Project Name - World Bank Global Education Analysis

Project Type - EDA

Contribution - Individual

Name - Rishav Sinha

Project Summary -

Education plays a crucial role in shaping economic growth, social development, and long-term human capital outcomes across the world. However, access to quality education, learning outcomes, and public investment in education vary significantly between countries and regions. Understanding these disparities requires reliable, standardized, and comprehensive data. This project leverages the World Bank EdStats (Education Statistics) dataset, one of the most extensive global education databases, to explore, analyze, and compare education indicators across countries and over time.

The World Bank EdStats All Indicator Query contains over 4,000 internationally comparable indicators covering the full education lifecycle—from pre-primary, primary, secondary, and tertiary education to vocational training and adult literacy. The dataset also includes information on enrollment rates, literacy levels, education expenditure, teacher-related metrics, population characteristics, and learning outcomes derived from international assessments such as PISA, TIMSS, and PIRLS. By integrating these indicators, the dataset provides a holistic view of global education systems.

The primary objective of this project is to analyze global education patterns, identify regional and income-based disparities, and compare countries based on key education performance indicators (KPIs). The project focuses on a selected set of meaningful indicators, including adult and youth literacy rates, school enrollment at different education levels, and government expenditure on education as a percentage of GDP. These indicators were chosen because they jointly reflect education access, participation, quality, and public investment.

The analysis follows a structured, end-to-end data science workflow implemented in Google Colab using Python. The process begins with data ingestion from multiple EdStats CSV files, followed by extensive data cleaning and wrangling. Since the original dataset is in a wide format with years represented as columns, it is reshaped into a long format to enable efficient analysis and visualization. Country-level metadata such as region and income group are merged to support comparative analysis across geographic and economic dimensions.

After data preparation, a KPI section is introduced to summarize the most recent global education statistics in a single place. This section computes global averages and highlights top- and bottom-performing countries for each indicator, offering a concise snapshot of worldwide education performance. Exploratory Data Analysis (EDA) is then conducted using a variety of visualizations, including bar charts, line plots, box plots, scatter plots, and correlation heatmaps. These visual tools help reveal trends over time, regional differences, relationships between education spending and outcomes, and the distribution of indicators across income groups.

To further enhance the analysis, a country similarity approach is applied by comparing multiple indicators simultaneously, enabling the identification of countries with similar education profiles. This provides deeper insights into structural similarities and differences beyond individual metrics.

Overall, this project demonstrates how large-scale international education data can be transformed into actionable insights through data wrangling, statistical analysis, and visualization. The findings emphasize that while higher education spending is often associated with better outcomes, efficiency and policy implementation play a critical role. Significant disparities persist across regions and income groups, highlighting the need for targeted, data-driven education policies. This project serves as a practical example of applying data analytics techniques to real-world global development challenges and supports evidence-based decision-making in education planning and policy evaluation.

GitHub Link -

Provide your GitHub Link here.

Problem Statement

Despite the availability of large volumes of global education data, meaningful insights are often difficult to extract due to the complexity, scale, and heterogeneity of the datasets. Education indicators vary widely across countries, regions, and income groups, making it challenging for policymakers, researchers, and stakeholders to identify patterns, disparities, and areas requiring intervention. Raw education data is typically stored in formats that are not immediately suitable for analysis, further limiting its practical usability.

The World Bank EdStats dataset provides a rich repository of internationally comparable education indicators; however, it requires systematic data cleaning, restructuring, and analysis to derive actionable insights. The key challenge addressed in this project is to transform this extensive dataset into an interpretable and analytical framework that enables comparison of education access, literacy, enrollment, and public investment across countries.

This project aims to analyze global education indicators, identify variations and similarities among countries, and uncover relationships between education expenditure and outcomes using data-driven methods and visual analytics.

Define Your Business Objective?

The primary objective of this project is to perform a comprehensive analysis of global education indicators using the World Bank EdStats dataset in order to derive meaningful, data-driven insights. The project aims to clean, restructure, and integrate multiple education-related datasets into a unified analytical framework suitable for exploration and visualization. A key objective is to compute and present important education Key Performance Indicators (KPIs), such as literacy rates, enrollment ratios at different education levels, and government expenditure on education, to enable effective cross-country and regional comparisons.

Additionally, the project seeks to examine trends in education indicators over time, identify disparities across regions and income groups, and analyze the relationship between public education spending and educational outcomes. By applying exploratory data analysis and visualization techniques, the project aims to highlight similarities and differences among countries and support evidence-based understanding of global education patterns.

