Education Infrastructure

•••

A Presentation by Team ESRGAN

Spring 24 Capstone Project

Unlocking Potential: Exploring Factors in Student Success

SECTION ONE

Introduction

This study investigates factors influencing educational outcomes.

We use PISA Scores as a metric and target variable.
The predictors include:

- Pupil-Teacher Ratio
- Annual Teacher Salary in USD (adjusted for ppp)
- Government Expenditure on Education (% of Total)
- Shortage of Learning Materials Index

-We collected data from and studied 86 countries.

-Our aim is to prescribe recommendations and policies to improve international education systems.

Methodology

Data Collection: The study datasets were downloaded from OECD (https://www.oecd.org/en/data.html) and World Bank(https://data.worldbank.org/)

Data Cleaning:

- Dropped duplicated data points
- Dropped Redundant and collinear variables
- Set every variable to the correct data format
- Handled missing values

Feature Engineering: We created two new features.

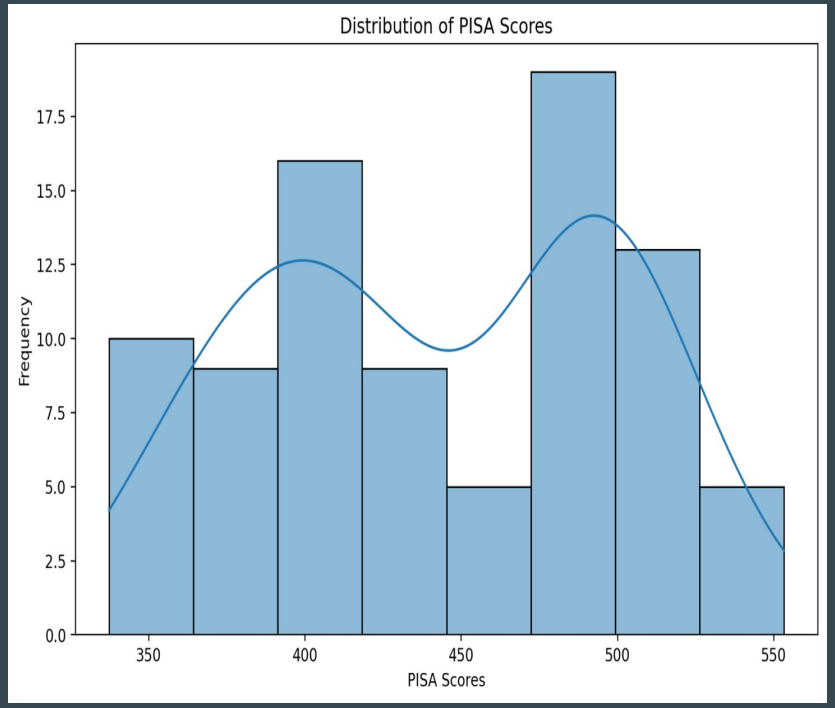
- Spending Efficiency: Ratio of PISA Scores to Govt expenditure on education
- Single PISA Score: mean of Math, Reading and Science PISA Scores

