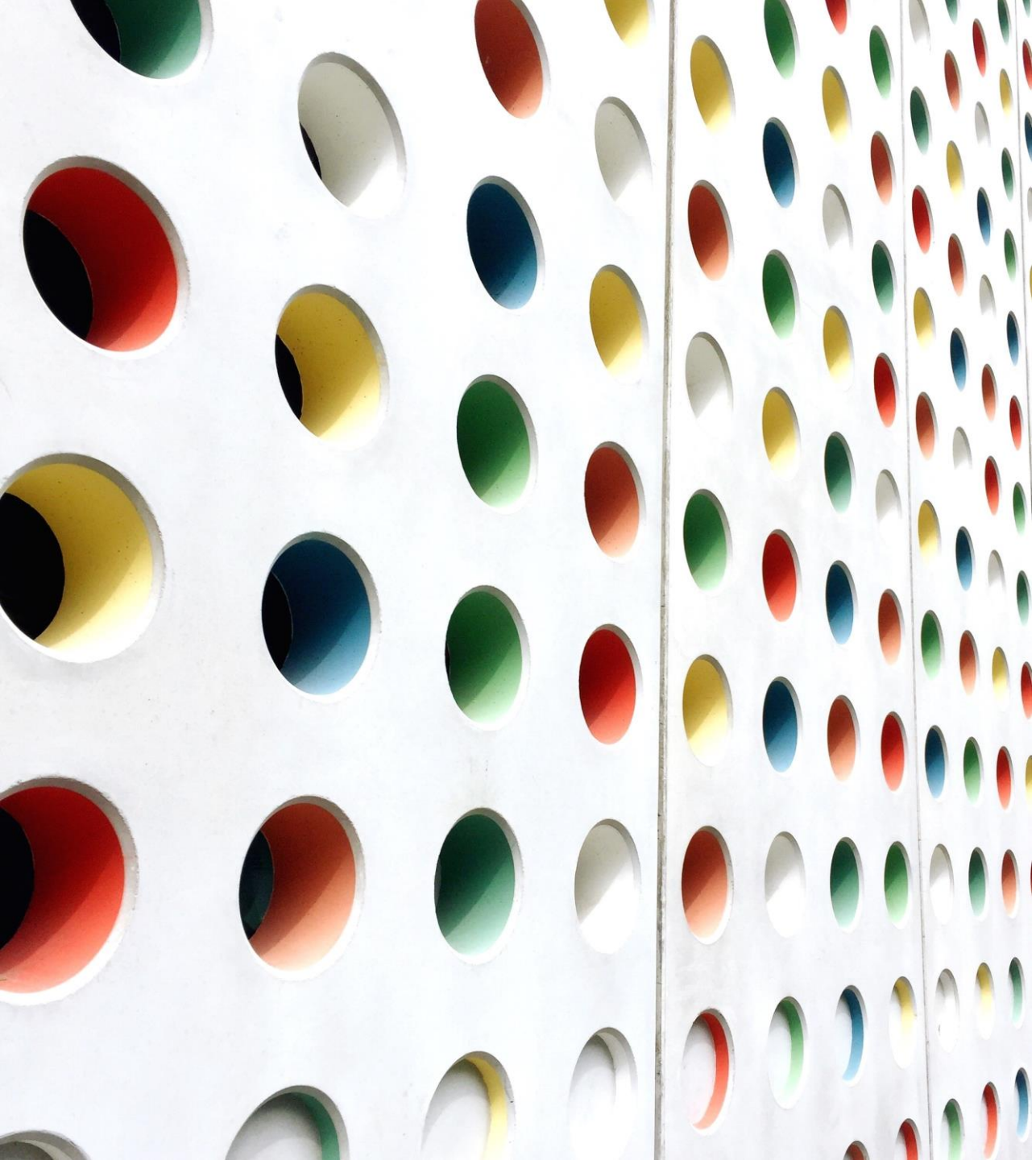


Title:
Cloud-Based Analytics for Student Mental Health Data

Subtitle:
A Microsoft Azure Implementation Project

INSTITUTION:
YORK ST JOHN UNIVERSITY
(DATA SCIENCE DEPT.)

DATE:
15TH APRIL 2025

BY: SAMUEL OYEWUNMI

INTRODUCTION

BRIEF OVERVIEW OF STUDENT MENTAL HEALTH CHALLENGES

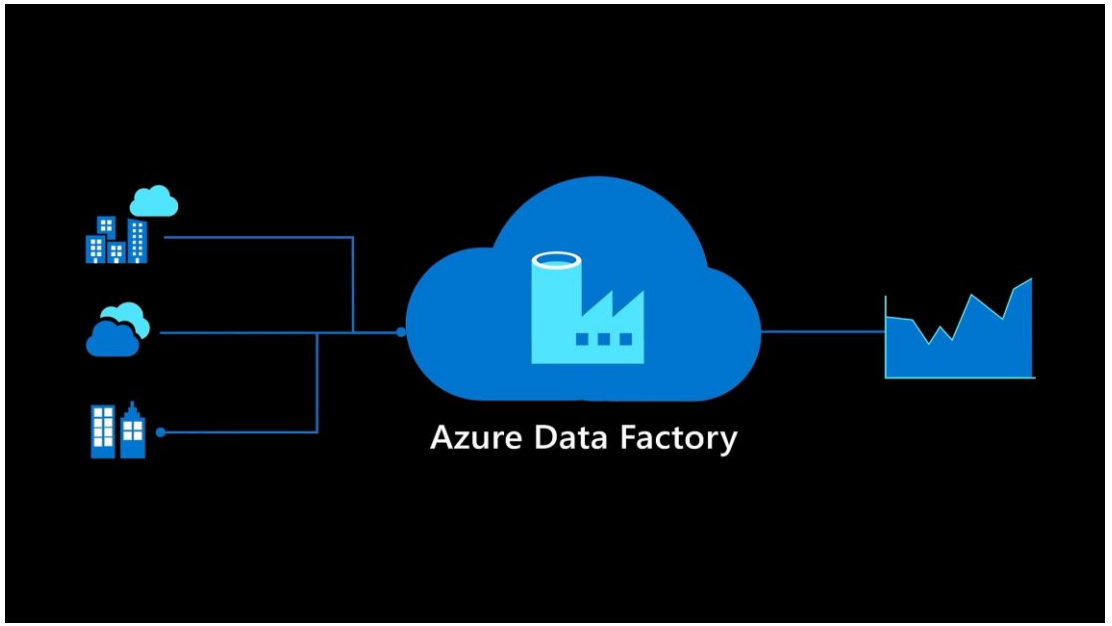
- Students today face increasing academic pressure, financial stress, social anxiety, and isolation.
- Mental health issues such as depression, anxiety, and burnout are becoming more common among university students.
- Early identification and intervention are critical to support student wellbeing and academic success.

IMPORTANCE OF DATA-DRIVEN APPROACHES

Data-driven strategies enable institutions to move beyond assumptions and anecdotal evidence. They help identify hidden patterns, predict mental health risks, and allocate resources efficiently. With real-time analytics, stakeholders can make informed decisions, tailor support services, and measure the effectiveness of interventions — ultimately improving student outcomes and wellbeing.

Project Objective

- To provide a concise introduction to the concept of Bigdata and its diverse applications across Cloud Computing.
- Leverage Microsoft Azure for data ingestion, analysis, and actionable insights.



