THEME: EDUCATION INFRASTRUCTURE

A Project by Team ESRGAN **Spring 24 Capstone Project**

Unlocking Potential: Exploring Factors in Student Success

Section One

Introduction

Education is the foundation of sustainable development and progress in every nation.
We use OECD’s Programme for International Student Assessment (PISA). PISA measures 15 year olds’ ability to use their reading, mathematics and science knowledge and skills to meet real life challenges. We choose the PISA Test Score to be a suitable metric to measure and compare educational outcomes in any given country and also serve as a measure of the overall quality of a country’s education system.

OBJECTIVES

This project aims to investigate the factors that influence educational outcomes across nations. We aim to understand the characteristics of nations’ education systems and prescribe strategies to strengthen fragile systems and further consolidate stable systems. In this study, we investigate data collected from 86 different countries on the following variables.

1) PISA Score (average pf maths, science and reading PISA test scores)
2) Pupil-Teacher Ratio (average number of pupils per teacher)
3) Annual Teacher Salary in USD (adjusted for purchasing power parity)
4) Government Expenditure on Education as % of Total Expenditure
5) Shortage of Learning materials Index (Average degree of lack of learning tools).

Data Collection

The raw data that constitutes the project dataset is available in the following open source databases:

1. OECD Database : https://www.oecd.org/en/data.html
2. World bank Database: https://data.worldbank.org/

Data Cleaning

1. We download the raw data in ‘.csv’ file formats and load them into pandas dataframes.

2. We remove columns containing redundant information and rows that were not representing countries.
3. We handle missing values by investigating shifts in the distribution of variables after filling missing values. Distributions that remain static before and after imputation, we fill with the median value. For distributions that shift after imputation we search and add the actual missing values from external data sources.
4. We do not alter outliers to avoid biassing the data.
5. We join multiple variables into a single DataFrame before exporting as a '.pdf' file.

Feature Engineering

- We create a new feature to convey information on the efficiency of government spending by computing the ratio between PISA Scores and Government Expenditure on education. We call this feature, ‘Spending Efficiency’.
- We observe a high collinearity between PISA Math, PISA Reading and PISA Science Scores. To avoid redundancy and improve computational efficiency, we create a single target variable from the average of maths, science and reading PISA Scores.