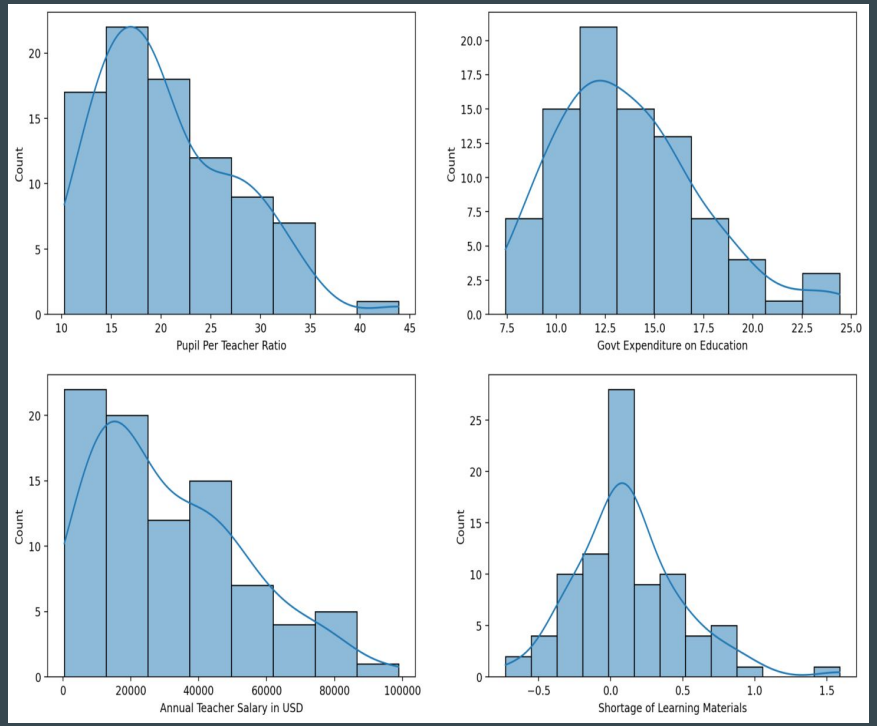
Exploratory Data Analysis

To uncover hidden insights and information from our data, we took the following analytical steps:

- Descriptive Analysis
- Distribution Analysis
- Unsupervised Statistical Learning
- Correlation Analysis

Distributions of Study Variables: Histogram





Unsupervised Statistical Learning

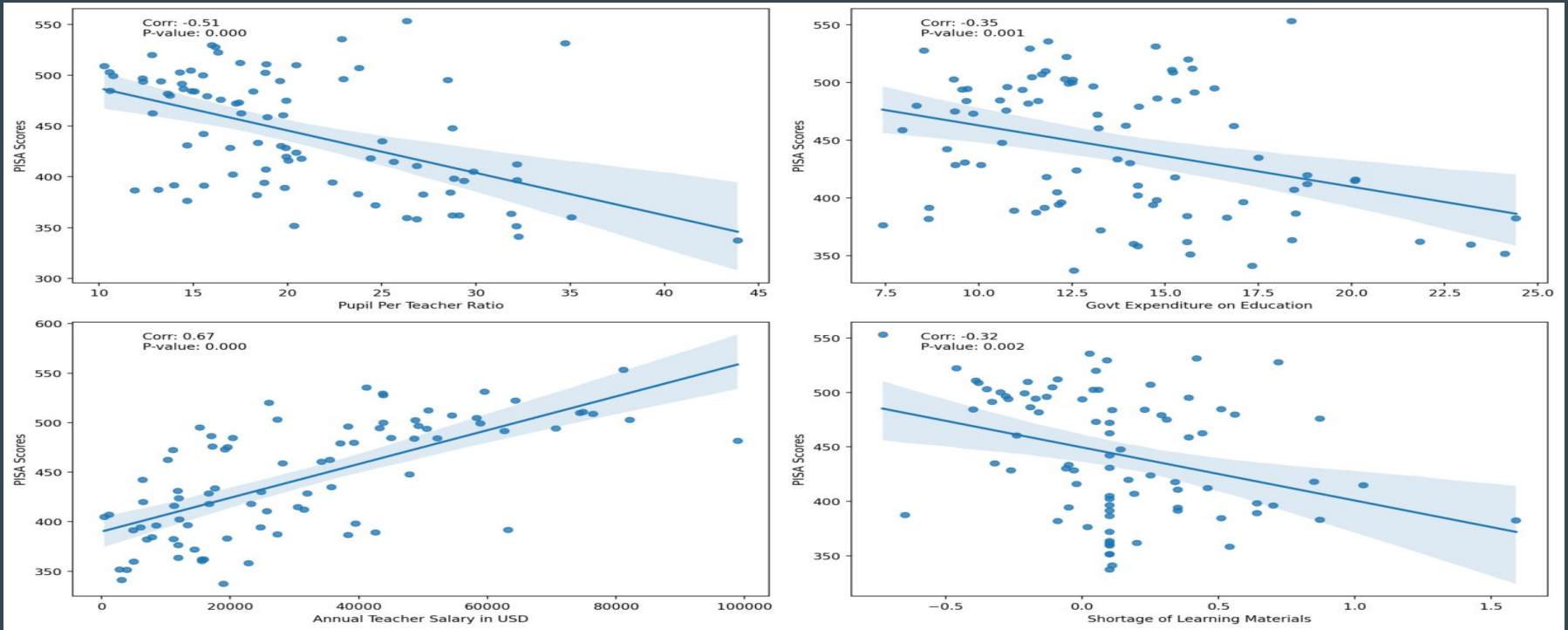
- ➔ To study the bimodal distribution of PISA Scores, we trained an unsupervised K-Means clustering machine learning algorithm.
- ➔ We set the number of centroids to match the number of peaks/modes in the histogram.
- ➔ Two clusters of countries were revealed.
- ➔ Cluster A have stronger education systems than Cluster B