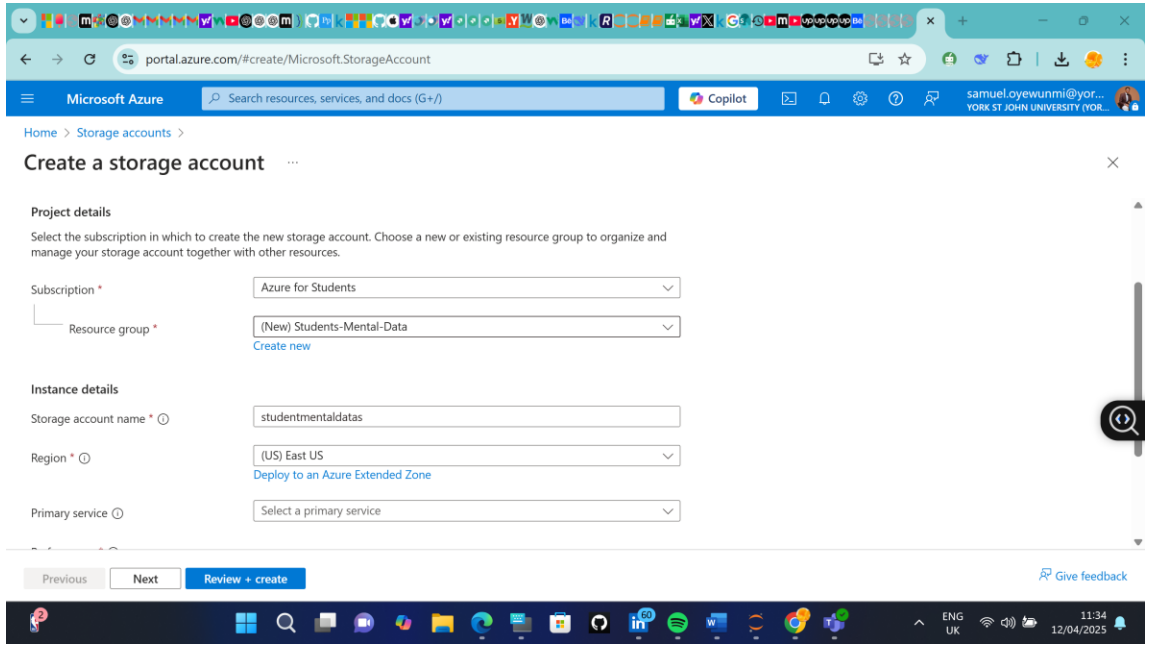
DATA INGESTION & STORAGE

- **Data Sources:** https://www.kaggle.com/datasets/shariful07/student-mental-health/data
- **Storage Solution:** Azure Blob Storage
- **Why Blob Storage:**
  - Cost-effective (Hot/Cool tiers) but for this project Hot tier is used.
  - Redundant (Geo-redundant storage)
  - Scalable for large datasets

Ingestion Process:

- Azure Data Factory pipelines
- Scheduled ingestion from secure endpoints

The image below shows the storage account page



SCALABLE PROCESSING ARCHITECTURE

Azure Services Used: Azure Data Factory (ADF), also Python & Azure SDK (Developer-Friendly) is used first because it allows automation and CI/CD pipelines integration. Its more like python-based ETL (Extract, Transform, Load) approach.

Install Azure SDK in Your Python Environment

```
[40]: pip install azure-storage-blob
```

```
Requirement already satisfied: azure-storage-blob in c:\users\homepc\anaconda3\lib\site-packages (12.25.1)
Requirement already satisfied: azure-core>=1.30.0 in c:\users\homepc\anaconda3\lib\site-packages (from azure-storage-blob) (1.33.0)
Requirement already satisfied: cryptography>=2.1.4 in c:\users\homepc\anaconda3\lib\site-packages (from azure-storage-blob) (43.0.0)
Requirement already satisfied: typing-extensions>=4.6.0 in c:\users\homepc\anaconda3\lib\site-packages (from azure-storage-blob) (4.11.0)
Requirement already satisfied: isodate>=0.6.1 in c:\users\homepc\anaconda3\lib\site-packages (from azure-storage-blob) (0.7.2)
Requirement already satisfied: requests>=2.21.0 in c:\users\homepc\anaconda3\lib\site-packages (from azure-core>=1.30.0->azure-storage-blob) (2.32.3)
Requirement already satisfied: six>=1.11.0 in c:\users\homepc\anaconda3\lib\site-packages (from azure-core>=1.30.0->azure-storage-blob) (1.16.0)
Requirement already satisfied: cffi>=1.12 in c:\users\homepc\anaconda3\lib\site-packages (from cryptography>=2.1.4->azure-storage-blob) (1.17.1)
Requirement already satisfied: pycparser in c:\users\homepc\anaconda3\lib\site-packages (from cffi>=1.12->cryptography>=2.1.4->azure-storage-blob) (2.21)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\homepc\anaconda3\lib\site-packages (from requests>=2.21.0->azure-core>=1.30.0->azure-storage-blob) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in c:\users\homepc\anaconda3\lib\site-packages (from requests>=2.21.0->azure-core>=1.30.0->azure-storage-blob) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\homepc\anaconda3\lib\site-packages (from requests>=2.21.0->azure-core>=1.30.0->azure-storage-blob) (2.2.3)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\homepc\anaconda3\lib\site-packages (from requests>=2.21.0->azure-core>=1.30.0->azure-storage-blob) (2025.1.31)
Note: you may need to restart the kernel to use updated packages.
```

Here, is the code used from ingest data to the pipeline;

Ingest and Read the Data Later

```
[44]: # Create blob client
blob_client = BlobClient.from_connection_string(connect_str, container_name, blob_name)

# Download and read CSV
stream = BytesIO()
blob_client.download_blob().readinto(stream)
df = pd.read_csv(BytesIO(stream.getvalue()))
df.head()
```

[44]:

	Timestamp	Choose your gender	Age	What is your course?	Your current year of Study	What is your CGPA?	Marital status	Do you have Depression?	Do you have Anxiety?	Do you have Panic attack?	Did you seek any specialist for a treatment?
0	08/07/2020 12:02	Female	18.0	Engineering	year 1	3.00 - 3.49	No	Yes	No	Yes	No
1	08/07/2020	Male	21.0	Islamic	year 2	3.00 - 3.49	No	No	Yes	No	No

ARCHITECTURE DIAGRAM (flow from ingestion to analysis)

