B report

In this part, our main focus is to obtain and process data that is typically difficult to extract. Ultimately, we extract several charts and tables(in jpg, png, and other image formats) from literature, and then convert them into DataFrames using methods such as WebPlotDigitizer and PaddleOCR. Subsequently, we analyze and process the data.

Architect:

1. **Obtaining Web Content**

In this stage, we implement the functionality to extract chart elements from web pages using Python. Firstly, we fetch the web content by sending an HTTP request, then parse the HTML document using the BeautifulSoup library, ultimately locating the elements containing the charts.

1. **Extracting Charts**

Repeat the steps of the first stage to extract charts from multiple documents.

1. **OCR (Optical Character Recognition)**

As the extracted charts mainly consist of chart types (line, bar, pie) and tables, we introduce two OCR methods respectively aimed at charts and tables using the indicators "total\_government\_expenditure\_on\_education\_gdp" and "work\_hour" as examples.

* 1. **Extracting chart data using the WebPlotDigitizer method**

1. Import Image: Import the image containing the desired data into WebPlotDigitizer. File formats can include jpg, jpeg, and others.
2. Calibrate Axes: Mark the starting and ending points of the axes on the imported image. This step informs WebPlotDigitizer how the pixels in the image map to actual data coordinates.
3. Extract Data Points: Using the point selection tool in the toolbar, WebPlotDigitizer converts each position into the corresponding data value based on the calibrated axes. Any number of data points can be extracted as needed.

d. Export Data: Finally, export the extracted data points to common data formats such as CSV or TXT for further analysis. This allows us to convert the line chart of "total\_government\_expenditure\_on\_education\_gdp" into CSV format, enabling further processing and analysis in Python using pandas.

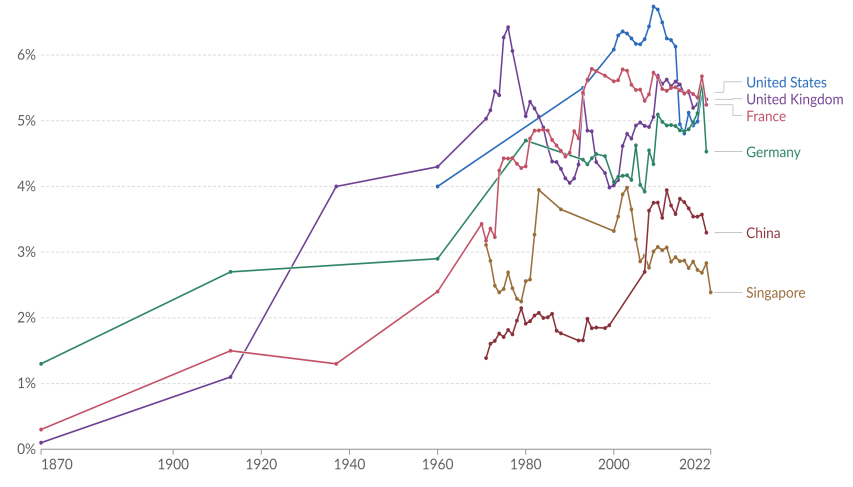


Figure1: total\_government\_expenditure\_on\_education\_gdp (line chart)

**3.2 Extracting table data using the PaddleOCR method**

a. Using work hour as an example, we first utilize the PPStructure module in PaddleOCR to extract text information from the images containing work hour. We load two image files, namely 2.1.1.png and 2.1.2.png, and perform text extraction operations on them. The extraction results are saved in the specified path.

b. Next, we merge the two text extraction results into one dataset. We use the Pandas library to read two Excel files and use the concat function to merge them into one DataFrame. The merged result is then saved in a new Excel file.



Figure2: Work hour (table)

1. **Data Processing**  
   After extracting data for each indicator using the above method, we merge the data for the 14 indicators obtained in the end into one combined dataset. Then, we can perform some basic data analysis on the dataset.