Cluster A vs Cluster B

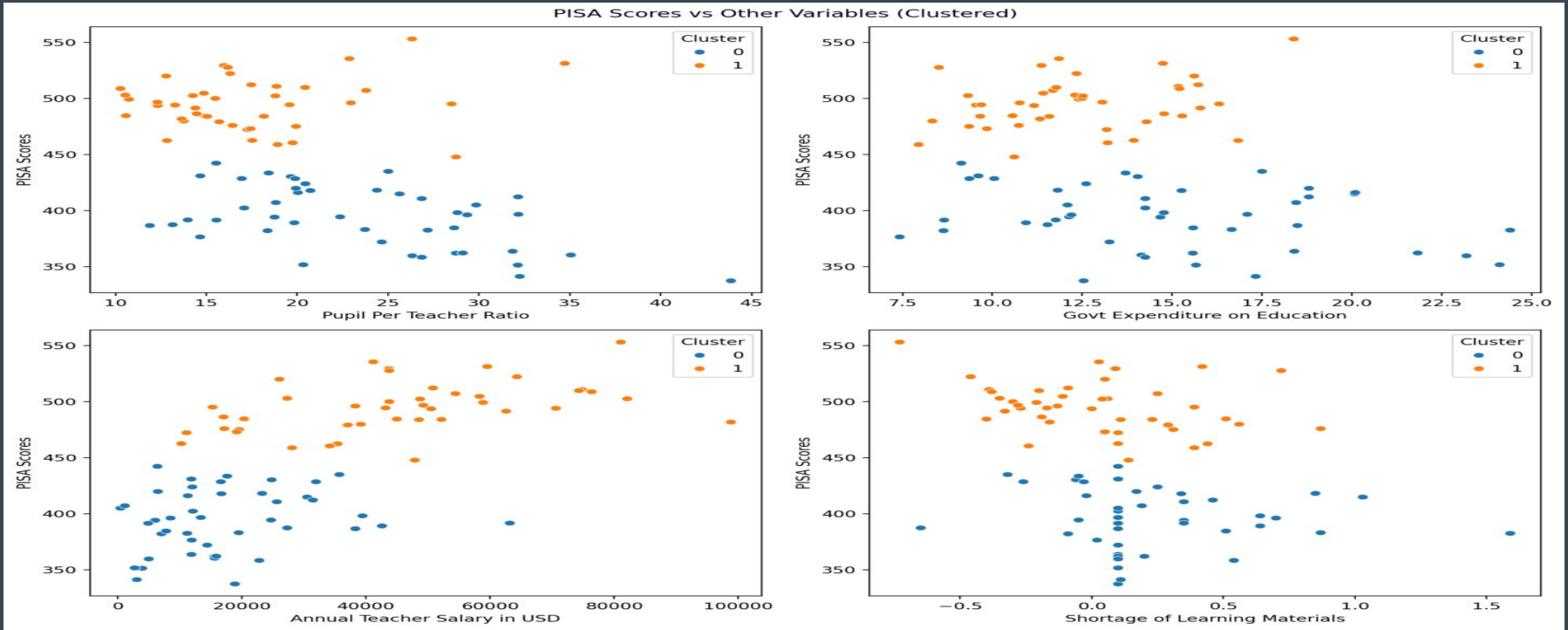
Group A	Australia	Austria	Belarus	Belgium	Canada	Croatia	Czech Republic	Denmark	Estonia	Finland	...	Slovenia	Spain	Sweden	Switzerland	Turkey
Group B	Albania	Algeria	Argentina	Azerbaijan	Bosnia and Herzegovina	Brazil	Brunei Darussalam	Bulgaria	Cambodia	Chile	...	Philippines	Qatar	Romania	Saudi Arabia	Thailand