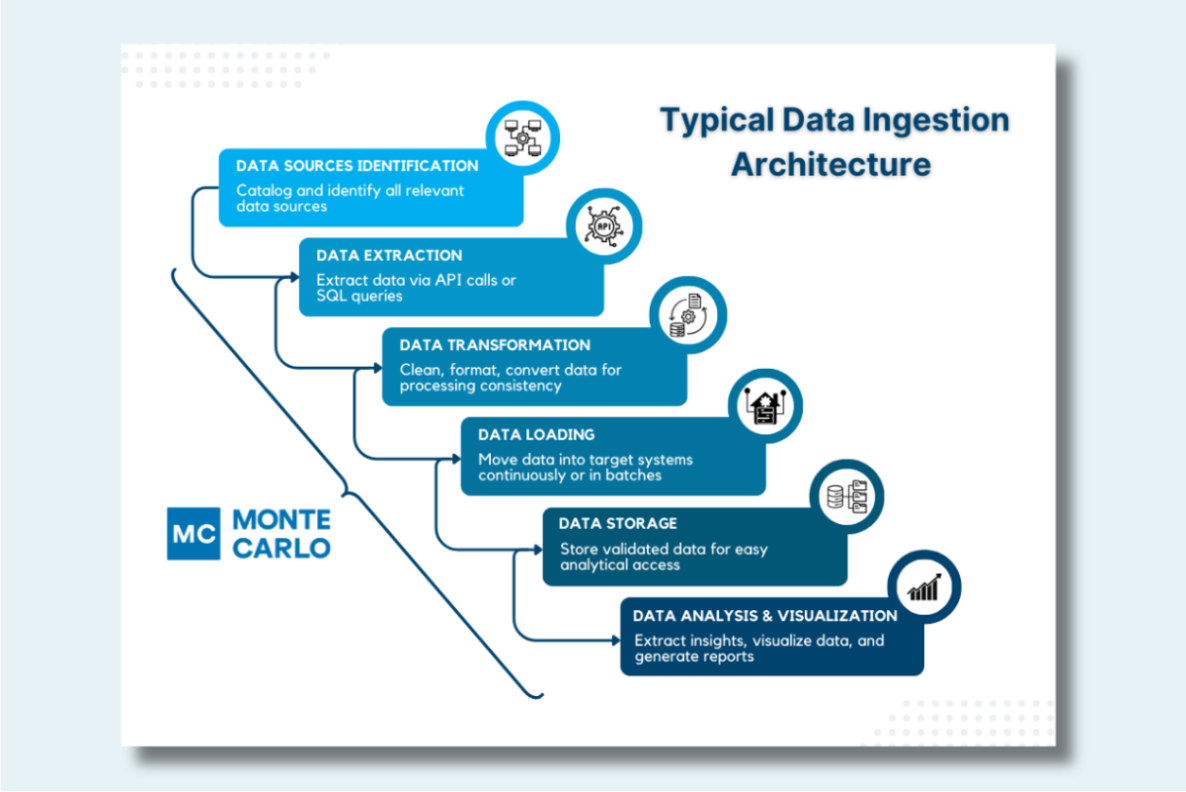


Image from Monte Carlo

STAGES TAKEN FOR THIS PROJECT

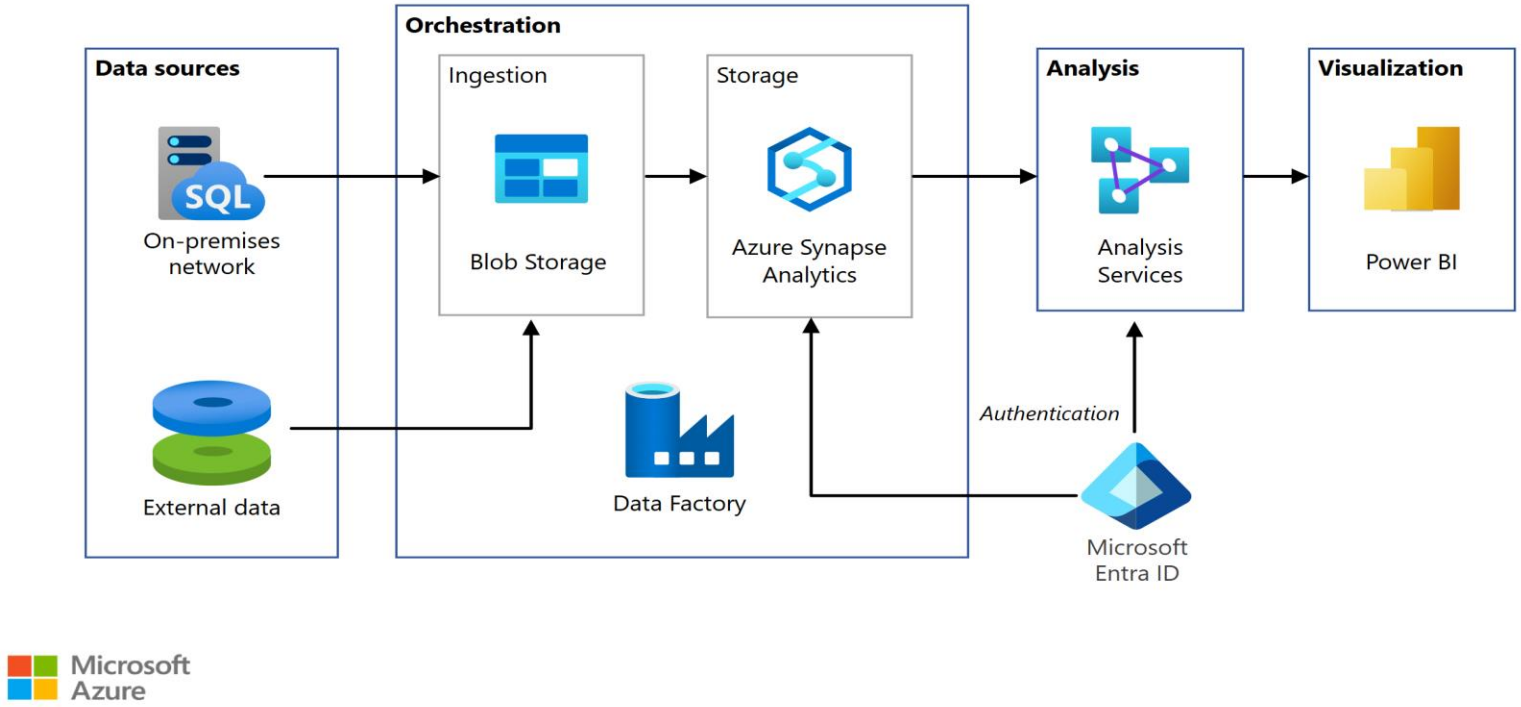
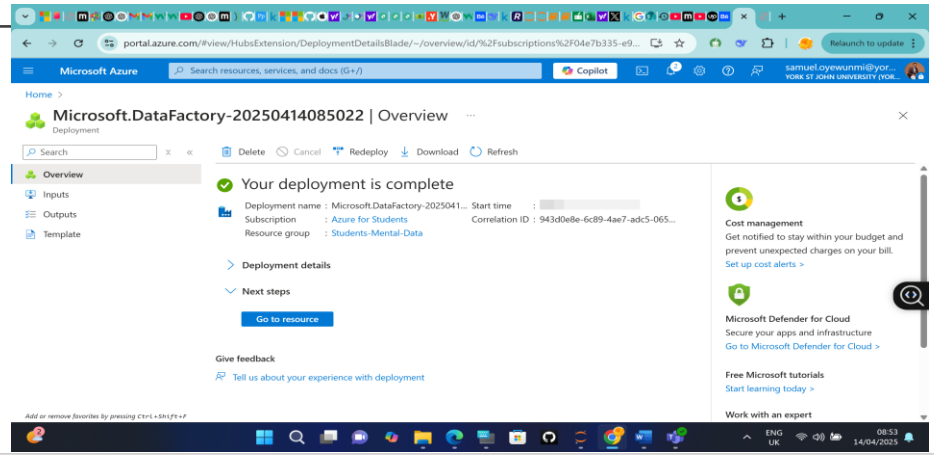


IMAGE FROM MICROSOFT AZURE

DATA EXTRACTION AND PREPROCESSING

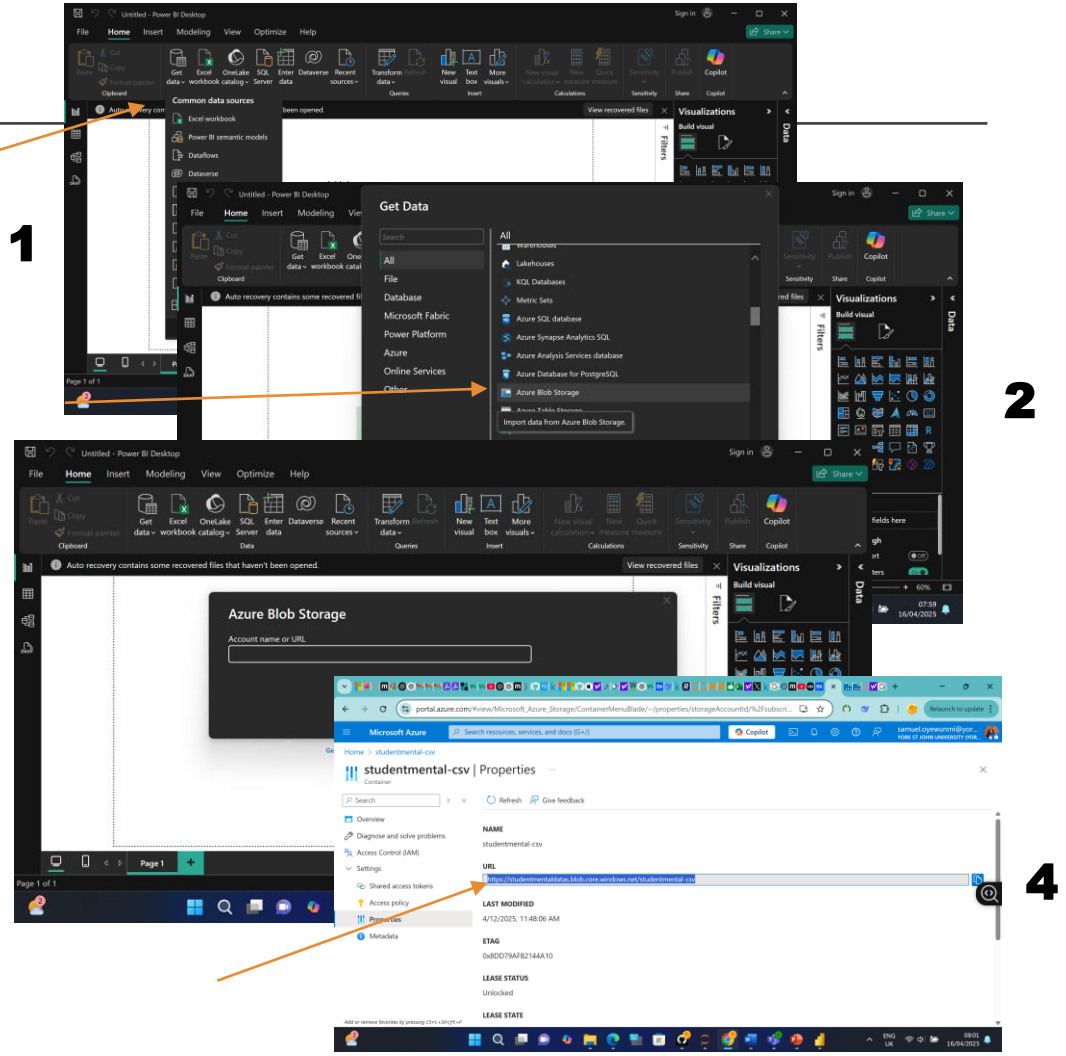
- **Extraction Tools:** Azure Data Factory, Azure Storage Explorer
- **Pre-processing Techniques:**
  - Handling missing values (nulls)
  - Standardizing field names and date format
  - Removing duplicates
- **Output:** Cleaned dataset saved in Azure Blob Storage as cleaned_student_mental_health.csv