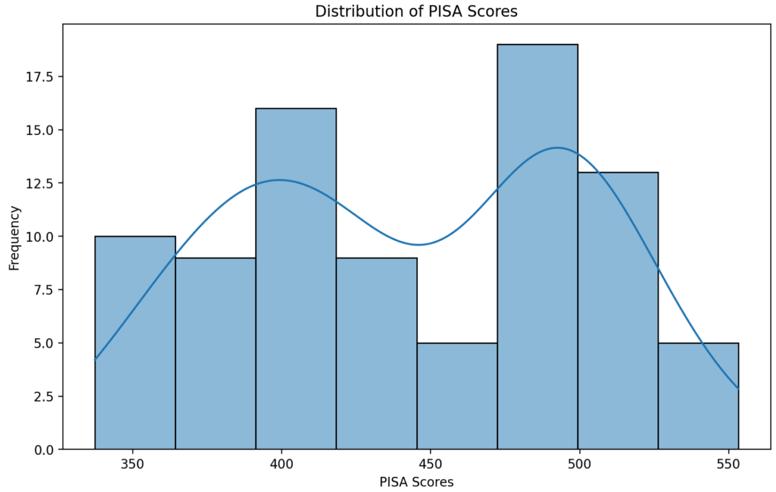
Exploratory Data Analysis

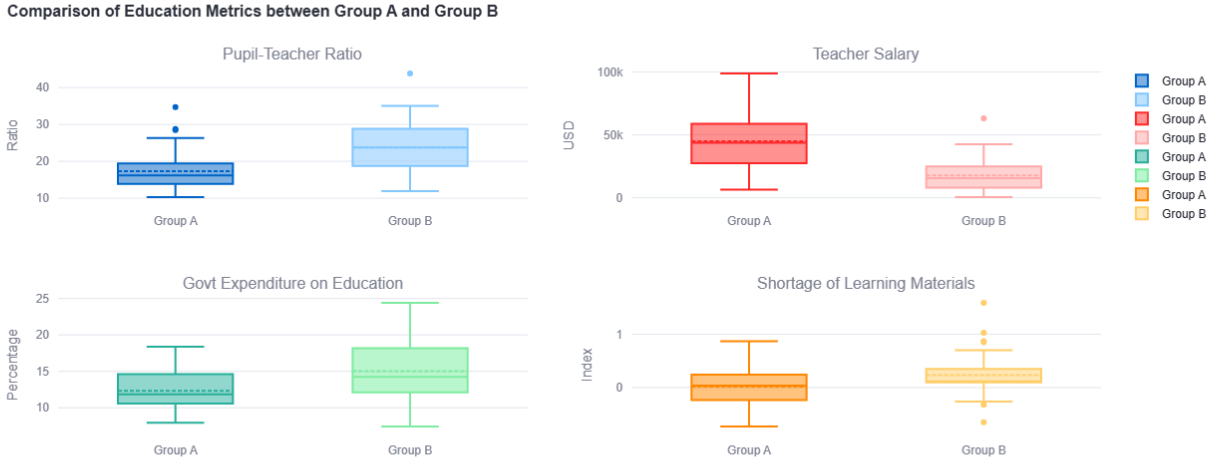
Highest vs Lowest Performers

- The top 5 performers based on PISA Scores are Singapore, China, Korea, Finland, and Japan.
- The worst 5 performers based on PISA Scores are Cambodia, Dominican Republic, Philippines, Uzbekistan, Kosovo.

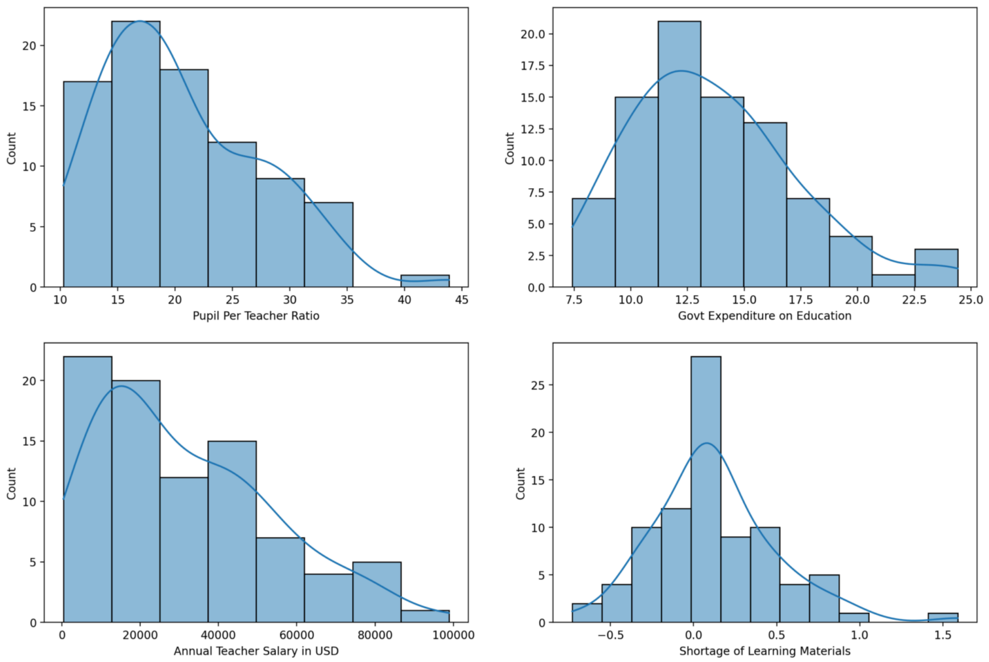
Distribution Plots

- **Pisa Scores** distribution as shown by the histogram is slightly positively skewed (mean > median) and bimodal, meaning that it has two peaks corresponding to two peaks. First peak(mode) is 395 - 405 and the second mode is 495 - 505. Most European and North American and East Asian countries fall within and beyond the second cluster (495 - 505). Most South American, West Asian and African countries fall within and below the first cluster (395 - 405). South East Asia is split between both clusters. There are no outliers for PISA Scores.



