IMPROVING ACADEMIC OUTCOME FOR SECONDARY SCHOOL EDUCATION

Emmanuel Aminu- Data Engineer
Amole Oluwaferanmi - Data Analyst(Team Lead)
Motunrayo Adeyemi - Data Scientist

School Background:

Federal Government College (Unity School) Lagos is a boarding school in Ijanikin, with about 12 Classrooms for the Senior School, 1 E-Library and 2 computer Labs (sponsored by CISCO), so every student have access to it from onset. They have about 2 labs each for Chemistry, Biology, and Physics and also have trading subject labs

The Senior School has three different disciplines, which falls into Art, Commercial and Science (although for the Science Discipline, students interested in Tech have the option of not seleting Biology). For the purpose of this work, we focussed on three Grades, SS1 to SS3 students, there are about 1400 students in total.

Data Collection Provision (Data Collection Plan: Outline the methods and tools used for data generation and collection.)

Each Classes has a classroom teacher and each students once admitted in SS1class would maintain the same class members until they pass out. So this made it easy to access historical records of the students from the classroom teachers, provided below are the list of data collected from the classroom teachers:

1. Students Bio data: The school gets info about Students and Parents during the time of Admission, and through PTA, with importance on knowing proof of funds, to avoid a student been properly taken care of.
2. Students Historical Results, and Current Results: Results on Subjects and Performances
3. Student Info (A survey done to understand Class struggles, and other informations, including attendance rate, which is high since it is a boarding school)

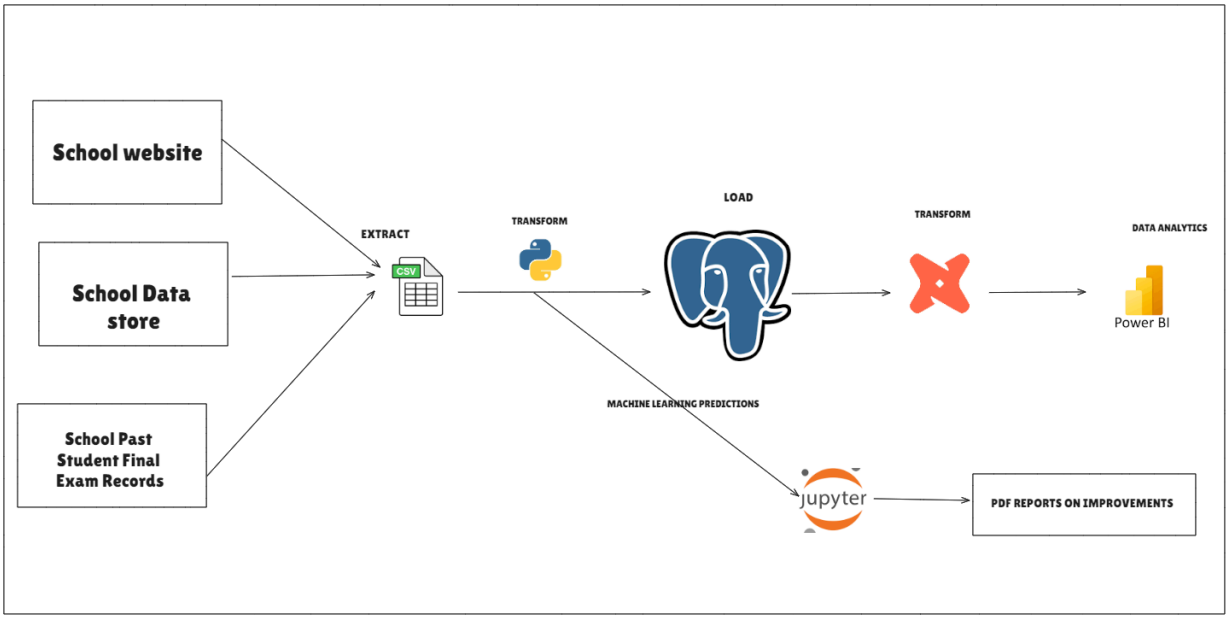
The school provided Data about their past WASSCE students/results data, and JAMB results data. Which is saved in an excel sheet.

Data Collection Plan and Data Pipeline Description:

The top-down methodology to data generation, first of all, focused on the problem we were trying to solve, from that we narrowed it to the specific data appoints available for Federal Government College (Unity School) Lagos.