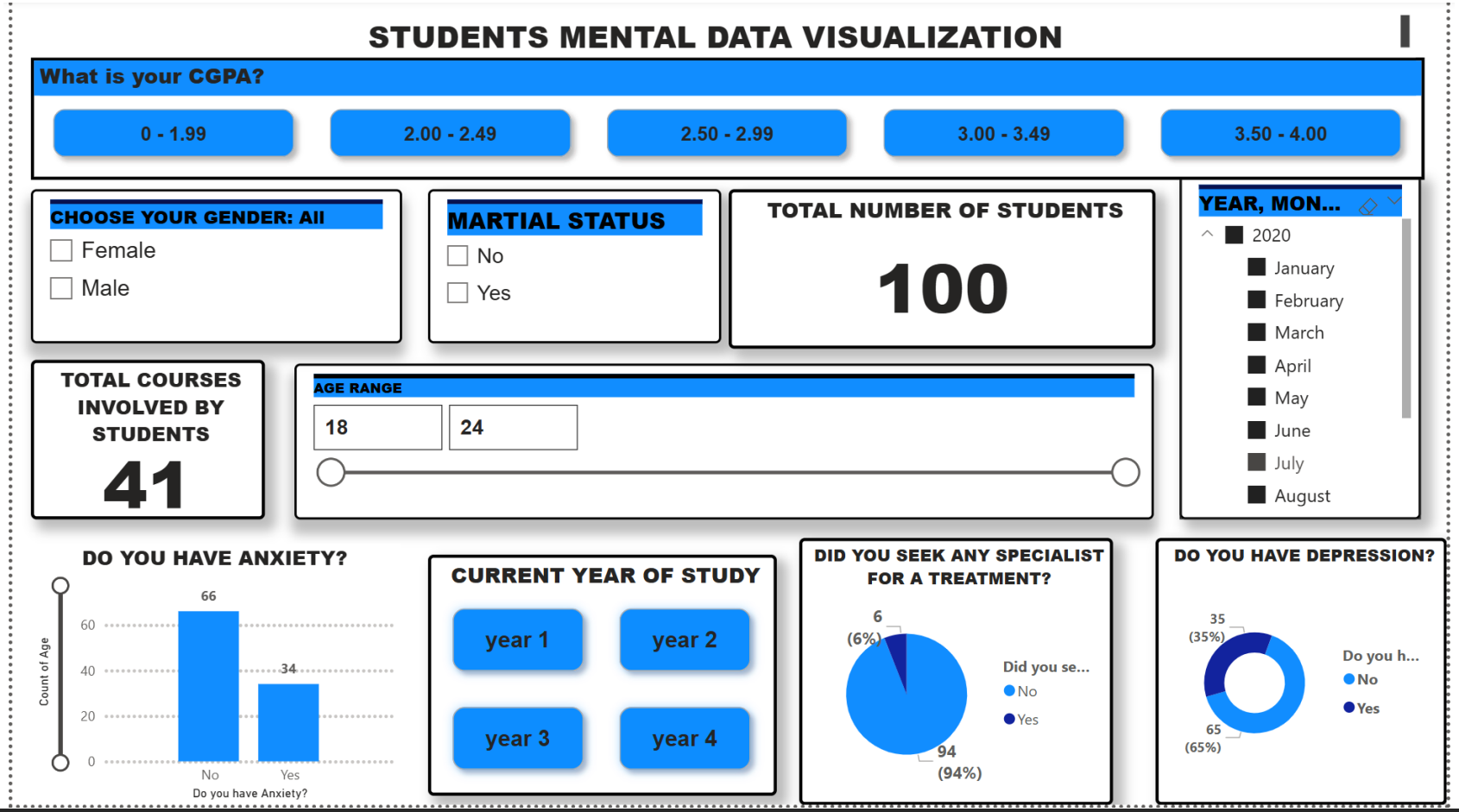

```
[89]: # Check for missing values
      df.isnull().sum()

[89]: Timestamp                                0
      Choose your gender                       0
      Age                                      1
      What is your course?                     0
      Your current year of Study               0
      What is your CGPA?                       0
      Marital status                           0
      Do you have Depression?                  0
      Do you have Anxiety?                     0
      Do you have Panic attack?                0
      Did you seek any specialist for a treatment?  0
      Month                                    0
      Weekday                                  0
      Hour                                     0
      dtype: int64

[91]: # Check for duplicate rows
      duplicates = df[df.duplicated()]

      # Print number of duplicates and preview them
      print(f"Number of duplicate rows: {duplicates.shape[0]}")
      print(duplicates.head())

      Number of duplicate rows: 0
```

ANALYSIS AND INSIGHTS

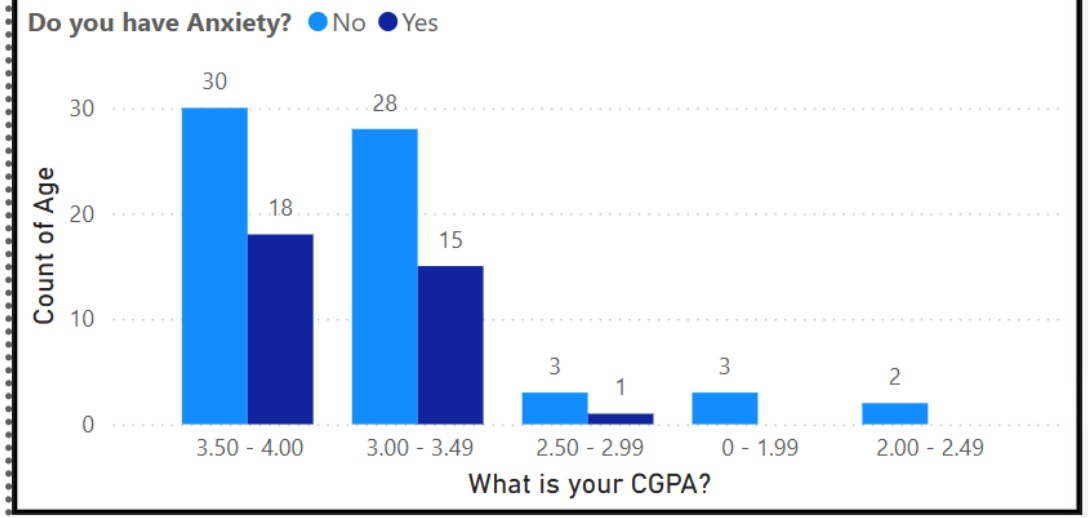
- **Platform:** Power BI connected to Azure Blob Storage
- **Steps used to link from Azure Blob storage to Power BI for visualization:**
  - Open Power BI Desktop
  - Get Data:
  - Click "Get Data" on the Home ribbon
  - Select "Azure" > "Azure Blob Storage“ then paste the copied url from Azure Blob storage properties
  - Click "Connect"
- **Main Analytics Performed:**
  - Anxiety prevalence across student CGPA and age groups
  - Age Distribution across Academics Years.
  - Mental Health Treatment seeking behavior across Academics Years
  - Mental Health trend across Academic year based on Panic Attack, Depression rate, and age groups