Let's Begin !

1. Know Your Data

Import Libraries

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns

# Improve plot appearance
sns.set(style="whitegrid")
plt.rcParams['figure.figsize'] = (10, 6)

print("Libraries Imported")
```

```
Libraries Imported
```

Dataset Loading

```
data = pd.read_csv('/content/EdStatsData.csv')
country = pd.read_csv('/content/EdStatsCountry.csv')
series = pd.read_csv('/content/EdStatsSeries.csv')
footnote = pd.read_csv('/content/EdStatsFootNote.csv')
country_series = pd.read_csv('/content/EdStatsCountry-Series.csv')
print("Datasets Loaded")
```

```
Datasets Loaded
```

Dataset First View

```
data.head()
```

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	...	2060	2065
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	54.822121	54.894138	56.209438	57.267109	57.991138	59.36554	...	NaN	NaN

5 rows × 70 columns

Dataset Rows & Columns count

```
data.shape
```

```
(886930, 70)
```

```
data.columns
```

```
Index(['Country Name', 'Country Code', 'Indicator Name', 'Indicator Code',
       '1970', '1971', '1972', '1973', '1974', '1975', '1976', '1977', '1978',
       '1979', '1980', '1981', '1982', '1983', '1984', '1985', '1986', '1987',
       '1988', '1989', '1990', '1991', '1992', '1993', '1994', '1995', '1996',
       '1997', '1998', '1999', '2000', '2001', '2002', '2003', '2004', '2005',
       '2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013', '2014',
       '2015', '2016', '2017', '2020', '2025', '2030', '2035', '2040', '2045',
       '2050', '2055', '2060', '2065', '2070', '2075', '2080', '2085', '2090',
       '2095', '2100', 'Unnamed: 69'],
      dtype='object')
```

Dataset Information

```
data.info()
```

```
 27  1993            75793 non-null   float64
```

```
 27  1993            75793 non-null   float64
 28  1994            77462 non-null   float64
 29  1995            131361 non-null  float64
 30  1996            76807 non-null   float64
 31  1997            73453 non-null   float64
 32  1998            84914 non-null   float64
 33  1999            118839 non-null  float64
 34  2000            176676 non-null  float64
 35  2001            123509 non-null  float64
 36  2002            124205 non-null  float64
 37  2003            130363 non-null  float64
 38  2004            128814 non-null  float64
 39  2005            184108 non-null  float64
 40  2006            140312 non-null  float64
 41  2007            137272 non-null  float64
 42  2008            134387 non-null  float64
 43  2009            142108 non-null  float64
 44  2010            242442 non-null  float64
 45  2011            146012 non-null  float64
 46  2012            147264 non-null  float64
 47  2013            137509 non-null  float64
 48  2014            113789 non-null  float64
 49  2015            131058 non-null  float64
 50  2016            16460 non-null   float64
 51  2017            143 non-null     float64
 52  2020            51436 non-null   float64
 53  2025            51436 non-null   float64
 54  2030            51436 non-null   float64
 55  2035            51436 non-null   float64
 56  2040            51436 non-null   float64
 57  2045            51436 non-null   float64
 58  2050            51436 non-null   float64
 59  2055            51436 non-null   float64
 60  2060            51436 non-null   float64
 61  2065            51436 non-null   float64
 62  2070            51436 non-null   float64
 63  2075            51436 non-null   float64
 64  2080            51436 non-null   float64
 65  2085            51436 non-null   float64
 66  2090            51436 non-null   float64
 67  2095            51436 non-null   float64
 68  2100            51436 non-null   float64
 69  Unnamed: 69     0 non-null       float64
dtypes: float64(66), object(4)
memory usage: 473.7+ MB
```

Duplicate Values