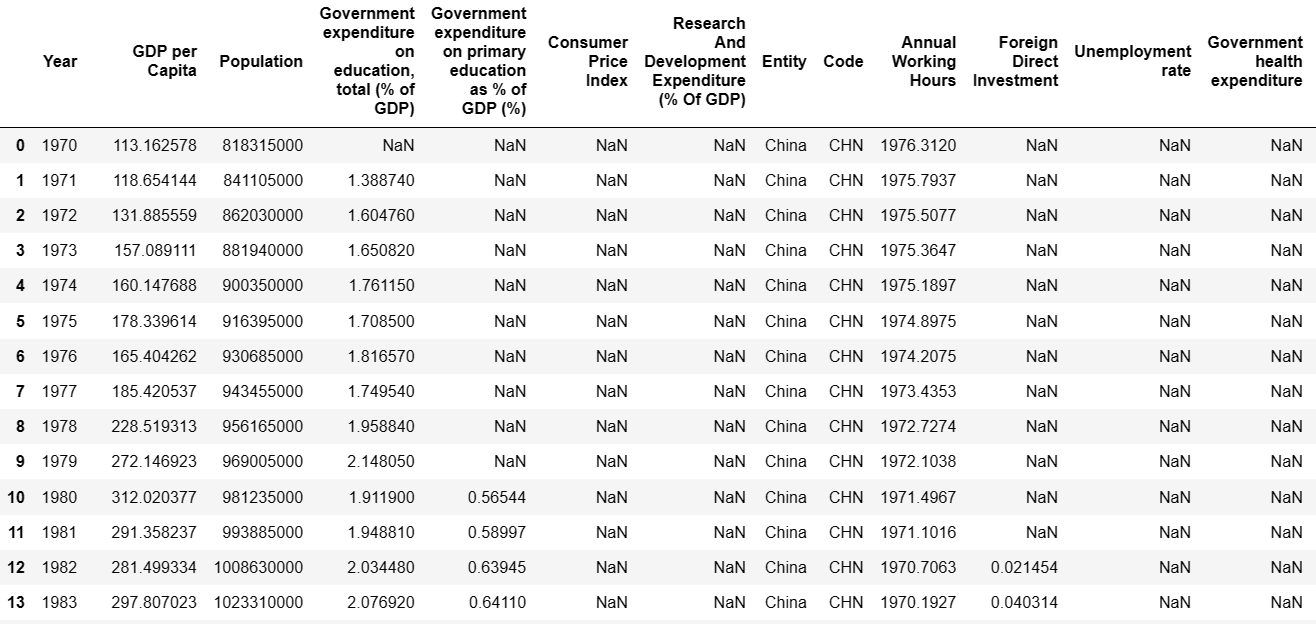


Figure3: Total Data

4.1 Data Cleaning

Data cleaning is a crucial step in data analysis, involving handling missing values, duplicates, outliers, and so on. When dealing with missing values, we use the .isnull() method to identify them. We observed that the missing values in this case are due to the unavailability of data for certain years, rather than being caused by human error or other unknown factors. Therefore, to ensure accuracy in subsequent processing, no action is taken regarding these missing values. We use the .duplicated() method to identify duplicate values. Statistical methods or visualization techniques such as .describe() can be employed to identify outliers.

4.2 Data Transformation

The numerical ranges of different features may vary significantly, potentially causing issues related to scale in subsequent processing. Therefore, we perform standardization to scale the feature values between 0 and 1.

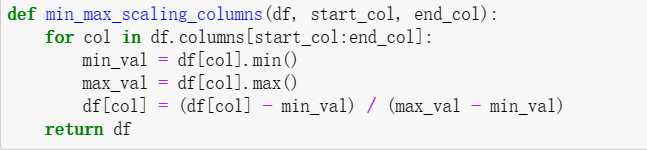


Figure4: Standardization

Then, we conducted some visualizations and correlation analysis. It can be observed that more government spending on education doesn't necessarily correlate with higher GDP. The GDP tends to peak when the expenditure accounts for around 3.3% of GDP, followed by a relatively high GDP when the expenditure is between 3-4% of GDP. Regarding the Consumer Price Index (CPI), it shows a positive correlation with GDP. Generally, a higher CPI corresponds to a higher GDP.

In terms of the correlation heatmap analysis, most features exhibit a positive correlation with each other. However, there are exceptions. The Age dependency ratio shows a negative correlation with other features, and the Research And Development Expenditure correlates negatively with Government expenditure on primary education.