Distribution of Study Variables: Boxplot

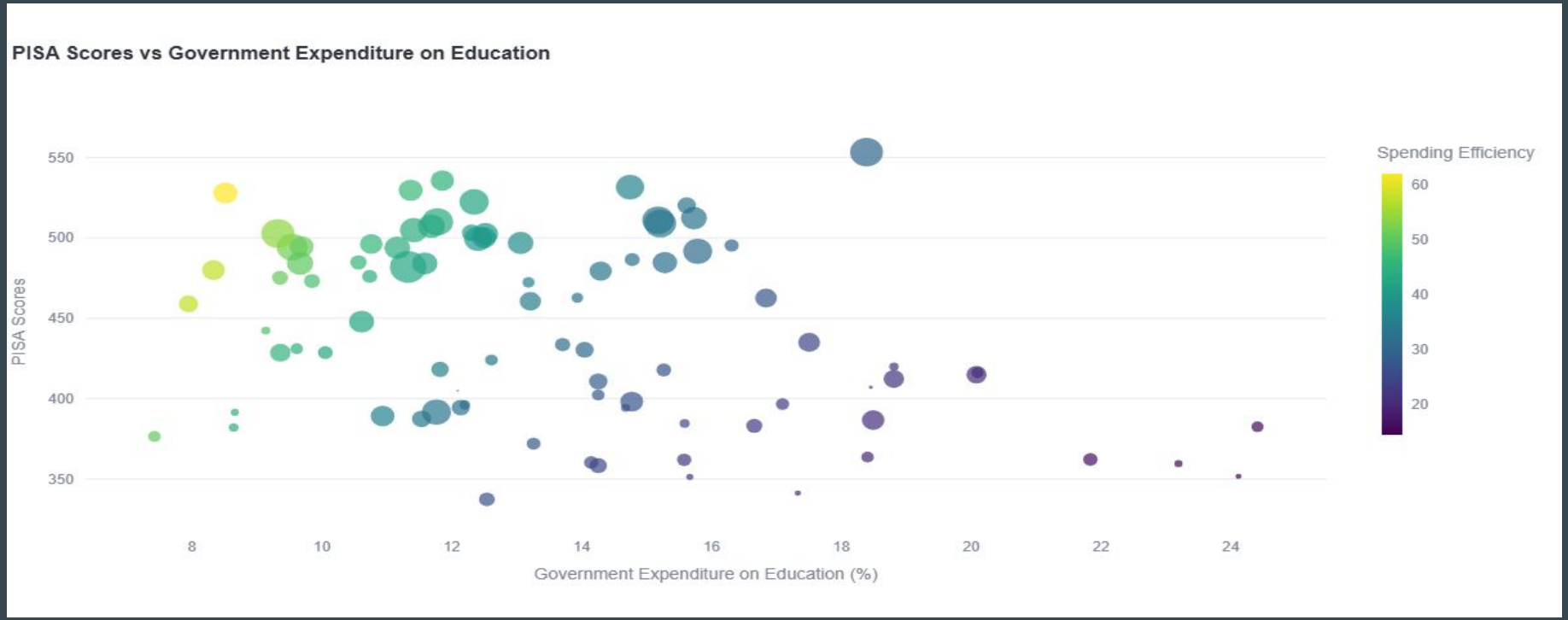
Correlation Analysis of Variables: Scatter Plots



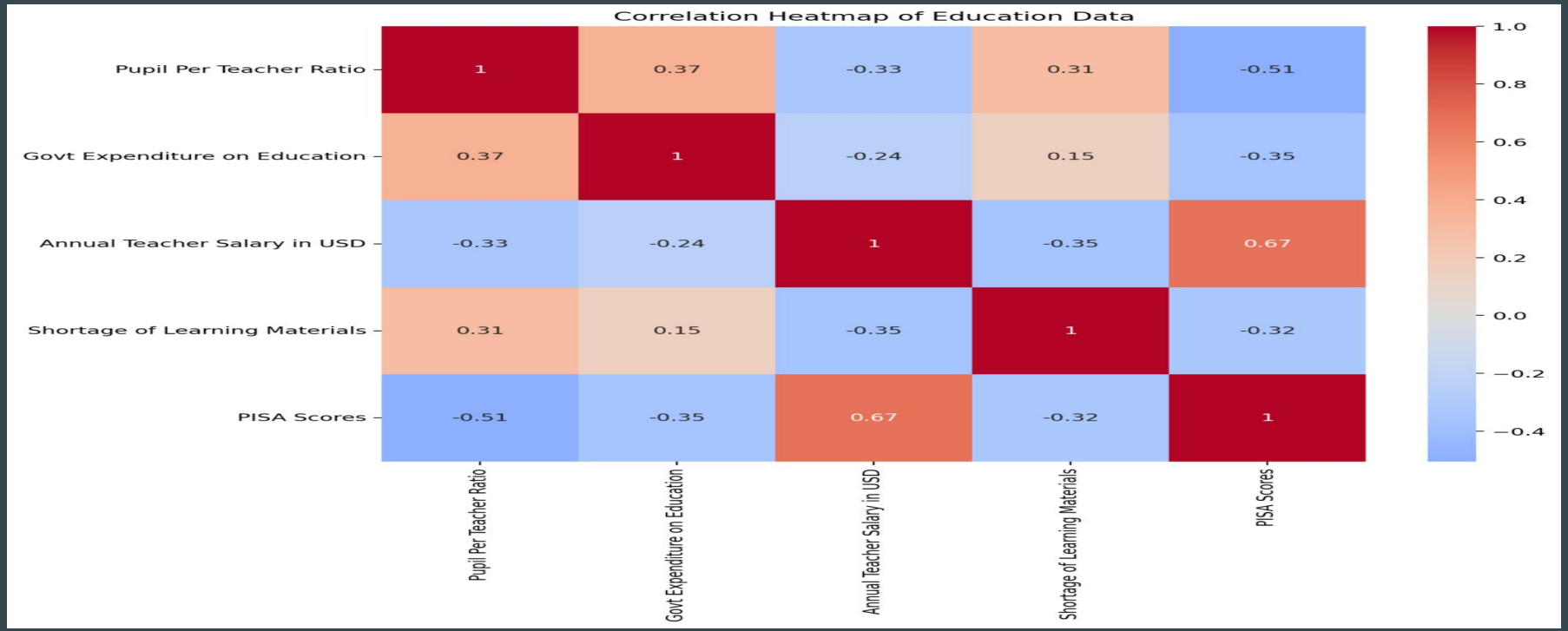
Correlation Analysis: Comparison of Clusters



