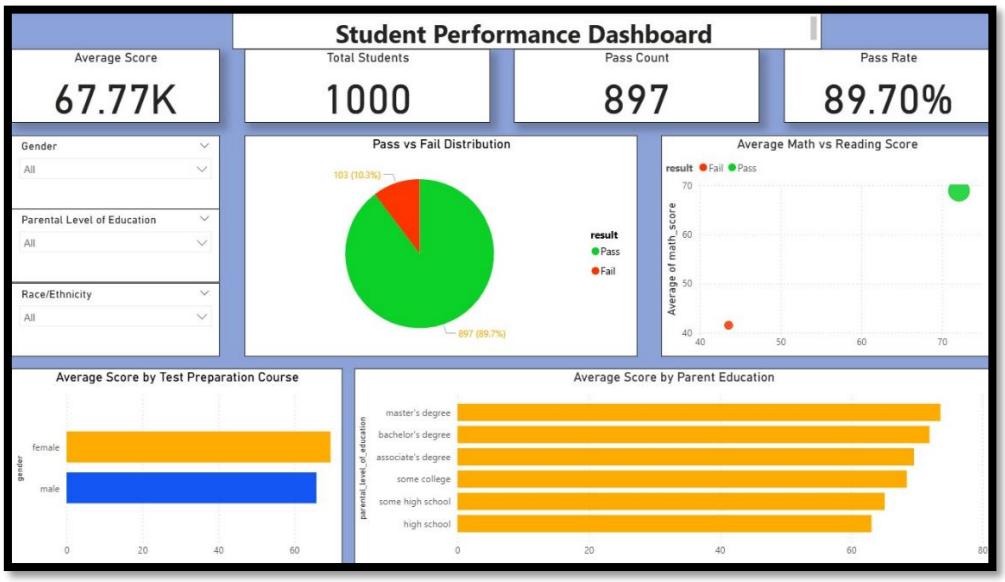
The Power BI dashboard reveals important findings regarding student performance as the following: **Overall Performance**: The average student score is approximately 67.77, with a pass rate of 89.7%. This indicates that most students perform successfully, though about 10% remain at risk of failing.

Subject-Level Findings: Weakness in mathematics is the primary reason for poor student performance. The low-performing mathematics students are among the failing students.

