

## Data Collection & Preparation **SIS Project Assignment**

**Group members:** Smetov Damir, Marat Rakhat

**Area of interest:** Relationship between education and economy

**Overall objective:** Collect data about countries from open sources, clean and merge them, and analyze key relationships between education and economy.

1. Introduction We chose this topic because we wanted to study whether education affects the economy. It is often said that countries investing more in education achieve higher economic results. To verify this, we analyzed the connection between GDP per capita and several education indicators such as education spending and the number of universities.

2. Sources of Data Two main sources were used:

**World Bank API:** provided structured information about GDP per capita, education expenditure (% of GDP), and population for different countries.

**Wikipedia:** contained lists of universities by country. Using Python and BeautifulSoup, we scraped data to count the number of universities in each country.

Combining these two data sources allowed us to create one dataset that included both economic and educational indicators.

3. Stages of Work The project was completed in several steps:

- **Data collection:** downloaded economic data from the World Bank API and scraped university lists from Wikipedia.
- **Data cleaning:** removed duplicates, handled missing values, and unified country codes (ISO-3 format).
- **Integration:** merged both datasets into a single table using the iso3 country code.
- **Processing:** added a new calculated metric — universities per million people — to compare countries of different sizes.

4. Analysis and Visualization After preparing the merged dataset, we analyzed the relationships between economic and educational indicators using different types of visualizations:

- **Bar Chart:** showed GDP per capita by country (e.g., USA, China, Japan, Germany, Kazakhstan). Developed countries like the USA, Canada, and Germany had the highest GDP per capita.
- **Heatmap:** displayed correlations between variables. A positive correlation was found between education expenditure and GDP per capita, as well as between the number of universities and tertiary enrollment.
- **Scatter Plot:** illustrated the relationship between the number of universities and GDP per capita — countries with more universities tended to have higher GDP.
- **Pie Chart:** showed the share of top 100 universities by country, where the United States had the largest share, followed by the UK, Japan, and Germany.

These visualizations confirmed that economic prosperity and educational development are connected.

5. Conclusion The analysis confirmed our hypothesis: education is positively related to economic growth. Countries that invest more in education and have more universities generally show higher GDP per capita.

However, the impact depends on many factors — efficiency of spending, quality of education, and population size. This project demonstrated how open data and Python tools can be used to explore real-world relationships between education and economy.