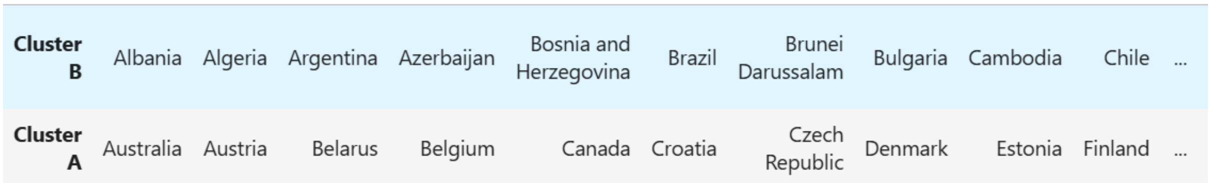


- **Pupil to Teacher ratio** distribution is skewed to the right showing that there's a global trend tending towards smaller class sizes which can be beneficial for learning outcomes. The median number of students per teacher is between 18 and 19. Cambodia is the sole outlier with a pupil-teacher ratio of 43.86 and it unsurprisingly has the lowest pisa score of 337.33.
- **Government expenditure on education** is positively skewed (mean > median), and Paraguay, Tunisia and Uzbekistan are all outliers, with governments spending more than 23% of their total expenditure on education.
- **Annual Teacher Salary** is strongly positively skewed (mean > median), and Luxembourg is an outlying nation with an annual teacher salary of over 98,000 USD.
- **Availability of learning materials index** is slightly positively skewed (mean > median). Costa Rica and Tunisia have indices of 1.03 and 1.59 respectively indicating extreme lack of learning tools while on the other end Singapore with an index of -0.73 has extreme levels of sufficiency and availability of learning tools.



Unsupervised Statistical Learning

- To investigate the bimodality of the histogram distribution of PISA Scores, we train a simple K-Means clustering machine learning model.
- We set the number of centroids (k) to be equal to the number of modes or peaks which is two.
- The model separates and labels the countries into two distinct groups, cluster A and cluster B. Each cluster reflects a mode in the histogram distribution.
- **Mean and Median PISA Scores**: Cluster A > Cluster B
- **Mean and Median Teacher Salary**: Cluster A > Cluster B
- **Mean and Median Pupil-Teacher Ratio**: Cluster A < Cluster B
- **Mean and Median Govt Expenditure on Education**: Cluster A < Cluster B
- **Mean and Median Shortage of Learning materials Index**: Cluster A < Cluster B

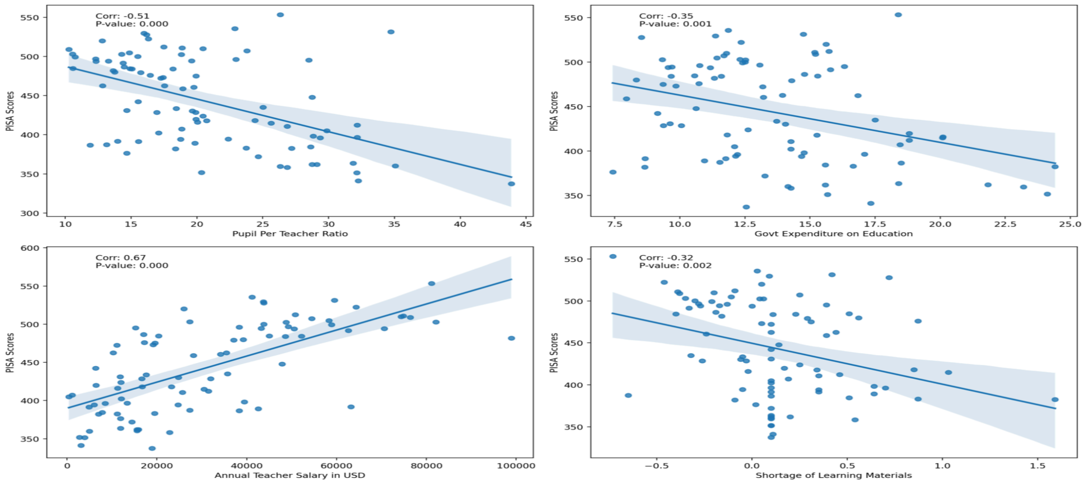


Cluster B	Albania	Algeria	Argentina	Azerbaijan	Bosnia and Herzegovina	Brazil	Brunei Darussalam	Bulgaria	Cambodia	Chile	...
Cluster A	Australia	Austria	Belarus	Belgium	Canada	Croatia	Czech Republic	Denmark	Estonia	Finland	...