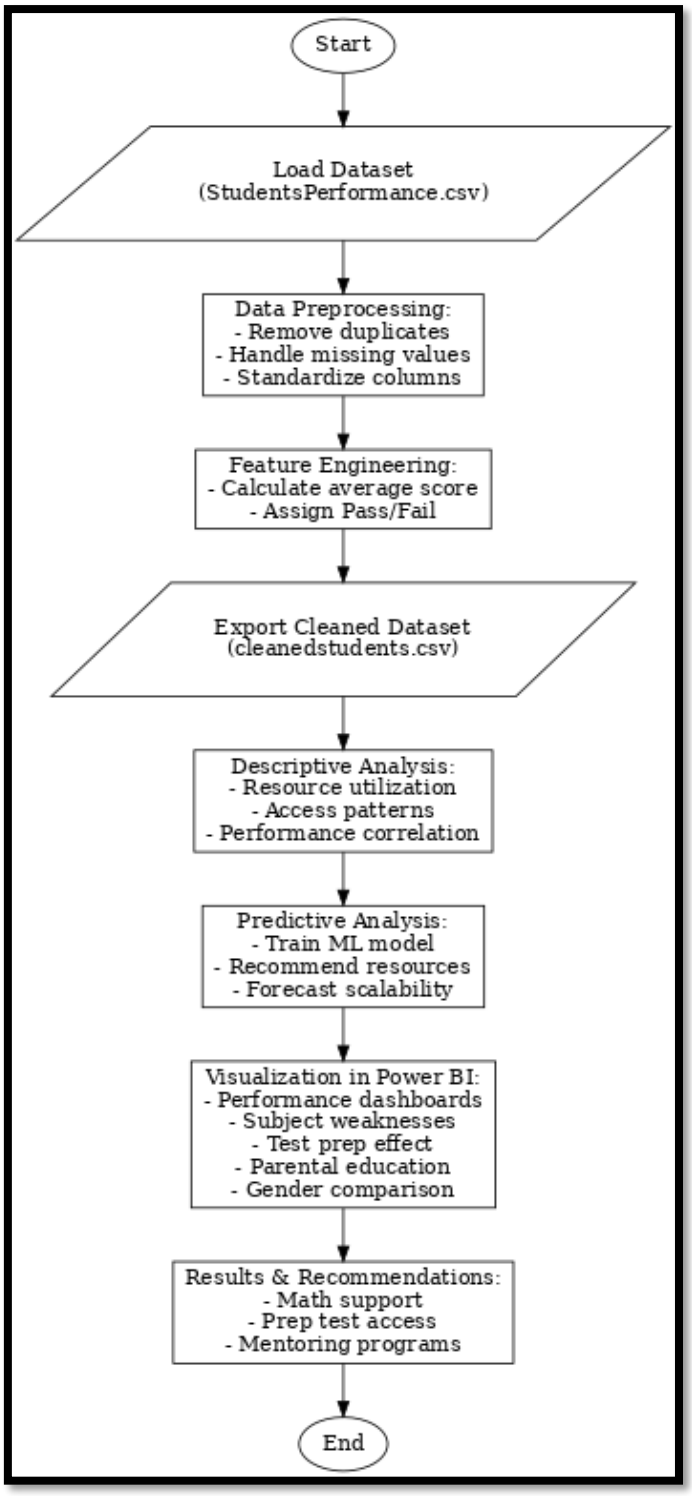
Test Preparation Influence: Those students who received a test preparation course had significantly higher scores than students who received no test preparation. This demonstrates the effectiveness of structured preparation in boosting performance.

Parental Education Influence: Student outcomes are positively correlated with parental education level. Learners whose parents hold higher qualifications, such as bachelor's or master's degrees, consistently record stronger results than those whose parents completed only high school.

Gender Differences: Both male and female students perform at similar levels, with no substantial gender-based performance gap observed in the dataset.



(Fig-4 that shows the power BI visualization)



(Fig-5 shows the flowchart visualization)

The flowchart above shows the whole process that is done in our group in order to produce a complete visualization dashboard in Power BI. These includes from finding and loading data which is the input until the visualization steps in Power BI which is the output.

Fourth: Result And Recommendations:

The results show that the majority of students are performing satisfactorily but that there is scope for improvement through tackling some specific critical risk areas. The interventions would have to address mathematics support, the extension of test preparation programs, and targeted mentoring for those from less wealthy family backgrounds. These would help in lowering failure levels, increasing equity, and student attainment.

Recommendations:

Based on students' performance data analysis, several recommendations for improving learning achievements at City University Malaysia can be presented:

First, mathematics support must be increased since lack of performance in mathematics has a direct link with failure rates. Targeted interventions such as remedial courses, tutoring, or internet-based learning aids will bridge this gap.

Second, the university needs to provide a means of expanding access to prep tests. The data indicate that students who took these courses actually scored higher on average, which indicates the value of good academic preparation for better success rates**.**

References

1. *Data Cleaning with Pandas*. (n.d.). KDnuggets. https://www.kdnuggets.com/data-cleaning-with-pandas

2. Samuel, O. (2024, January 30). *How to Use Pandas for Data Cleaning and Preprocessing*. FreeCodeCamp.org. https://www.freecodecamp.org/news/data-cleaning-and-preprocessing-with-pandasbdvhj/

3. Anello, E. (2023, August 3). *7 Steps to Mastering Data Cleaning and Preprocessing Techniques*. KDnuggets. https://www.kdnuggets.com/2023/08/7-steps-mastering-data-cleaning-preprocessing-techniques.html

4. Maher, M. A. (2024, December 31). *1.Introduction Universities often find themselves accumulating data without taking effective action, which can lead to passivity.* Linkedin.com. https://www.linkedin.com/pulse/challenges-opportunities-implementing-data-management-mohamed-a--vlm5e

Enhancing Student Performance and Resource Management at City University Malaysia.