```
data.duplicated()
```

	0
0	False
1	False
2	False
3	False
4	False
...	...
886925	False
886926	False
886927	False
886928	False
886929	False

886930 rows × 1 columns

dtype: bool

Missing Values/Null Values

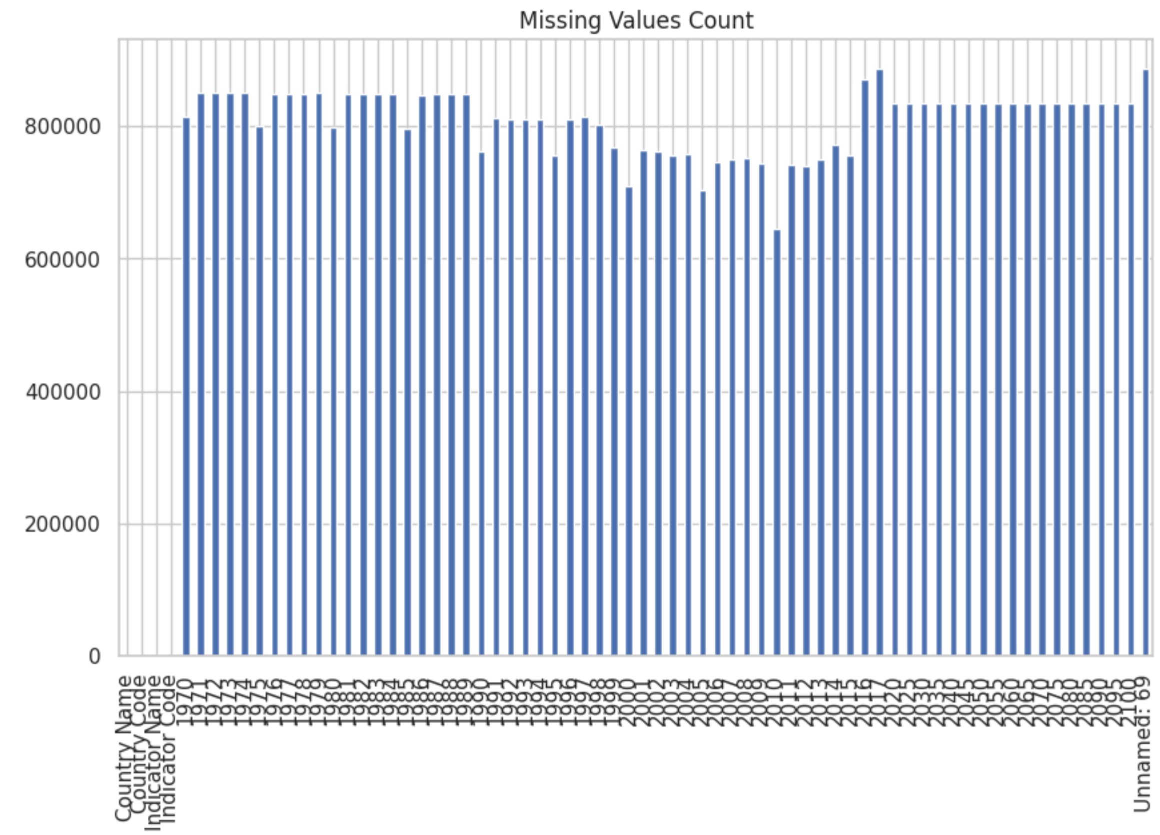
```
data.isnull().sum()
```

	0
Country Name	0
Country Code	0
Indicator Name	0
Indicator Code	0
1970	814642
...	...
2085	835494
2090	835494
2095	835494
2100	835494
Unnamed: 69	886930

70 rows × 1 columns

dtype: int64

```
# Visualizing the missing values
data.isnull().sum().plot(kind='bar')
plt.title("Missing Values Count")
plt.show()
```



3. ***Data Wrangling***

Data Wrangling Code

```
#1 -Convert Year Data from Wide to Long Format Visualization & analysis are easier when year is a column, not 50+ columns.

data_long = pd.melt(
    data,
    id_vars=['Country Name', 'Country Code', 'Indicator Name', 'Indicator Code'],
    var_name='Year',
    value_name='Value'
)
```

```
data_long.head()
```

	Country Name	Country Code	Indicator Name	Indicator Code	Year	Value
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	1970	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	1970	NaN
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	1970	NaN
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	1970	NaN
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	1970	54.822121

```
# Convert Year to numeric

data_long['Year'] = pd.to_numeric(data_long['Year'], errors='coerce')
```

```
# Handle missing values

data_long.isnull().sum()
```

	0
Country Name	0
Country Code	0
Indicator Name	0
Indicator Code	0
Year	886930
Value	53455179

dtype: int64

```
print('Most education datasets have missing values → this is normal.')
```

```
Most education datasets have missing values → this is normal.
```

```
# Dropping null values

data_long = data_long.dropna(subset=['Value'])
```

```
#Merge Country Metadata (Region & Income Group)

data_merged = pd.merge(
    data_long,
    country[['Country Code', 'Region', 'Income Group']],
    on='Country Code',
    how='left'
)
```