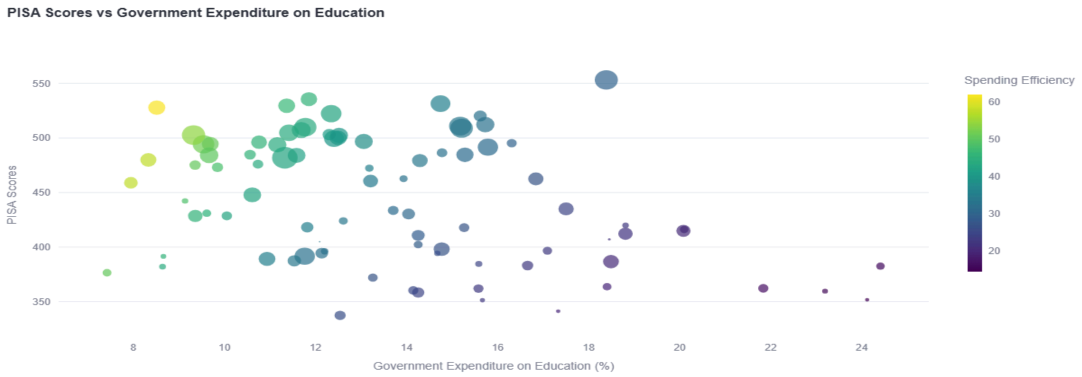
Correlation Analysis

- There is a strong positive correlation between ‘Teacher Salary’ and ‘Total Pisa Score’ (corr=0.61, p-value=0.0) suggesting that countries investing more in teacher salaries tend to have better student performance. This is so because such education systems are able to attract the best talents to take up careers in teachings, and effectively motivate them.

- ‘Pupil-per-teacher ratio’ also has a strong negative correlation with the ‘Total Pisa Score’ (corr=-0.51, p-value=0.0), implying that smaller class sizes might be associated with better learning outcomes. Compact learning environments enable instructors and teachers to effectively tailor teaching strategies to meet the needs of every student in the class space.



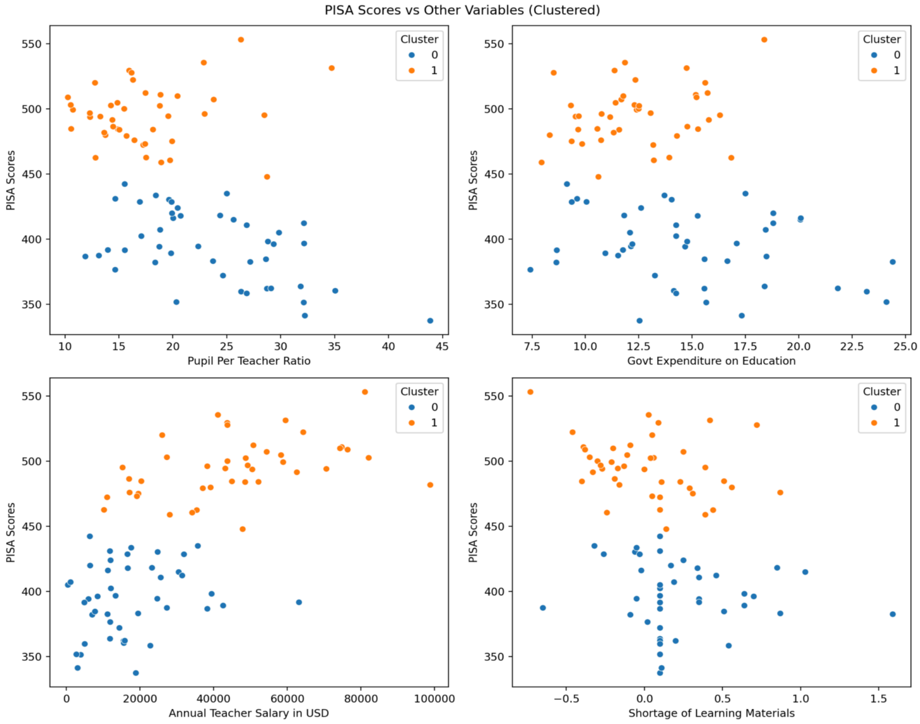
- Paradoxically, ’Government Expenditure on Education’ correlates slightly negatively (corr=-0.35, p-value=0.001) with Pisa Score’, indicating that ‘simply’ spending more on education does not guarantee better results. To explore this unusual relationship, we engineer a new feature called ‘Government Education Spending Efficiency’ which is the ratio between ‘PISA Scores’ and’ Govt Expenditure on Education’ and we deduce that countries with higher pisa scores despite lesser spendings on education are more efficient spenders than countries with lower pisa scores despite higher expenditure on education



Possible Explanations:

- **Other Factors**: Pisa scores may be influenced by other factors beyond government spending such as cultural attitudes towards education, teaching quality and other socio-economic factors.
- **Efficiency of Spending**: Countries who spend less may use their resources more efficiently.
- **Time Lag**: Current expenditure might not immediately reflect in test scores, as educational investments often take time to show results.
- **Quality vs Quantity**: The data doesn't account for how the money is spent.