With the data available to us, we decided to start working with what we had, for sourcing of the student results, the school has a portal whereby each student can access their results from: https://fgclagos.myskoolportal.com.ng/login

We sourced from the backend, which we had raw csv files. Below is the Pipeline we followed:

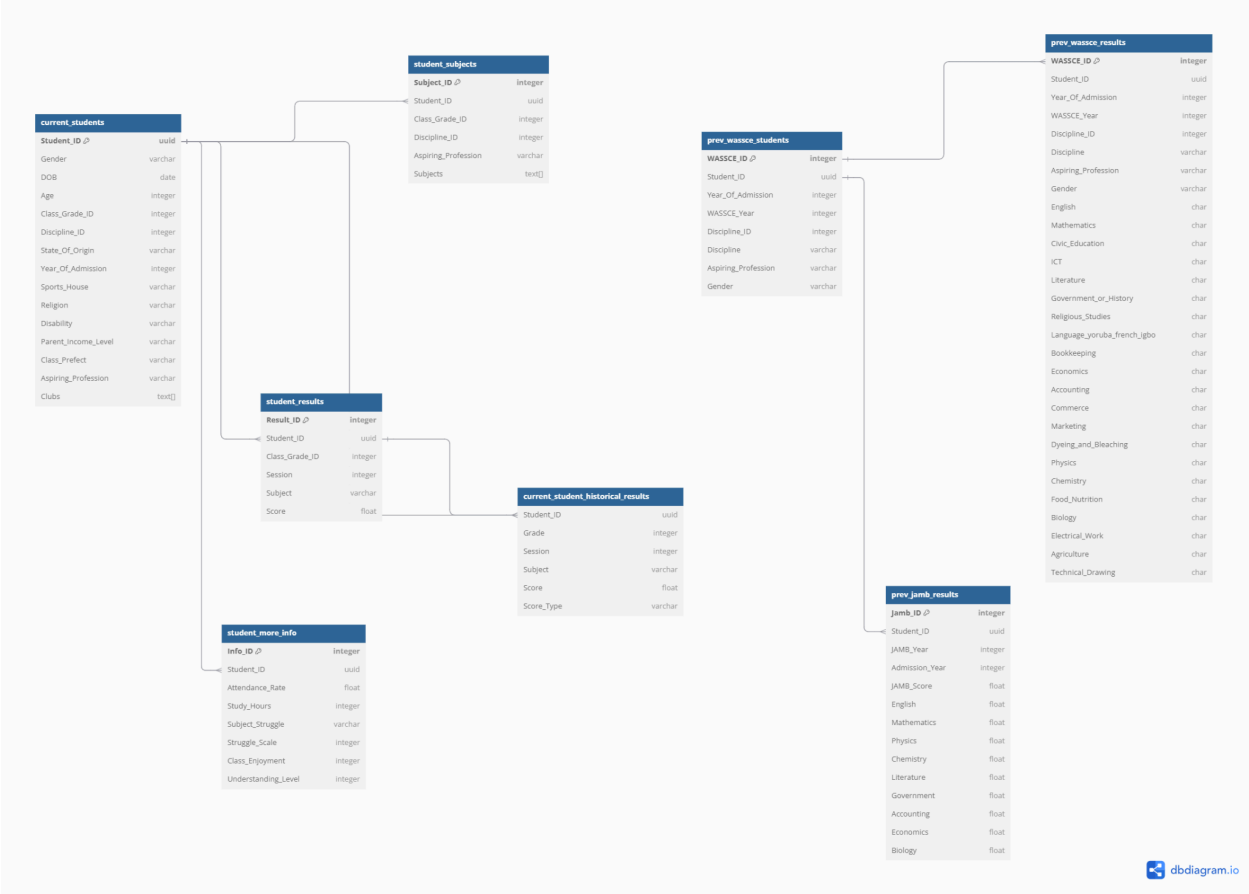


Data Pipeline of Ignite Datathon

Data warehouse dictionary: Provide detailed information about your data warehouse design, including database, schema, tables, relationships, data types etc

Warehouse Architecture:

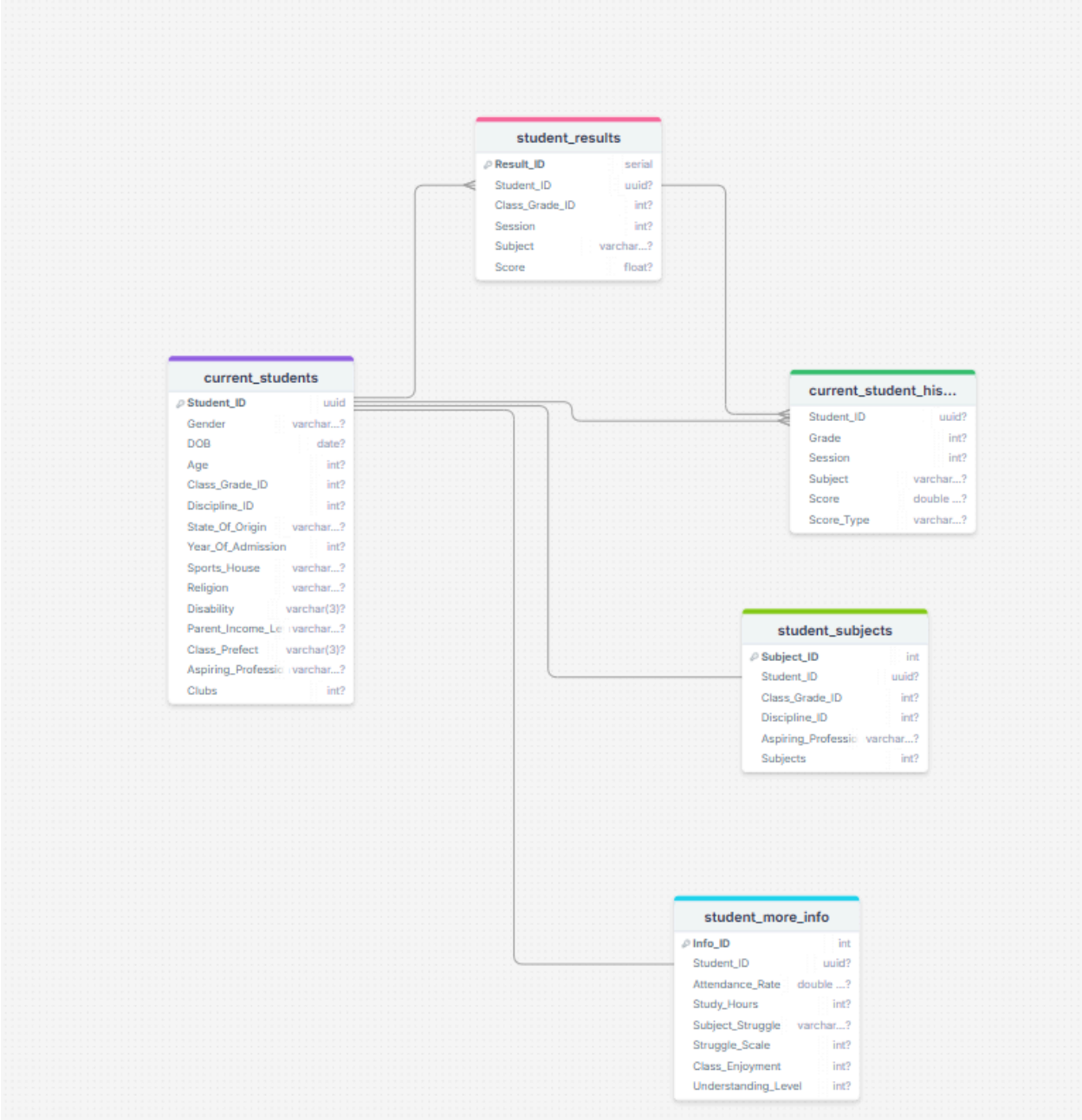
The data warehouse is a local data warehouse, which is spinned up ny docker ,on port 5434, we created a database named unity_school_datawarehouse. With the schema named dwh, first thing was to generate the data based on the right information, and give each students a unique id, using UUID. The overview of th schema is below:



Overview of Unity_School_DataWarehouse

In more details we first focused on the current students, their bio data, results, subject, other info such as attendancerate from class teachers, and their study hours (timed by the hours they read during the provided school Prep time in the afternoon and also in the Night). We created a survey on how much the student understood their courses, classes engagements, and the teachers gave ideas on understanding level.