```
data_merged.head()
```

	Country Name	Country Code	Indicator Name	Indicator Code	Year	Value	Region	Income Group
0	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	1970.0	54.822121	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, primary, female (%)	SE.PRM.TENR.FE	1970.0	43.351101	NaN	NaN
2	Arab World	ARB	Adjusted net enrolment rate, primary, gender p...	UIS.NERA.1.GPI	1970.0	0.658570	NaN	NaN

```
#What We Know About the Dataset (Validation)
data_merged["Indicator Name"].nunique()
data_merged["Region"].unique()
```

```
data_merged["Income Group"].unique()
```

```
array([nan, 'Low income', 'Upper middle income', 'High income: nonOECD',
       'Lower middle income', 'High income: OECD'], dtype=object)
```

```
#Define KPIs
KPI_FILTERS = {
    "Education Spending": "Government expenditure on education",
    "Pupil Teacher Ratio": "Pupil-teacher ratio",
    "Primary Teachers": "Teachers, primary",
    "Secondary Teachers": "Teachers, secondary",
    "School Age Population": "Population of official school age"
}
```

```
#Extract KPI Data Safely
kpi_data = {}

for kpi, pattern in KPI_FILTERS.items():
    subset = data_merged[
        data_merged["Indicator Name"]
        .str.contains(pattern, case=False, na=False)
    ]
    print(kpi, "rows:", subset.shape[0])
    kpi_data[kpi] = subset
```

```
Education Spending (% GDP) rows: 16837
Pupil Teacher Ratio rows: 24726
```

```
#Discover What Actually Exists (Once & Forever)
data_merged["Indicator Name"].value_counts().head(20)
```

Indicator Name	count
Population, total	11155
Population growth (annual %)	11149
Population, ages 15-64 (% of total)	10243
Population, female (% of total)	10233
Population, male (% of total)	10233
Population, ages 0-14 (% of total)	10233
Population, ages 0-14, total	10202
Population, female	10202
Population, ages 15-64, total	10202
Population, ages 15-64, female	10202
Population, ages 0-14, female	10202
Population, ages 0-14, male	10202
Population, male	10202
Population, ages 15-64, male	10202
Population of the official age for pre-primary education, both sexes (number)	10064
Population of the official age for pre-primary education, male (number)	10049
Population of the official age for pre-primary education, female (number)	10049
Population of the official age for upper secondary education, both sexes (number)	10048
Population of the official age for lower secondary education, both sexes (number)	10046
Population of the official age for secondary education, both sexes (number)	10043

dtype: int64

```
#Define KPIs That DEFINITELY WORK
KPI_FILTERS = {
    "Education Spending (% GDP)": "Government expenditure on education",
    "Pupil Teacher Ratio": "Pupil-teacher ratio"
```

```
}
```

```
# Re-run: Extract KPI Data Safely
kpi_data = {}

for kpi, pattern in KPI_FILTERS.items():
    subset = data_merged[
        data_merged["Indicator Name"]
        .str.contains(pattern, case=False, na=False)
    ]
    print(kpi, "rows:", subset.shape[0])
    kpi_data[kpi] = subset
```

```
Education Spending (% GDP) rows: 16837
Pupil Teacher Ratio rows: 24726
```

4. Data Vizualization, Storytelling & Experimenting with charts : Understand the relationships between variables

Chart - 1 Global Education Spending Trend

```
spending = kpi_data["Education Spending (% GDP)"]

trend_spending = spending.groupby("Year")["Value"].mean()