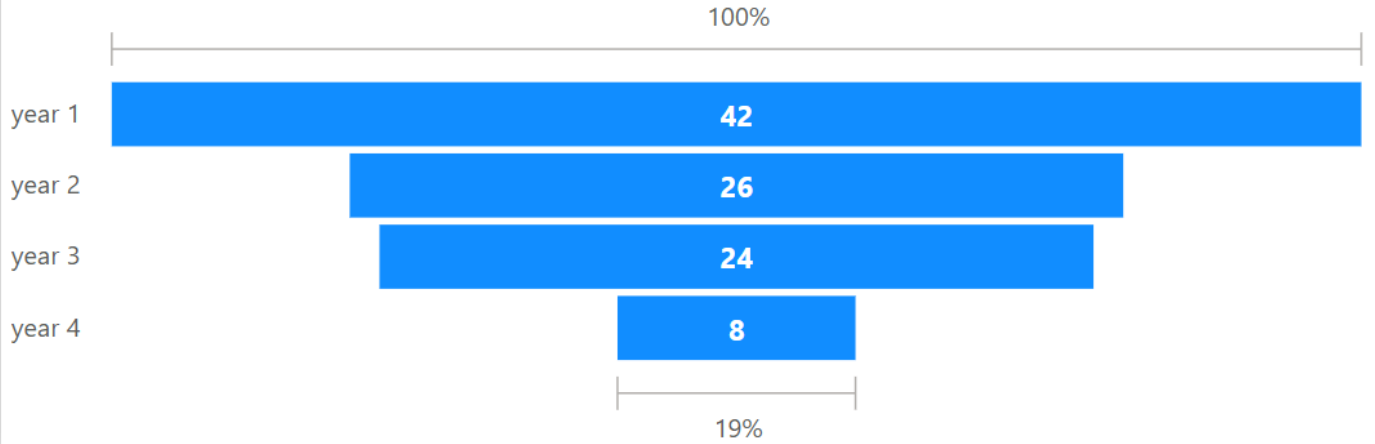
VISUALIZATION



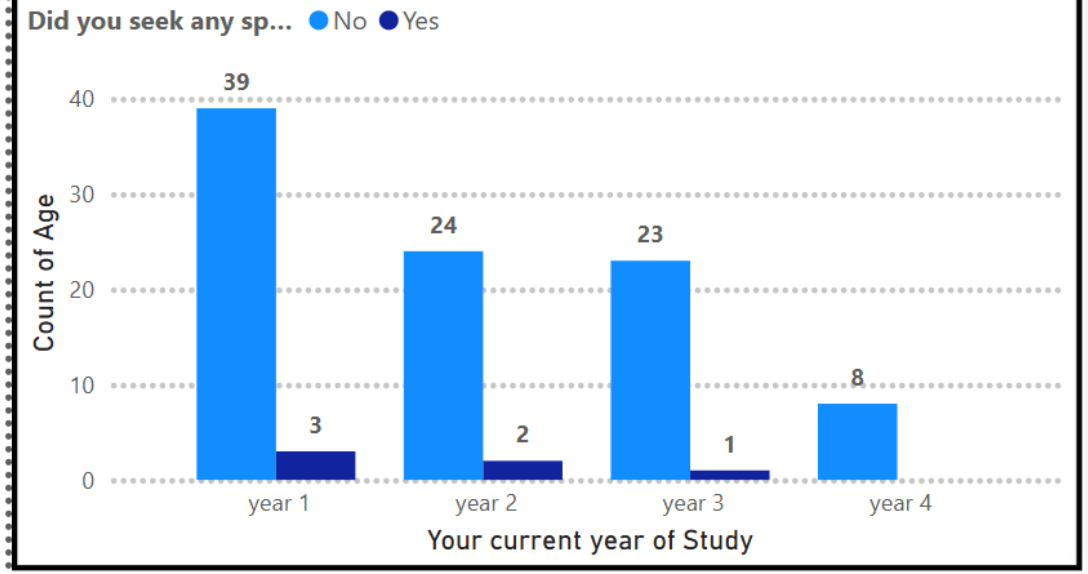
WHAT IS YOUR CGPA BY COUNT OF AGE ? and DO YOU HAVE ANXIETY?



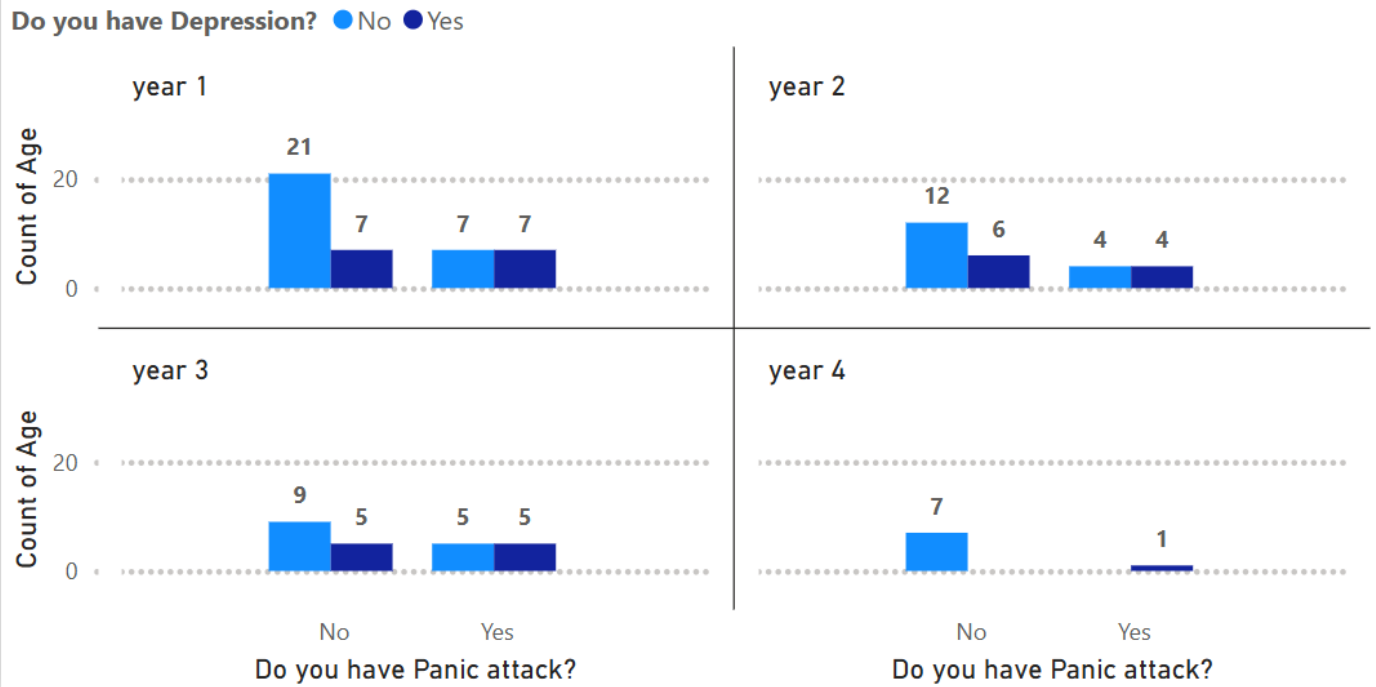
COUNT OF AGE BY CURRENT YEAR OF STUDY



Count of Age by Your current year of Study and Did you seek any specialist for a treatment?