Due to the time factor of this hackathon, we created these datas. The database design is shown below. Our design shows One to Many, and Many to Many Relationship as the DIM table such as the Student_Subjects, Current_Students lead to FCT such as Student_Results, and Historical_Student_Results.



School Student Relationships

More details on the Table

Current_Students: Comprises of the Student_ID, Gender, Date of Birth, Age, Class_Grade_ID (SS1 to SS3), Discipline_ID(1 is for Arts, 2 is for Commercials, 3 is for Scence), State of Origin, Year of Admission, Sports_House (This houses also serve as dorm rooms), Religion, Disablity (Yes or No), Parent_Income_Level (This was filled by the Parents during admission with the Parent Teachers Association Body, shared with the school: Low Income, Middle Class, Upper

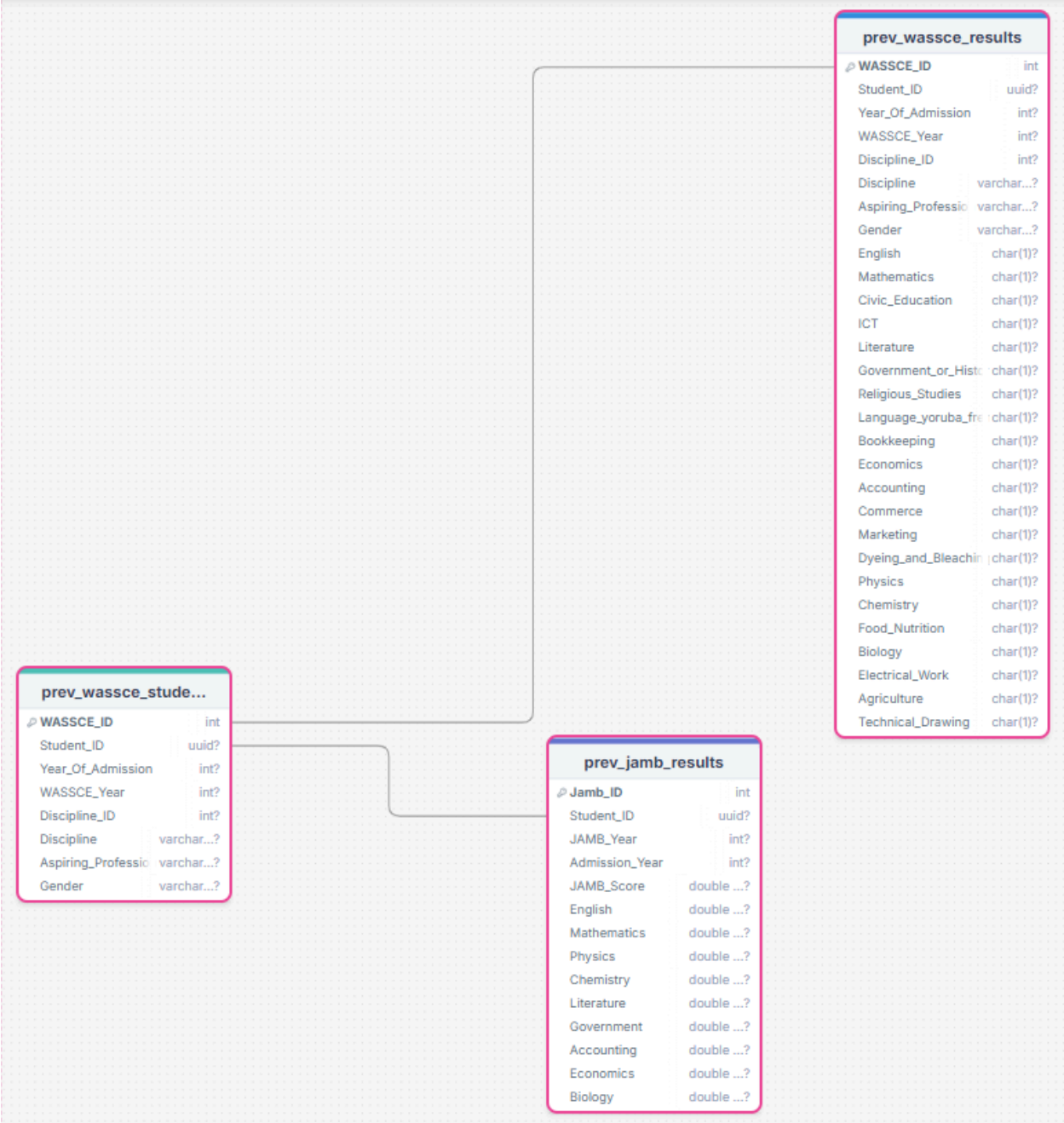
Class), Class Prefect (Awarded mostly to 10% of the SS3 (Grade 3) students), Aspiring Professions and Clubs.