- ‘Shortage of Learning Tools’ has a moderate negative but statistically significant correlation with PISA scores (corr=-0.32,p-value=0.002). This implies that countries having sufficient amounts of learning tools tend to have higher PISA scores. However, the relationship is not strong, and there are likely other factors influencing PISA scores that are not accounted for in this analysis.



EDA Summary

Top Performers: Singapore, China, Korea, Finland, and Japan.
Bottom Performers: Cambodia, Dominican Republic, Philippines, Uzbekistan, Kosovo.
PISA Scores:Bimodal distribution, suggesting two groups of countries with differing performance levels.
Pupil-Teacher Ratio: Generally, smaller class sizes are associated with better PISA scores.
Government Expenditure on Education: Higher spending doesn't directly translate to better scores; efficiency is key.
Teacher Salary: A strong positive correlation with PISA scores suggests that well-paid teachers contribute to better outcomes.
Shortage of Learning Materials: Countries with sufficient learning tools tend to perform better.

Section Two

Data Science AI Assistant with Retrieval-Augmented Generation (RAG)

We develop a data science AI assistant powered by RAG (Retrieval-Augmented Generation).

Functions

- We develop a RAG AI assistant to perform complex statistical analytics on our datasets.
- It will provide accurate insights from the data by leveraging on the statistical analysis of the data and its knowledge of the study background and other contextual information that we have not provided explicitly.

Capabilities

- We design a RAG-based AI Assistant that generates Python code and executes the code automatically on the local Python interpreter of the user to generate the code output.
- This allows us to create any complex interactive visualisation and any advanced Python code for data analysis by simply asking the system directly and clicking a button.
- The AI Assistant is capable of reasoning and responding to some questions beyond the scope of the study.
- It also retains memory of past conversations and uses that as context to respond to new questions.
- Every Chart and graph provided in Section One report was generated by the RAG AI Assistant.

LIMITATIONS

- The AI Assistant can only work reliably with our education dataset in PDF document format.
- It is unable to fetch raw data from some external databases

Key Components

Large Language Models (LLMs): We offer three Anthropic LLM models from the Claude family, Claude Sonnet, Claude Haiku, and Claude Opus) for text generation and task completion.

Embedding Models: We also offer two different embedding models, Google Generative AI embedding model and VoyageAI embedding model. They encode text from documents into numerical vector representations.

Document: The document for our RAG application is a PDF file containing structured tabular data on factors that affect education outcomes across countries.

Vector Database: We use FaceBook AI Similarity Search (FAISS) as our vector database for storage of embedding vectors and efficient retrieval.
RAG Framework: Langchain provides the tools that facilitate building the RAG system.

Python Dependencies

langchain==0.1.11
langchain-core==0.1.30
langchainhub==0.1.15
voyageai==0.2.1
pypdf==4.1.0
Streamlit==1.36.0

langchain-community==0.0.27
angchain-anthropic==0.1.4
anthropic==0.19.1
langchain-google-genai==1.0.1
faiss-cpu==1.8.0
Python==3.12.2

NB: Current Anthropic APIs work with recent versions of Langchain only in Google Vertex AI and Amazon Bedrock, both non-open source platforms.
To use Langchain with Anthropic APIs on VS Code and Jupyter NoteBook, we installed older Langchain versions and their dependencies.