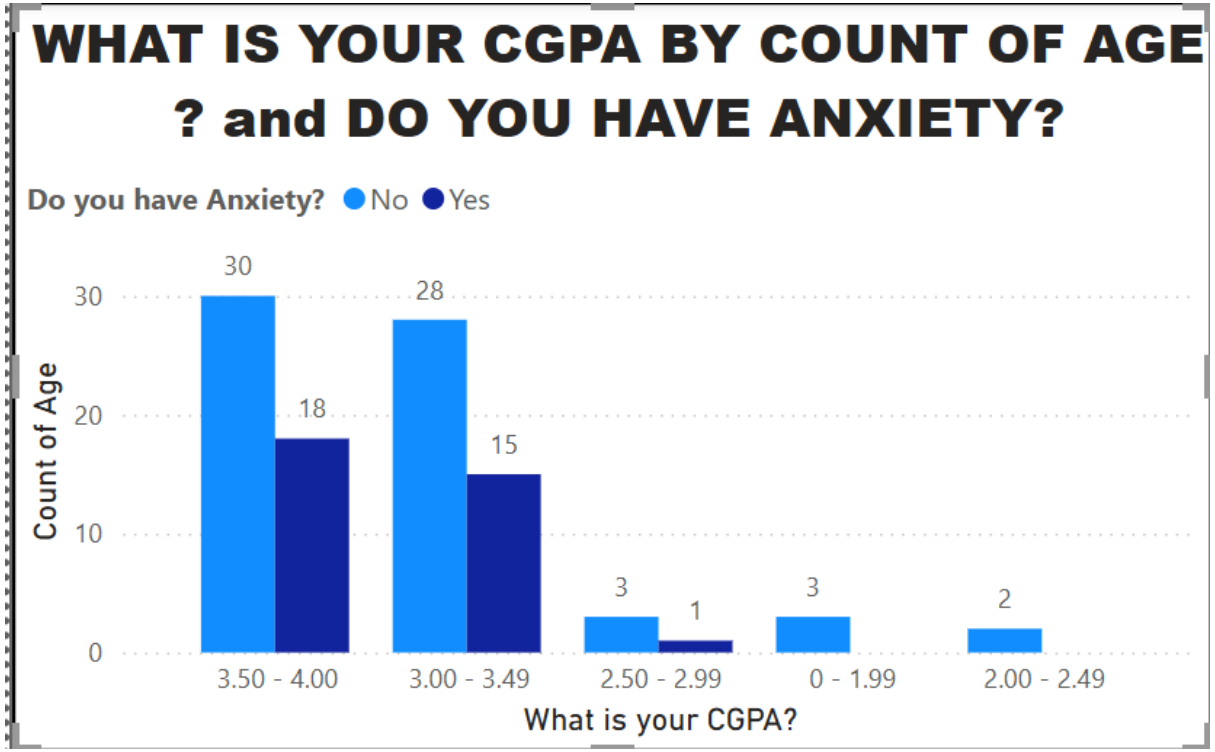
DO YOU HAVE PANIC ATTACK BY AGE COUNT?, DO YOU HAVE DEPRESSION? and YOUR CURRENT YEAR OF STUDY



INSIGHTS

ANXIETY PREVALENCE ACROSS STUDENT CGPA AND AGE GROUPS

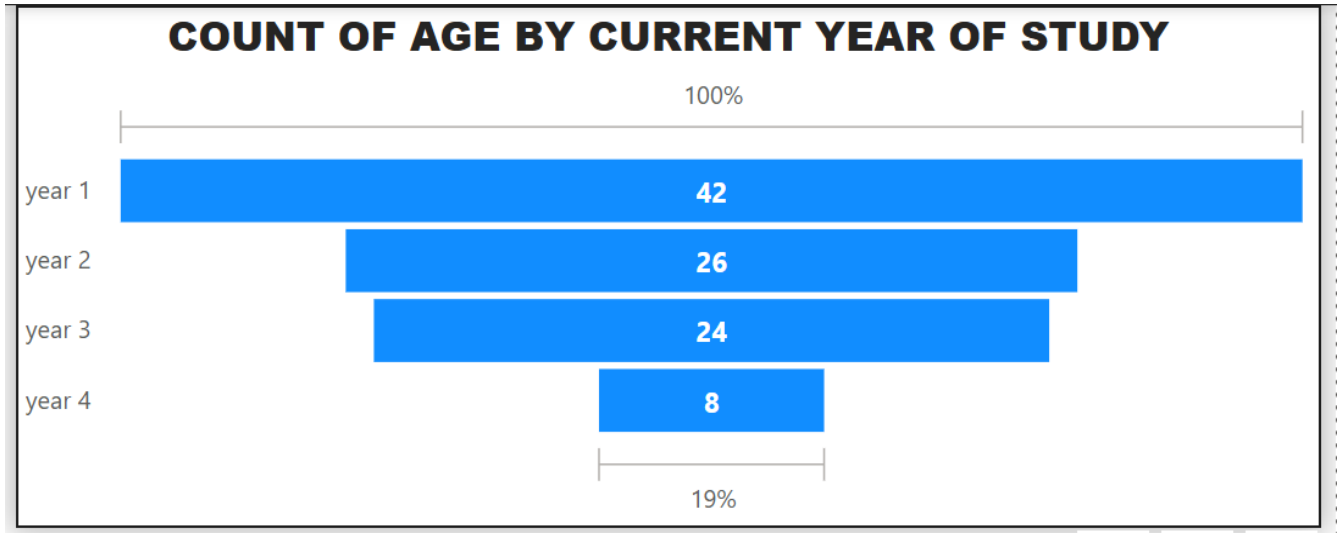
Goal: *To assess whether academic performance (CGPA) and age correlate with self-reported anxiety levels.*
Insight: *Respondents with higher CGPAs (3.50–4.00) dominate the sample, but the chart lacks clear anxiety trendlines—suggesting further breakdown by age is needed.*
Visualization: *A clustered column chart showing CGPA ranges (x-axis) vs. age count (y-axis), with anxiety (Yes/No) as color-coded segments*



INSIGHTS

AGE DISTRIBUTION ACROSS ACADEMICS YEARS.

Goal: *To understand how student ages are distributed by year of study (e.g., fresh to senior).*
Insight: *Year 1 has the smallest age diversity (only 8 students, 19% of the sample), while Years 2–4 dominate the population—suggesting higher retention or enrollment in later years.*
Visualization: *A stacked bar chart showing proportions of students per year (Years 1–4), with age count annotated (e.g., "Year 1: 8 students, 19%").*



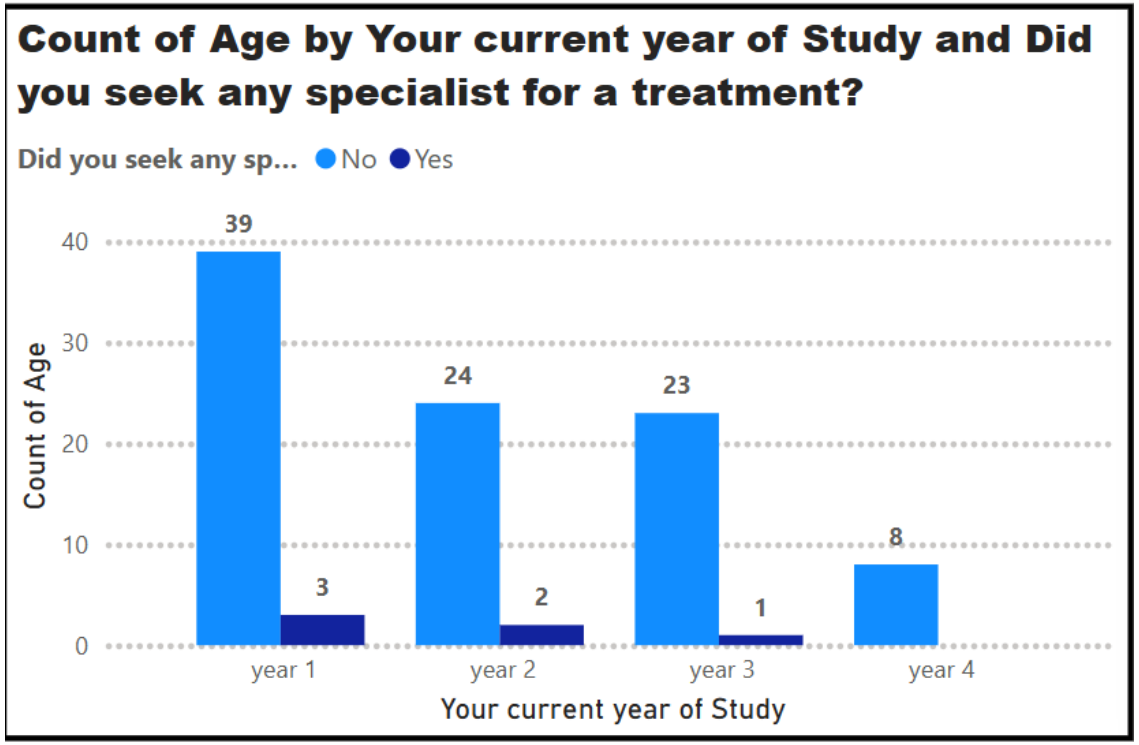
INSIGHTS(Continuation)