Student_Results: Comprises of Student_ID, Class_Grade (1 for SS1, 2 for SS2 and 3 for SS3), Session (1 for 1st Term, 2 for 2nd Term and 3 for 3rd Term), Subject and Score. This provides results of the current students and their scores across various classes and the corresponding term.

Current Student Historical Results: This table contains columns such as: Student_Id (unique identifier across the different table), Grade (1 for SS1, 2 for SS2 and 3 for SS3), Session which serves as the term (1 for 1st term, 2 for 2nd term and 3 for 3rd term), Subject, Score and Score type (Current for the student's current score, Historical_Grade_1 for the student's scores when they were in SS1 and this applies to students from SS3 and SS2 then we have Historical_Grade_2 which shows the student's score in SS2 and this applies to students in SS3). Essentially, this table shows the performance of students from when they got into the school up until the point they are now i.e for SS3 students, we get to see their performance in SS2 and SS1 and so on.

Student Subjects: Basically, this tables contains Student_ID, Class_Grade_ID, Discpline_ID, Aspiring_Profession, Subjects (this is in array and contains all the subjects a student is taking right from SS1 till SS3).

Student_more_info: Consists of columns such as Student_ID, Attendance_Rate with max value as 1, Study_Hours which ranges from 1 - 5 and more if applicable, Subject_Struggle which essentially talks about the subjects students are struggling wit, Struggle_Scale (1 - 5), Class_Enjoyment (how well students enjoy their classes in general on a scale of 1 - 5), Understanding_Level (1 - 5 which shows much much students understand what's being taught in their respective classes).



Previous WASSCE and JAMB Student Relationships

More details on the Table

Prev_WASSCE_Student: Provides details for of old students who did their WASSCE in this school and it consists of column such as: Student_ID, Year_Of_Admission, WASSCE_Year (the year each student took their WASSCE exams), Discipline_ID (1 for Art, 2 for Commercial and 3 for Science), Discipline (Art, Commercial and Science), Aspiring_Profession and Gender.

Prev_wassce_results: With the **prev_wassce_students** providing some information about the students, the **prev_wassce_results** is providing details of the score of each of the student and contains columns such as: Studen_ID, Year_Of_Admission, WASSCE_Year, Discipline_ID, Discipline, Aspiring_Profession, Gender and the student's grade (A - F) in various subjects.

Note: This table was further transformed.

Prev_jamb_results: Contains columns such as the Student_ID, JAMB_Year, Admission_Year, JAMB_Score (total score over 400) and columns coontaining scores in the various subjects each student did.

This was basically the process it took during data generation, warehouse design and DB creation, schema, tables and relationships.

Analytical Models: Present the predictive models (if any) and analysis performed on the data.

JAMB Models: Using machine learning models (RandomForest, XGBoost Regressor, and LightGBM) we were able to use this models to predict how the final outcome of the students can be like based on historical analysis of their results, and other features, such as Understanding level, Study Hours, Disability (health issues), parental income, Class struggle scale, and subject struggle.

We merged the old students JAMB results based on Discipline with the historical results of current students, to find the average normalized scores to determine their particular average scores, by creating a Syntethic JAMB Score,

XGBoost performed well under this circumstances as shown below:

```
------------------------------
XGBoost Model:
  Mean Squared Error: 51.01
  Mean Absolute Error: 5.16
  R² Score: 0.94
------------------------------
LightGBM Model:
  Mean Squared Error: 84.31
  Mean Absolute Error: 6.71
  R² Score: 0.90
```