Methods:

1. **Create and Set Up the Virtual Environment:**
    a. Create and activate a new virtual environment
    b. Create a requirements plain text file listing all necessary libraries
    c. install the required libraries listed in the requirements file.
    d. Import every necessary library, module and class

2. **Code Structure**
    a. **Initialise Embedding Models and LLMs**: Set the API Keys to the environment variable and initialise instances of the models

    b. Parse the Document and Create the Vector Retriever:
        i. Load PDF documents using PyPDF loader.
        ii. Split text into manageable chunks using Recursive Character Text Splitter.
        iii. Generate vector embeddings using the chosen embedding model.
        iv. Store embeddings in the FAISS vector database.
        v. Create a vector retriever to retrieve relevant document embeddings based on user queries.

    c. Create a History-Aware Retriever:
        i. We create a chain to combine contextual vector embeddings from FAISS, the chat history, the current user query, and the LLM, which in turn returns a history aware vector retriever.

ii. The history-aware vector retriever takes into account past chat history while retrieving vectors from the database. This enables the model to use past conversations (history awareness) as context when responding to new queries.

d. **Prompt Engineering:**

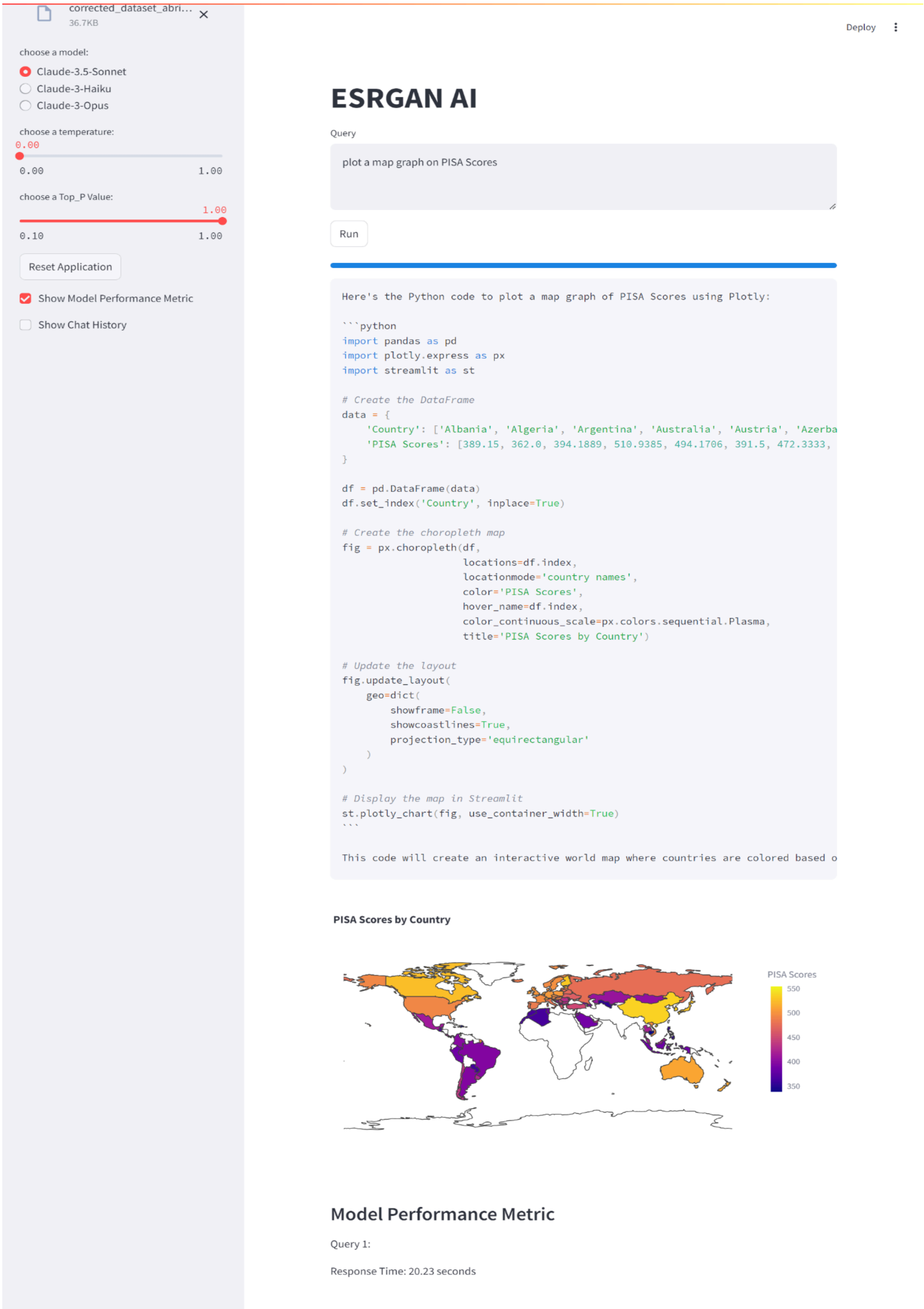
i. **Contextualize-question Prompt**: We design an instructional prompt to guide the LLM to leverage the chat history for context before re-writing the user query.