MENTAL HEALTH TREATMENT SEEKING BEHAVIOR ACROSS ACADEMICS YEARS

Goal: *To analyze whether students' likelihood of seeking specialist treatment for mental health varies by their year of study (Year 1 to Year 4) and age count distribution.*

Insights: Year 1 Dominance and Year 4 having the least

1. *Year 1 has the highest count of students (39 "No," 3 "Yes"), suggesting either lower treatment-seeking behavior or better mental health in freshmen.*
2. *No student seek any specialist for treatment in year 4 at all(8 “No”, 0 “Yes”*



INSIGHTS(Continuation)

MENTAL HEALTH TREND ACROSS ACADEMIC YEAR BASED ON PANIC ATTACK, DEPRESSION RATE, AND AGE GROUPS

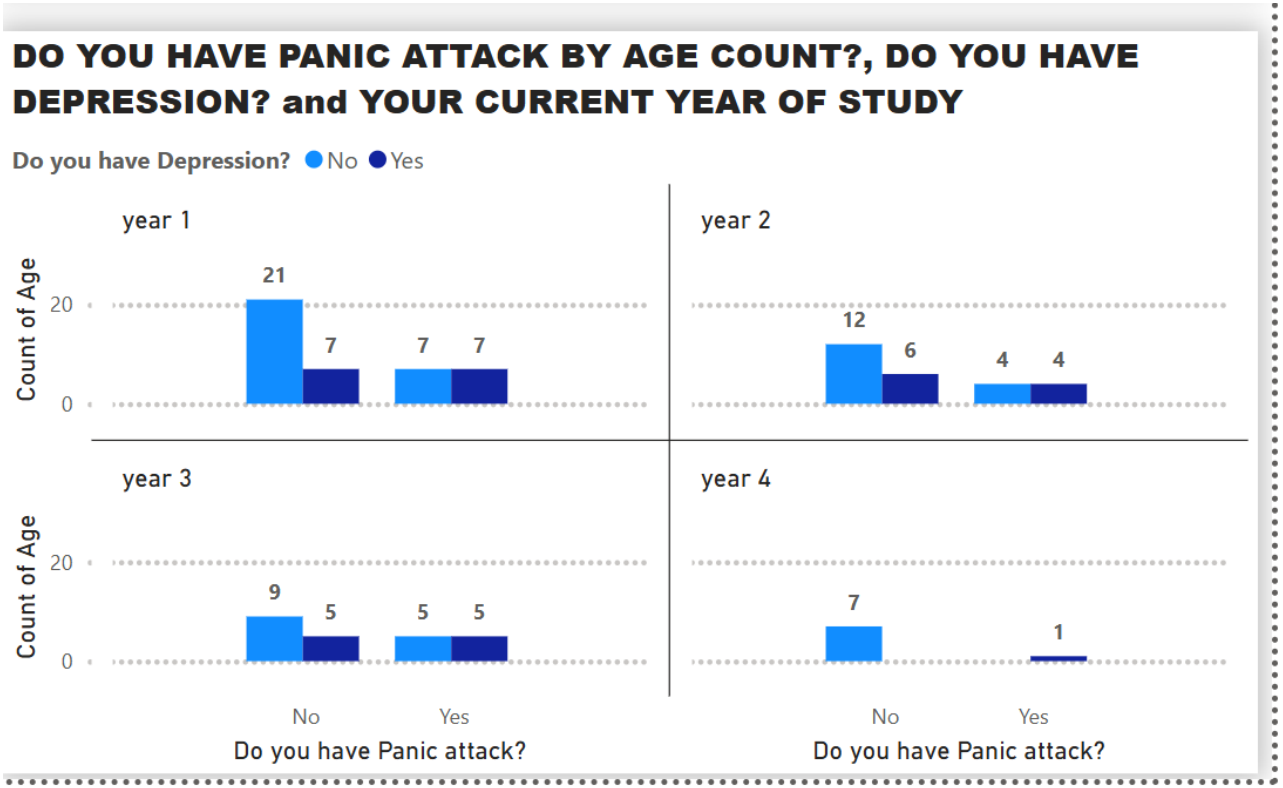
Goal: *To examine the prevalence of panic attacks and depression among students in different academic years (Year 1 to Year 4) and how these correlate with age demographics.*

Key Insights: Year-Specific Mental Health Trends:

1. ***Year 1:*** *Shows a high count of students reporting no depression (21 "No") but significant panic attacks (7 "Yes").*
2. ***Year 4:*** *Lowest sample size, with mixed depression (7 "No,") and panic attack (1 "Yes") responses.*

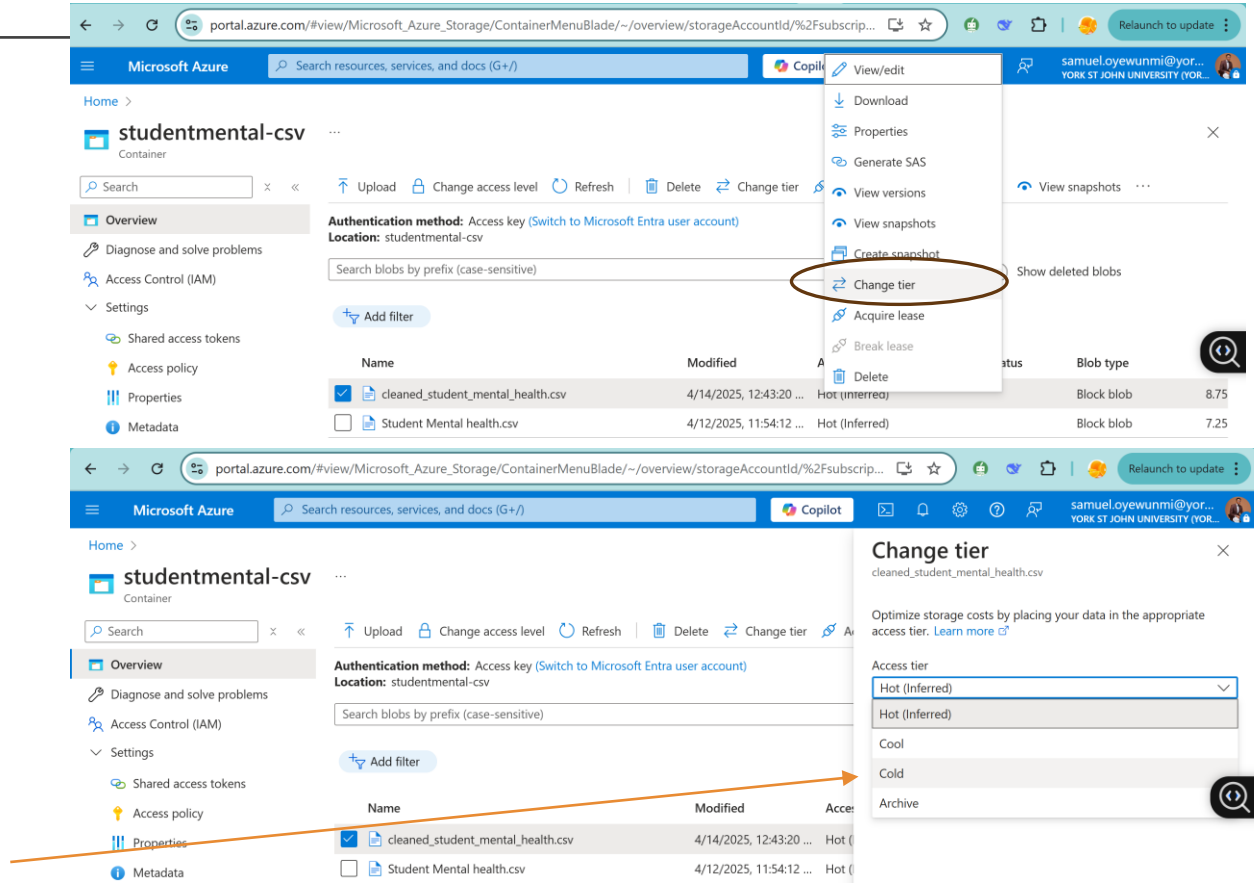
Potential Patterns:

1. ***Panic Attacks:*** *Peaks in Years 1 and 3, suggesting stress during transition years (e.g., freshman adjustment, junior-year pressures).*
2. ***Depression:*** *Higher in Year 2 (33% "Yes") compared to Year 1 (25% "Yes"), possibly linked to academic workload or social factors.*



COST OPTIMIZATION STRATEGIES

- **Methods Implemented:**
  - **Cold Storage:** Archived data stored in Azure Blob Cool/Archive tiers to minimize long-term storage costs.
  - **Reserved Instances:** Pre-purchased capacity for Azure Data Factory and Synapse reduces compute costs significantly.
  - **Auto-Pause Synapse:** Automatically pauses resources during inactivity to avoid paying for idle time.
  - **Power BI Free Tier:** Utilized free Power BI features to eliminate reporting and dashboard costs.