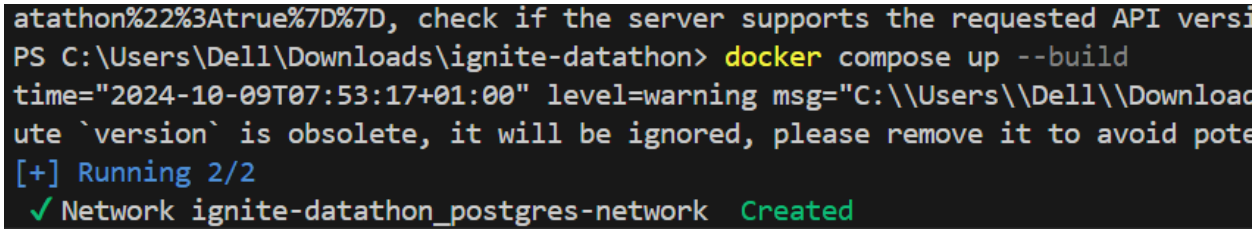
For any student at risk of scoring below 200 we made the model cluster those students out and send reports to the teachers on what to do about through PDF reports.

WASSCE Models: This followed a different approach by only using the current historical results and applying other features using only XGBoost model, although the model wasn't perfect for prediction, it gave good view of students who perform well, and sends report to the school.

Final Solution: Build a solution (report, app, visual, web page etc) that highlights the key insights, shows trends and correlations, and provides actionable recommendations for stakeholders to improve the performance of candidates.
For this part of the datathon, we decided to create a Power BI report to visualize our data to draw insights and proffer recommendations to the school on what to do to improve the grades of their students.

Interact with the report **here.**

Approach Taken:
To connect to the DB, the github repo which contains all that we did was clone into VS Code then a line of code was run in the terminal after installation of **docker** to create a local connection.



Line of code run (docker compose up –build)

The local connection was created in docker and so we could connect to the PostgreSQL DB that was created by inputting the server and database name in the next prompt.

Prompt from PBI (Server: 127.0.0.1:5434 and Database: unity_school_datawarehouse)

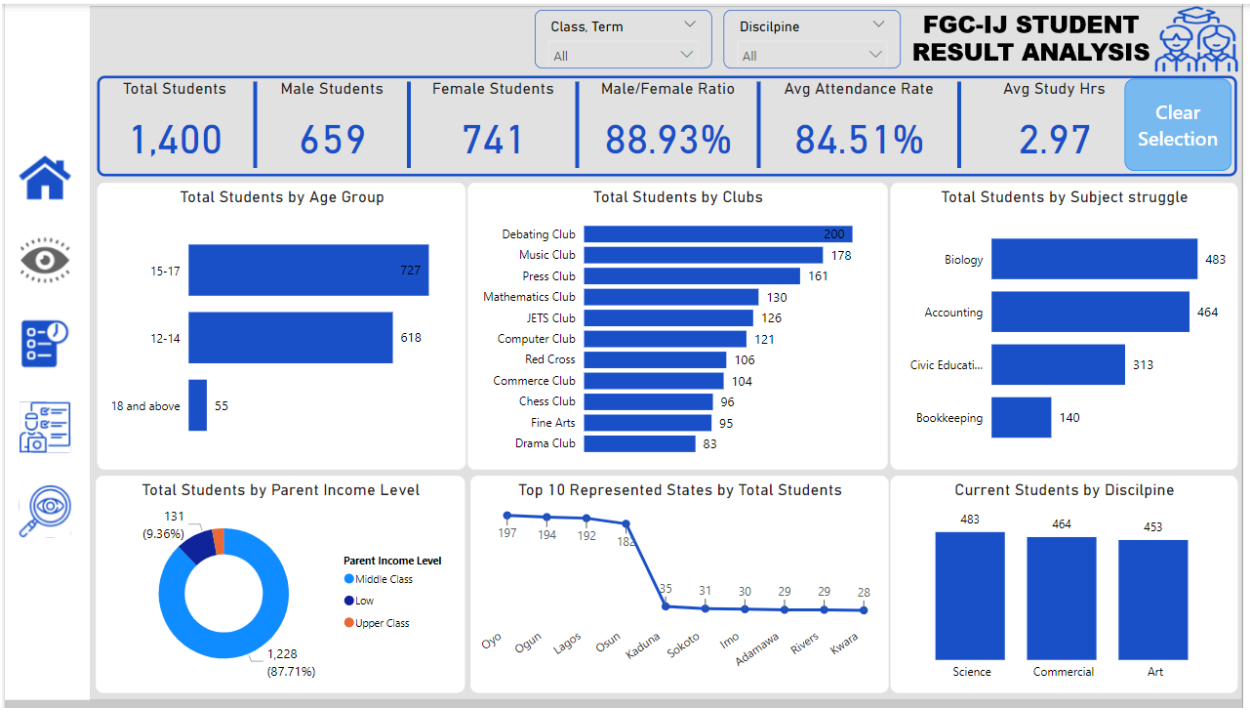
Then the next popup comes in where we connect to the tables which were in .csv format and did some transformations such as:
Current_students table and the **student_more_info** was merged as new to basically serve as our fact table and contains all the information of the current students.