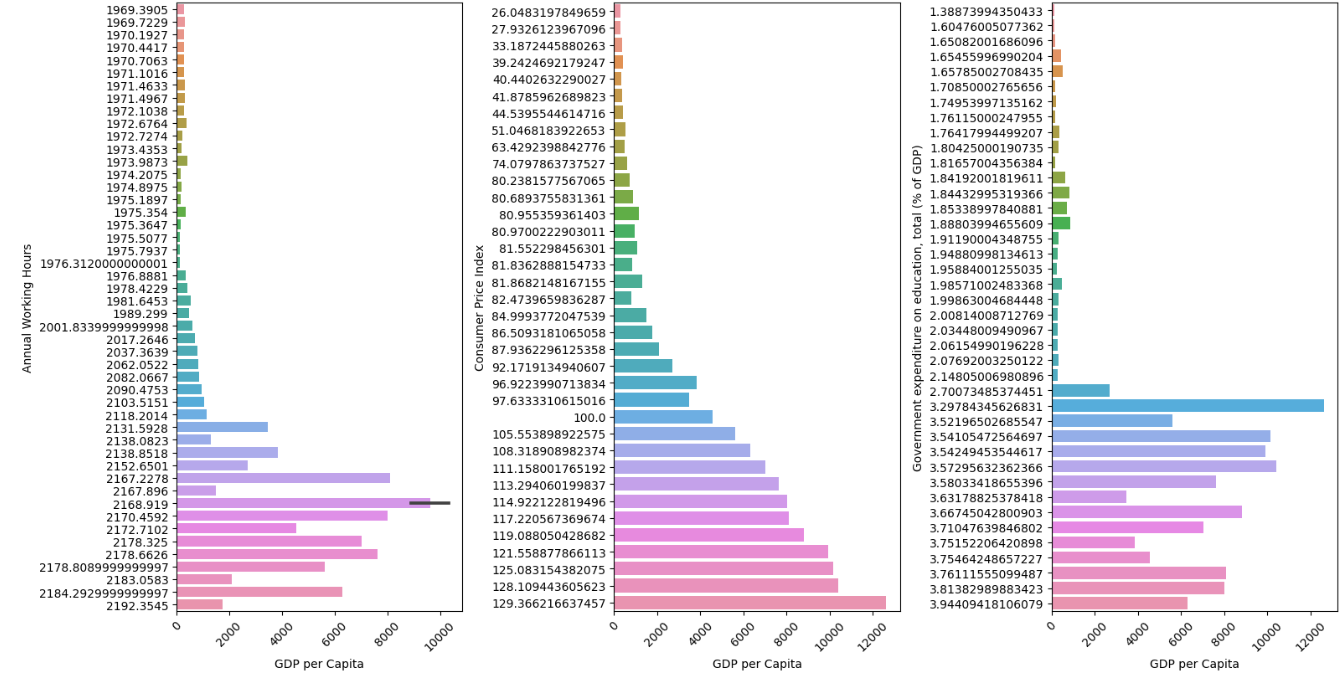


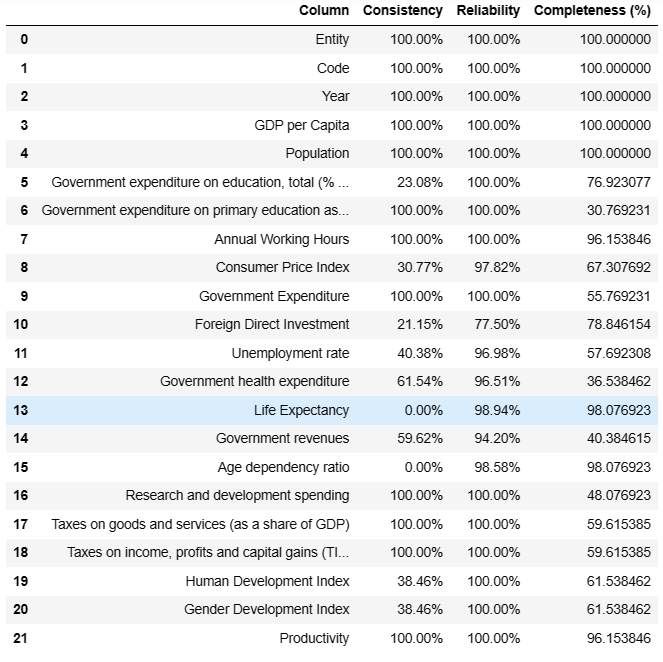
Figure5: Some Visualizations



Figure6: Heatmap Correlation Analysis

4.3 Data quality test

To evaluate the data quality, we also calculated the Consistency, Reliability, and Completeness of the data in the dataset. The final results are as follows:

Figure7: Compute the KPIs