SECURITY AND COMPLIANCE

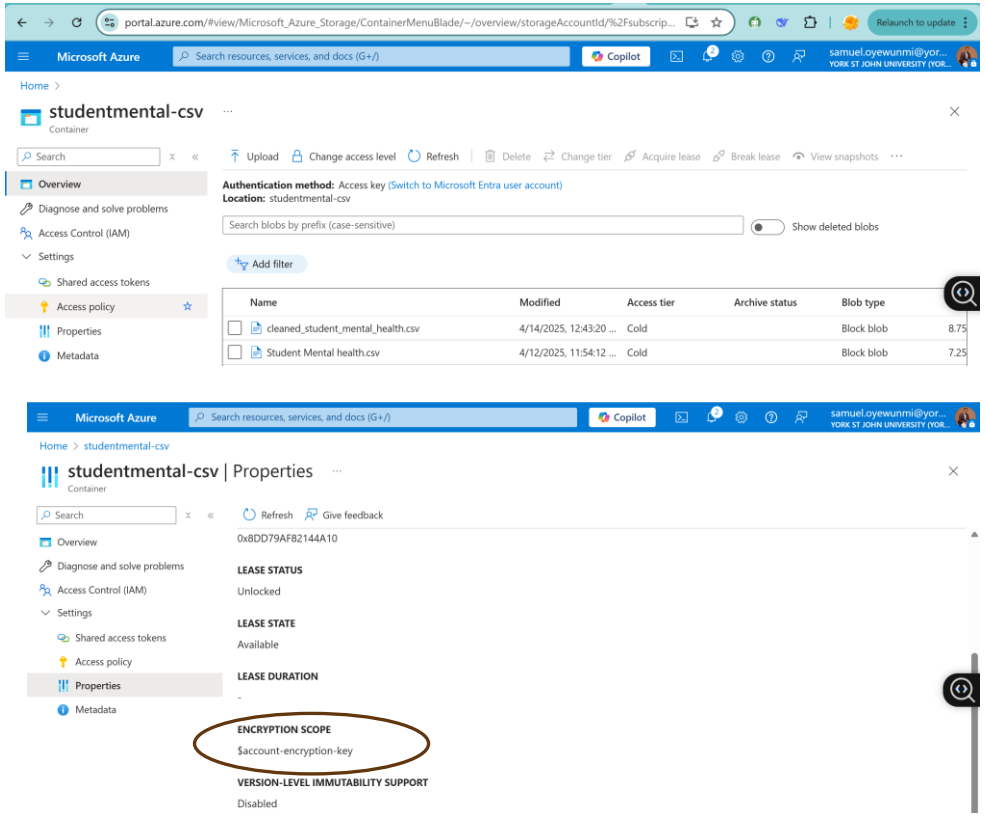
Security Features

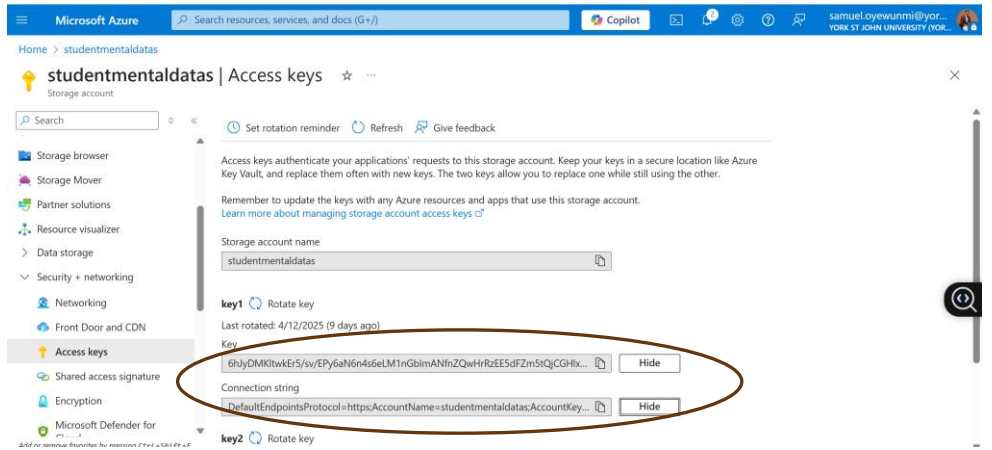
- **Role-Based Access Control (RBAC):**
  Access to data and Azure services is tightly managed based on user roles, ensuring that only authorized personnel can view or manipulate sensitive student mental health records.
- **Azure Key Vault:**
  Secrets such as database credentials and API keys are stored securely using Azure Key Vault, preventing unauthorized access and enhancing system integrity.
- **Blob-Level Encryption (AES 256):**
  All data stored in Azure Blob Storage is encrypted using AES 256-bit encryption, a high-security standard that ensures confidentiality of stored student data.

Compliance Considerations

- **GDPR Compliance:**
  Data storage and processing are aligned with GDPR regulations, ensuring that students' personal data is handled legally and ethically, with consent and privacy in mind.
- **Custom Access Policies:**
  Fine-grained access control policies are defined to safeguard student data, limiting exposure based on academic roles (e.g., counselor, admin) and preventing data misuse.

SECURITY AND COMPLIANCE





This shows the access key and confidential key that can only be used for authentication purpose of the data stored in the storage

PERFORMANCE MONITORING AND MANAGEMENT

Monitoring Tools

- **Azure Monitor:**
  Provides real-time visibility into the health and performance of Azure resources like Data Factory, Storage, and Synapse.
- **Log Analytics:**
  Aggregates and queries log data across services, enabling deep diagnostics and performance tuning.
- **ADF Activity Runs:**
  Tracks the success/failure of each step in Azure Data Factory pipelines, helping identify delays or data flow issues.

Troubleshooting Strategy

- **Alerts on Pipeline Failures:**
  Automatic alerts are triggered on failures, ensuring fast response and minimal downtime.
- **Diagnostic Logs Enabled:**
  Logs are captured for all key services, making root cause analysis efficient and auditable.
- **Data Quality Checks Post-Ingestion:**
  Validations are performed after ingestion to catch missing or corrupt values, ensuring the reliability of insights derived from student data.

PERFORMANCE MONITORING AND MANAGEMENT

Monitoring Tools

- **Azure Monitor:**
  Provides real-time visibility into the health and performance of Azure resources like Data Factory, Storage, and Synapse.
- **Log Analytics:**
  Aggregates and queries log data across services, enabling deep diagnostics and performance tuning.
- **ADF Activity Runs:**
  Tracks the success/failure of each step in Azure Data Factory pipelines, helping identify delays or data flow issues.

Troubleshooting Strategy

- **Alerts on Pipeline Failures:**
  Automatic alerts are triggered on failures, ensuring fast response and minimal downtime.
- **Diagnostic Logs Enabled:**
  Logs are captured for all key services, making root cause nalysis efficient and auditable.
- **Data Quality Checks Post-Ingestion:**
  Validations are performed after ingestion to catch missing or corrupt values, ensuring the reliability of insights derived from student data.

SUMMARY AND CONCLUSION

- Microsoft Azure offers scalable and secure tools for data management
- Effective insights can help universities address student mental health proactively
- Project demonstrates technical capability, business alignment, and cost-effectiveness