Also, **Prev_wassce_students** and **Prev_wassce results** were also merged as one table.

The Data Engineer did a great job making sure the data was clean so after the merging, we checked the datatype for each of the columns to make sure they all have the right format, then some values were replaced for better understanding by users.

In this solution, we decided to take the approach of creating an overview page which basically shows a summary of our tables across board and it displays informations such as: some important KPIs (total students, male and female students, male/female ratio, average attendance rate and the average study hours).

We further created created some important visuals to make sense of our data by showing the age range of students in this school, the parent income level of the students, the clubs the students participate in, the subjects students struggle with, the state that had representation in this school and considering this school is in Lagos (South-West), more students form south-west states were represented here, the number of students in various discipline.



Overview Page

We also created a page to show the performance level of the current students in the school which we called the **Current Result Analysis** page which basically displays the detailed performance of the schools's current student across various discipline.

We also created a page for the **historical result analysis** where we displayed the the historical results of students from SS1 up until SS3 to see how their performance in the school's internal exams affected their WAEC and JAMB grades.

Finally, a page was dedicated to the Insights and Recommendations that would be given to the school to better improve the scores of their students.

Note: This solution can be created as an app or website for teachers and parents to see their students/ward's performance from time to time to help them have an idea of which areas they need to further help their student/ward to boost their grades across board.

SCIENTIFIC APPROACH

The dataset is split across 10 sheets, each bringing different insights into the student experience.
The data contains:
Demographic data like Gender, Disability, and Parent Income provide context about a student's background.
Academic performance data such as Score, Subject, and Attendance Rate give us a glimpse into their achievements and areas of struggle.
Behavioral data like Study Hours and Struggle Scale help us understand the students' effort, resilience, and challenges they face on a daily basis.
This data spans from 2019 to 2022, giving us a wide time frame to analyze performance trends.
With Student_ID serving as the common key across all sheets, we can combine data sources into a unified view for each individual.

Historically, interventions often occur too late after a student has already failed or dropped out. Since we can build a predictive model, we can shift this paradigm. Identifying students who are at risk of poor performance early on, educators can provide personalized interventions that increase the likelihood of success. The challenge, however, lies in the diversity of factors that

contribute to performance. From the socioeconomic status especially in the context of Nigeria to attendance rates, each column in the data tells part of the story.

Exploratory Data Analysis (EDA)

Correlation Analysis: It was observed that Study Hours had a strong positive correlation with Score, while Struggle Scale had a moderate negative correlation with Score. This highlights the dual role of both effort and difficulty in student performance.
Parent_Income and Attendance: Students from lower-income families had higher absenteeism rates, reinforcing the importance of considering socioeconomic factors in the predictive model.
Supporting column for prediction includes; studyhours, attendance_rate and parent's income.

Modeling

We employed the Linear regression Model which splitted the data into 70% to train the data and 30% to test the data.

A threshold was set for the score which was;
Pass: If predicted score > 50
Fail: If predicted score ≤ 50

Below is the link to the code:
http://bit.ly/3U0Z40c

RESULT AND INTERPRETATION

TOOLS AND TECHNOLOGY USED
Python is the major tool used in this project. However, other libraries, such as pandas, numpy, and sci-kit, were employed, Git-hub. The working environment was a Jupyter notebook via VScode.

MODEL EVALUATION
The accuracy of the model was evaluated using the outcome of the r-value

CONCLUSION AND STEPS TO TAKE:

- It would be great to deploy this model within the school's learning management system (LMS) to provide real-time performance prediction.

- Also, we would like to collaborate with educators to design intervention programs based on model outputs.
- We would be glad if this model can be expanded by incorporating other relevant additional data sources.