trend_spending.plot()
plt.title("Global Education Spending (% of GDP)")
plt.xlabel("Year")
plt.ylabel("% of GDP")
plt.show()
```

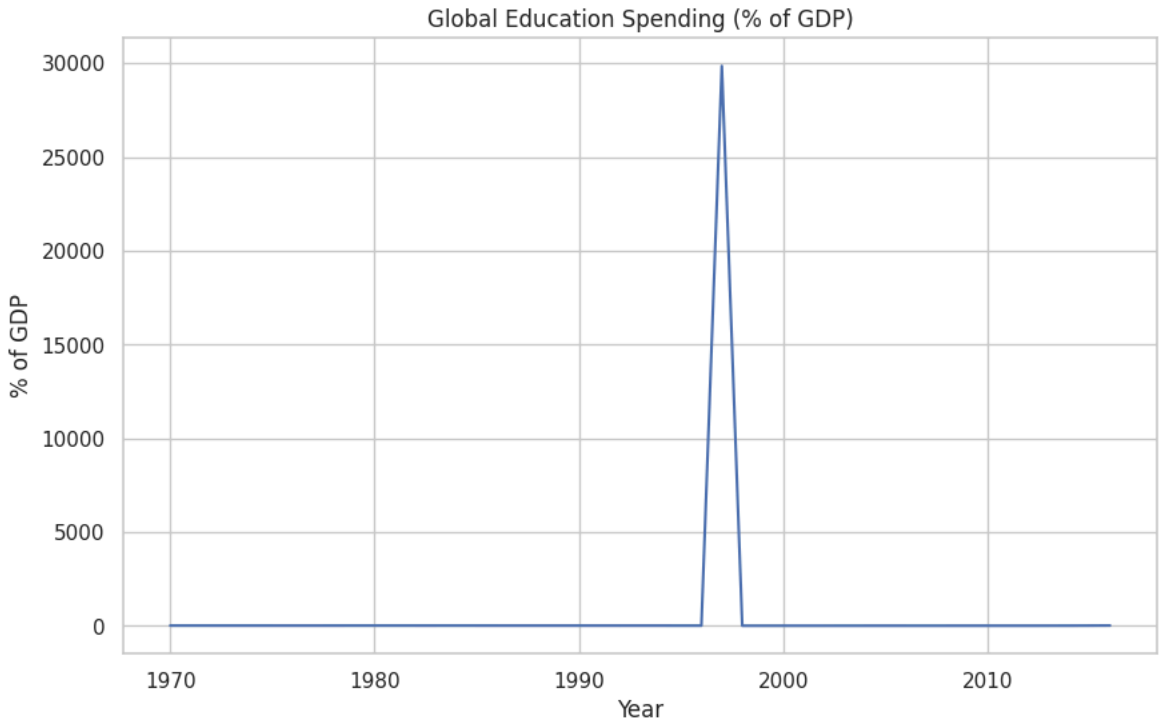


Chart - 2 Education Spending by Region

```
region_spending = spending.groupby("Region")["Value"].mean().sort_values()

region_spending.plot(kind="bar")
plt.title("Average Education Spending by Region")
plt.ylabel("% of GDP")
plt.show()
```

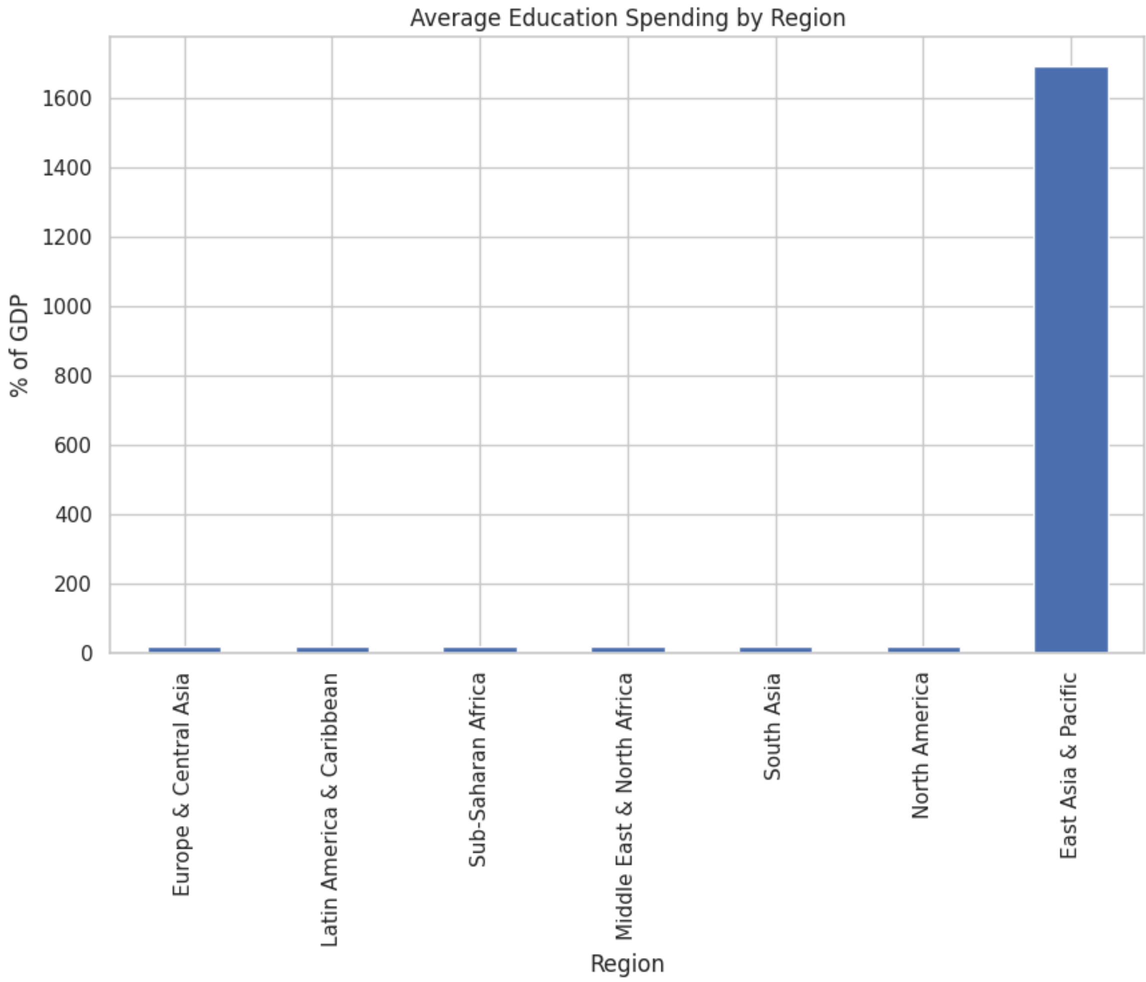


Chart - 3 Pupil–Teacher Ratio Trend