Correlation analysis: Spending Efficiency



Correlation Analysis: Correlation Map



Key Findings

1. **Top performers** : Singapore, China, Korea, Finland, Japan

Bottom performers : Cambodia, Dominican Republic, Philippines, Uzbekistan, Kosovo

2. **PISA Scores** : Bimodal distribution (peaks at 395-405 and 495-505)

Higher cluster : Most European, North American, Australian, East Asian countries

Lower cluster : Most South American, West Asian, South East Asia, African countries

3. **Correlations** :
Teacher Salary and PISA Score : Strong positive (0.61)

Pupil-Teacher Ratio and PISA Score : Strong negative (-0.51)

Government Expenditure and PISA Score : negative (-0.35)

4. **K-Means Clustering:**
Cluster A (Higher performers): Lower pupil-teacher ratio, better learning materials, higher teacher salaries, lower government spending

Policy Recommendations

For Underperforming Countries:

1. Reduce class sizes
2. Invest in teacher development and competitive salaries
3. Ensure access to learning resources
4. Improve spending efficiency
5. Foster a culture of learning

For High-Performing Countries:

1. Share best practices
2. Address equity gaps
3. Promote innovation in education

Data Science AI Assistant

Section Two

Introduction

- ➔ Developed a data science AI assistant powered by RAG.
- ➔ Scope: Perform statistical analysis, generate Python code, create interactive visualizations, and provide insights.
- ➔ All charts and graphs in this report were generated by the RAG application.

Key Components

- ➔ LLMs: Claude Sonnet, Haiku, and Opus
- ➔ Embedding Models: Google Generative AI and VoyageAI
- ➔ Document: PDF with structured data on education outcomes
- ➔ Vector Database: Face AI Similarity Search (FAISS)
- ➔ RAG Framework: Langchain

Application Dependencies

- langchain==0.1.11
- langchain-community==0.0.27
- langchain-core==0.1.30
- langchain-anthropic==0.1.4
- langchainhub==0.1.15
- anthropic==0.19.1
- voyageai==0.2.1
- langchain-google-genai==1.0.1
- pypdf==4.1.0
- faiss-cpu==1.8.0
- streamlit==1.36.0
- python==3.12.2

Code Structure

- ➔ Environment Setup: Create virtual environment, install required packages
- ➔ Import libraries
- ➔ Initialize embedding models and LLMs
- ➔ Parse document and create vector retriever
- ➔ Design prompts (contextualize-query prompt and system prompt)
- ➔ Create history-aware retriever
- ➔ Create retrieval chain
- ➔ Send a query and invoke a response with Langchain Runnable interface
- ➔ Build UI with Streamlit and deploy on local server

Conclusion

- ➔ Stronger international cooperation amongst nations can help improve global education systems
- ➔ Every country should implement data-driven decision making

DEMO