ii. **System Prompt**: This is the main prompt of the RAG system. We carefully engineer a role-playing and instructional prompt that transforms the LLM into an expert data scientist capable of solving a diverse range of problems on the reference document using several Python libraries. We also use in-context one-shot learning to guide the general nature of the LLMs responses and the format of its outputs.

e. **Create a Retrieval Chain**: We create a retrieval chain that receives inputs from the System Prompt and history-aware retriever, before transmitting to the LLM. The responses from the LLM are then passed on to the Runnable interface which gives the final RAG system output.

f. **Invoke the Runnable interface**: The runnable interface takes the user query, passes it on to the retrieval chain which retrieves vectors from FAISS before sending them along with the user query to the LLM to trigger a response. The response is in turn returned to the user via the runnable interface.

g. **Build the User Interface (UI) and Deploy**: We use Streamlit to create a user-friendly interface for interacting with the chatbot, and we deploy it to a local server.



choose a model:

● Claude-3.5-Sonnet
○ Claude-3-Haiku
○ Claude-3-Opus

choose a temperature: 0.00 (0.00 – 1.00)

choose a Top_P Value: 1.00 (0.10 – 1.00)

☑ Show Model Performance Metric

□ Show Chat History

ESRGAN AI

Query

plot a map graph on PISA Scores

~~~
Here's the Python code to plot a map graph of PISA Scores using Plotly:

```python
import pandas as pd
import plotly.express as px
import streamlit as st

# Create the DataFrame
data = {
    'Country': ['Albania', 'Algeria', 'Argentina', 'Australia', 'Austria', 'Azerba
    'PISA Scores': [389.15, 362.0, 394.1889, 510.9385, 494.1706, 391.5, 472.3333,
}

df = pd.DataFrame(data)
df.set_index('Country', inplace=True)

# Create the choropleth map
fig = px.choropleth(df,
                    locations=df.index,
                    locationmode='country names',
                    color='PISA Scores',
                    hover_name=df.index,
                    color_continuous_scale=px.colors.sequential.Plasma,
                    title='PISA Scores by Country')

# Update the layout
fig.update_layout(
    geo=dict(
        showframe=False,
        showcoastlines=True,
        projection_type='equirectangular'
    )
)

# Display the map in Streamlit
st.plotly_chart(fig, use_container_width=True)
```

This code will create an interactive world map where countries are colored based o
~~~

PISA Scores by Country

Model Performance Metric

Query 1:

Response Time: 20.23 seconds

This figure above is a simple demonstration of the RAG AI Assistant.

Section Three

Policy Recommendations

For Underperforming Countries (cluster B countries)

1. **Optimise Class Sizes**: We recommend reducing pupil-teacher ratios by propagating smaller class size, particularly in early education, to enhance individual attention.
2. **Invest Strategically in Teacher Development**:
   - We advise countries to offer competitive salaries and benefits to attract and retain qualified teachers.
   - We also recommend providing continuous professional training opportunities focused on effective teaching methods and student engagement.
3. **Ensure Access to Learning Resources**: Countries must address shortages in learning materials, including textbooks, technology, and laboratory equipment.
4. **Improve Spending Efficiency**:
   - Governments should evaluate existing education programs and prioritise those with proven effectiveness.
   - They should also be transparent and accountable in budget allocation and utilisation.
5. **Foster a Culture of Learning**: Both private and public sectors of nations must encourage parental involvement, community engagement, and societal value on education.

For High-Performing Countries (cluster A countries)

1. **Share Best Practices**: We recommend that higher performing nations collaborate with and mentor underperforming nations, sharing successful educational models and strategies.
2. **Promote Innovation**: We also advise that they continue investing in research and development to explore new pedagogical approaches and technologies that enhance learning.

Conclusion

- All countries must cooperate to facilitate knowledge exchange and collaboration on education policy and research.
- Every country should utilise robust data collection and analysis to inform policy interventions and track progress. We hereby encourage other countries like Nigeria to participate in PISA Test Assessments.