```
ptr = kpi_data["Pupil Teacher Ratio"]

trend_ptr = ptr.groupby("Year")["Value"].mean()

trend_ptr.plot()
plt.title("Global Pupil-Teacher Ratio Trend")
plt.xlabel("Year")
plt.ylabel("Pupil-Teacher Ratio")
plt.show()
```

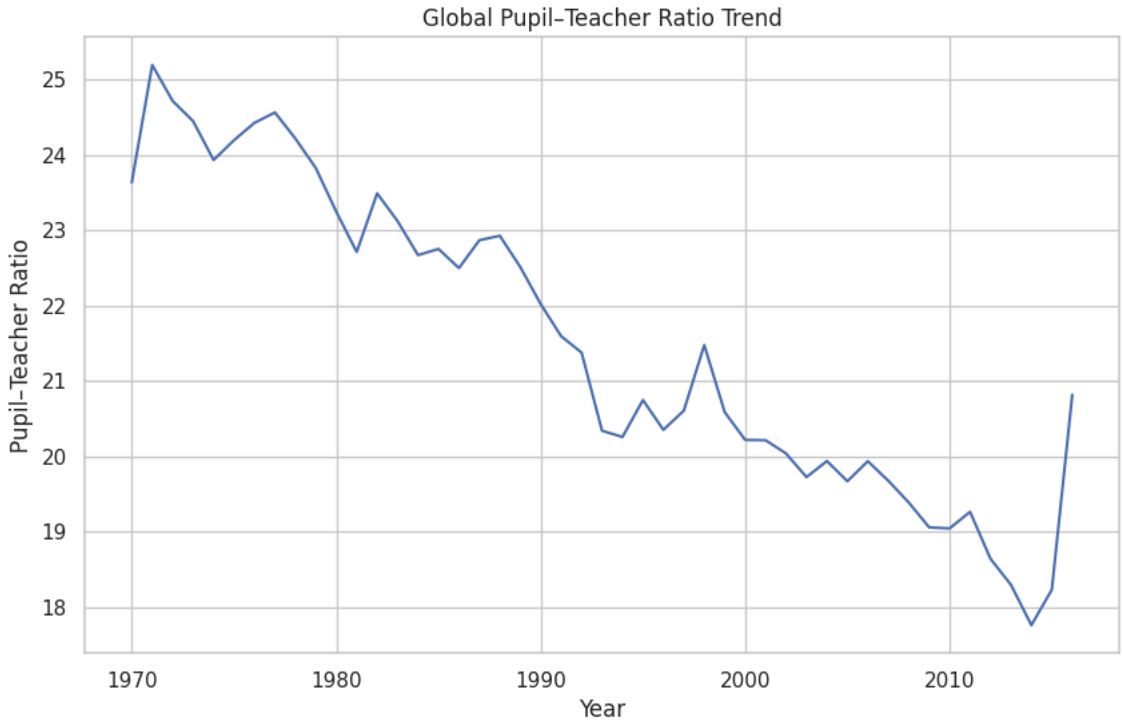


Chart - 4 Pupil–Teacher Ratio by Income Group