First: Problem Identification

Student Performance Tracking

Trends across semesters are not properly monitored.
Early identification and intervention for struggling students is limited.

Course Effectiveness Assessment

No systematic evaluation of high-failure courses.
Missed opportunities to improve course design and learning outcomes.

Resource Allocation

Faculty, facilities, and learning resources often misaligned with demand.
Leads to inefficiency and impacts overall institutional performance.

Second: Computational Thinking Concepts and Solution Plan

We used the three concepts of computational thinking to break the problem down and find proper solutions

- Decomposition of the problems
- Abstraction
- Algorithmic Thinking Preparation and analysis of data

Third: Methosdology

1. Dataset source and data collection:

Dataset: Students' Performance in Exams.

- Contains student marks across multiple subjects
- Includes factors influencing achievement:
  - Parental background
  - Test preparation
- Useful for analyzing
  - Patterns in performance
  - Correlations between demographics & outcomes
- Enables the university to:
  - Identify key drivers of academic success
  - Design targeted student interventions

2-Data Preprocessing and Cleaning

Data Preprocessing Steps

- Duplicate Removal
  - Eliminated repeated records to reduce errors and biases.
- Handling Missing Values
  - Numeric attributes: replaced with median.
  - Categorical attributes: replaced with mode.
- Column Naming
  - Standardized by using lowercase, removing spaces, and using underscores.
  - Improved usability and consistency in analysis.
- Feature Engineering
  - Created average score
  - Generated categorical outcome

3- Data Analysis

1. Descriptive Analysis (What is happening now?)
   - Resource Utilization: Track student usage of online libraries & platforms, including peak times.
   - Access Patterns: Identify most/least used learning resources.
   - Performance Correlation: Compare cloud usage with academic results.

2. Predictive Analysis (What could happen in the future?)
   - Performance Prediction: Machine learning to flag students at risk of failing (e.g., low log-ins, short usage times).
   - Personalized Recommendations: Suggest tailored resources (e.g., extra practice in weak subjects)
   - Scalability Forecasting: Predict future demand for cloud storage & processing as enrollment rises.

4- Visualization And Power BI Dashboards

Key Findings and Observations

- Overall Performance
  - Average student score: 67.77
  - Pass rate: 89.7% → ~10% at risk of failing
- Subject-Level Insights
  - Mathematics is the weakest subject
  - Low math scores are strongly linked to student failure
- Test Preparation
  - Students with a test prep course scored significantly higher
  - Shows effectiveness of structured preparation
- Parental Education
  - Higher parental qualifications = stronger student outcomes
  - Clear positive correlation
- Gender Differences
  - Male & female students perform at similar levels

5- Result And Recommendations

- The majority of students perform satisfactorily.
- Critical risk areas remain:
  - Weakness in mathematics.
  - Need for wider test preparation access.
  - Students from less wealthy backgrounds require additional mentoring.
- Addressing these areas will reduce failure rates, improve equity, and raise student attainment.

Recommendations:

- Strengthen Mathematics Support
  - Offer remedial classes, tutoring, and online learning aids.
  - Target students consistently underperforming in mathematics.
- Expand Test Preparation Programs
  - Increase availability of structured prep courses.
  - Ensure broader access for all students, not just a few.
- Targeted Mentoring & Equity Support
  - Provide mentoring and academic guidance to students from less advantaged backgrounds.
  - Encourage peer learning and faculty-student engagement.

THANK YOU VERY MUCH!