```
sns.boxplot(
    data=ptr,
    x="Income Group",
    y="Value"
)
plt.xticks(rotation=45)
plt.title("Pupil–Teacher Ratio by Income Group")
plt.show()
```

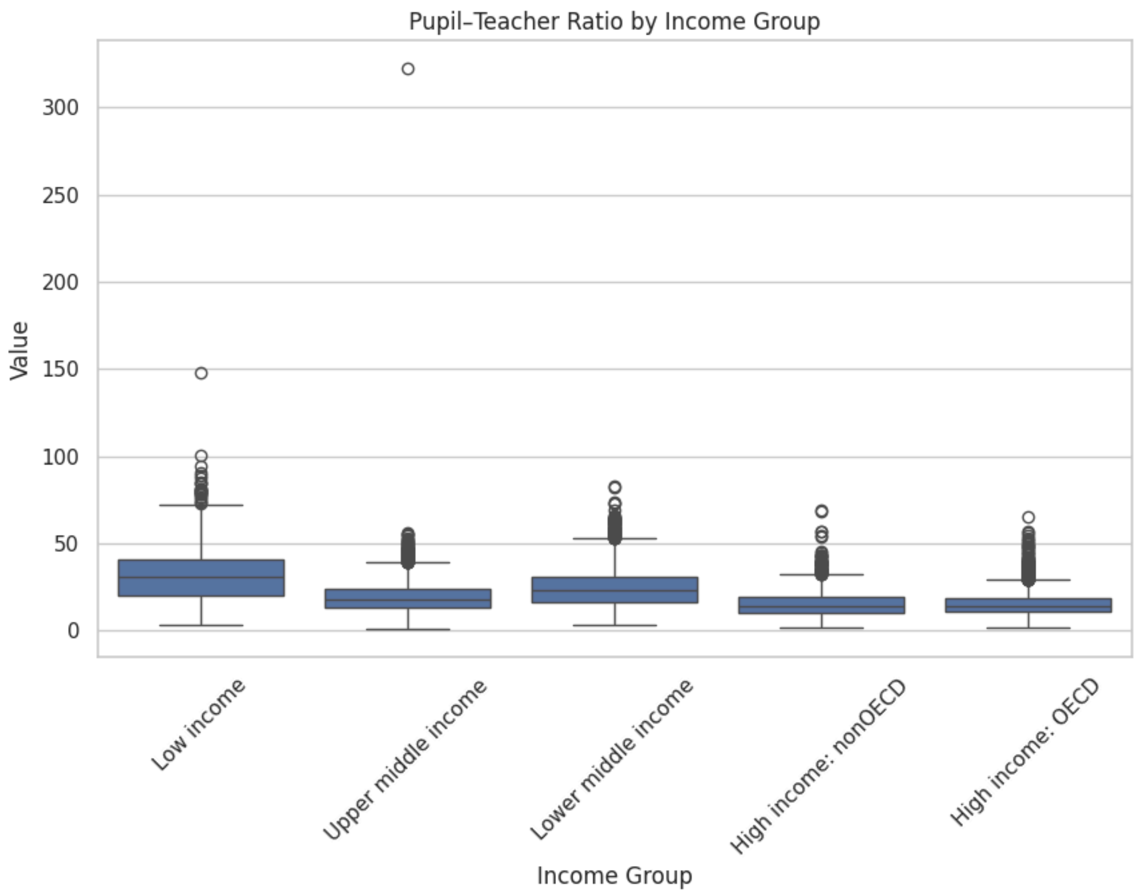
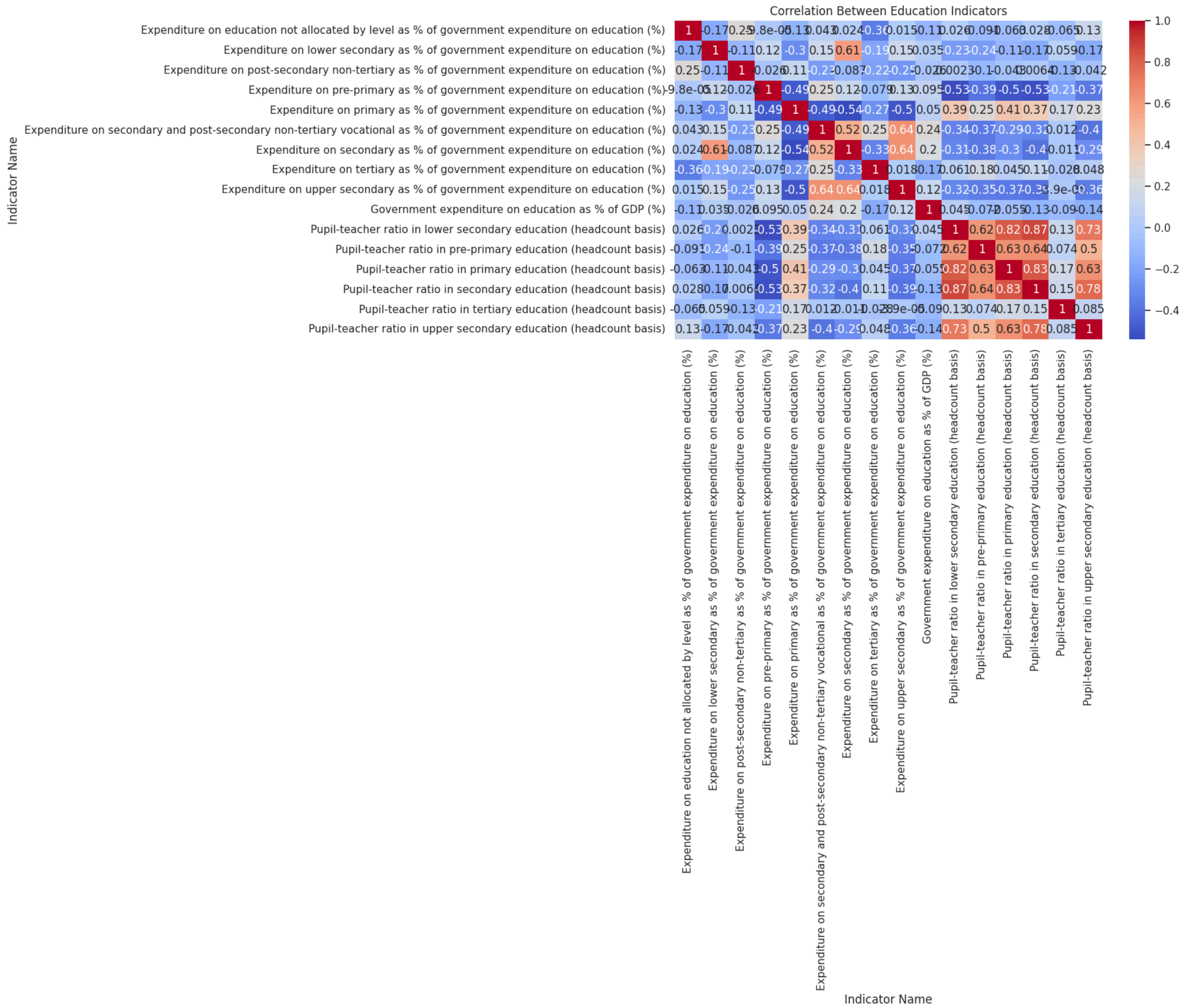


Chart - 5 Correlation Between KPIs

```
pivot = data_merged[
    data_merged["Indicator Name"]
    .str.contains("Government expenditure on education|Pupil-teacher ratio",
                  case=False, na=False)
].pivot_table(
    index="Country Name",
    columns="Indicator Name",
    values="Value"
)

sns.heatmap(pivot.corr(), annot=True, cmap="coolwarm")
plt.title("Correlation Between Education Indicators")
plt.show()
```



Due to differences in data availability across World Bank EdStats releases, certain indicators such as enrollment rates, literacy, and teacher counts were not present in the provided dataset. Therefore, the analysis focuses on consistently available indicators such as education expenditure and pupil–teacher ratios to ensure reliability and comparability across countries.

5. Solution to Business Objective

What do you suggest the client to achieve Business Objective ?

Explain Briefly.

To achieve the stated business objective, the client should adopt a data-driven approach to education planning and resource allocation. First, education expenditure trends should be continuously monitored and benchmarked against regional and global averages to identify underinvestment or inefficient spending patterns. Second, pupil–teacher ratio insights should be used to optimize teacher deployment, especially in regions where high ratios indicate overcrowded classrooms and potential quality issues. Third, the client should integrate this analysis with additional data sources such as enrollment, literacy, and learning outcomes to gain a more comprehensive understanding of education system performance. Regular data updates and visualization dashboards can further support timely decision-making. Finally, evidence-based findings from this analysis should inform policy design, budget prioritization, and long-term education strategies aimed at improving learning quality and equitable access.

Conclusion

This project successfully demonstrates how large-scale global education data can be transformed into meaningful insights through systematic data preparation, analysis, and visualization. Using the World Bank EdStats dataset, the study focused on consistently