**HO CHI MINH UNIVERSITY OF SCIENCE**

**FACULTY OF INFORMATION TECHNOLOGY**

**REPORT HW2**

**Subject: Data Visualization Class:** 22KHDL1

**Students:**

Trương Thuận Kiệt – 22127224 Phạm Minh Mẫn - 22127257

**TP. Hồ Chí Minh, tháng 3, năm 2025**

Contents

[**1.** **Group information and task assignments** 3](#_Toc192441760)

[**2.** **External Used Libraries** 3](#_Toc192441761)

[**3.** **Data Crawling** 3](#_Toc192441762)

[**a.** **Spotify API** 3](#_Toc192441763)

[**b.** **Mystreamcount** 6](#_Toc192441764)

[**c.** **Getgenre and Musicbrainz** 8](#_Toc192441765)

[**4.** **Data Overview and Preprocessing** 8](#_Toc192441766)

[**a.** **Shape of data** 8](#_Toc192441767)

[**b.** **Meaning of each row** 8](#_Toc192441768)

[**c.** **Meaning of each column** 8](#_Toc192441769)

[**d.** **Duplicate** 9](#_Toc192441770)

[**e.** **Data Retyping** 9](#_Toc192441771)

[**f.** **Missing Values** 10](#_Toc192441772)

[**g.** **Distribution of Numerical Values** 10](#_Toc192441773)

[**i.** **Songs Dataset** 10](#_Toc192441774)

[**ii.** **Artists Dataset** 11](#_Toc192441775)

[**h.** **Distribution of Categorical Values** 12](#_Toc192441776)

[**i.** **Songs Dataset** 12](#_Toc192441777)

[**ii.** **Artists Dataset** 14](#_Toc192441778)

[**5.** **Questions** 15](#_Toc192441779)

[**a.** **Question 1** 15](#_Toc192441780)

[**b.** **Question 2** 20](#_Toc192441781)

[**c.** **Question 3** 23](#_Toc192441782)

[**d.** **Question 4** 25](#_Toc192441783)

[**e.** **Question 5** 27](#_Toc192441784)

1. **Group information and task assignments**

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Full Name** | **Tasks** | **Rate of completion** |
| 22127224 | Trương Thuận Kiệt | * Crawl data from World Bank API. * Provide the overview of dataset (numerical columns). * Pose and answer problems 1, 2 and 3. * Write report. | 100% |
| 22127257 | Phạm Minh Mẫn | * Crawl data from World Bank API. * Provide the overview of dataset (categorical columns). * Preprocessing data * Pose and answer problems 4, 5 and 6. * Write report. | 100% |

1. **External Used Libraries**
   1. **Requests**

* Used to crawl data from API.
  1. **Geopandas**
* To plot the map of the world.

1. **Data Crawling**
   1. **Initialize the features to get**

* A screenshot of a computer program

  Description automatically generated
* In this Lab, we will focus on socio-economic features and crawl 2 continents **Asia** and **Europe**.
  1. Get information of countries
* Since the World Bank API does not provide API to get countries by continent therefore, we had to crawl all countries at first.

A screen shot of a computer program

Description automatically generated

* 1. **Get Asian and European Countries**
* After getting the ISO Code of all countries, we would get only Asian and European Countries based on the continent.

A computer screen with colorful text

Description automatically generated

* 1. **Get Data**
* By running loadDataContinent we will start crawling data of each continent from 2020 to 2023.

1. **Data Overview and Preprocessing**
   1. **Shape of data**

* Raw data has 2472 rows and 32 columns. (2 countries removed since the names are not relevant)
* European countries data has 744 rows.
* Asian countries data has 1200 rows.
  1. **Meaning of each row**
* Each row represents a country-year record, meaning it provides various **economic, demographic, and social statistics** for a given country in a specific year.
  1. **Meaning of each column**

|  |  |
| --- | --- |
| Columns | Description |
| Total Population | Total number of people in the country. |
| Female Population | Total number of females in the country. |
| Male Population | Total number of males in the country. |
| Birth Rate | Number of births per 1000 people per year. |
| Death Rate | Number of deaths per 1000 people per year. |
| Compulsory Education Dur. | Number of years of mandatory education. |
| Employment in Industry (%) | Percentage of the workforce employed in industry. |
| Employment in Agriculture (%) | Percentage of the workforce employed in agriculture. |
| Employment in Services (%) | Percentage of the workforce employed in services. |
| Female Employment in Agriculture (%) | Percentage of employed females working in agriculture. |
| Female Employment in Industry (%) | Percentage of employed females working in industry. |
| Female Employment in Services (%) | Percentage of employed females working in services. |
| Unemployment (%) | Total unemployment rate as a percentage. |
| GDP in USD | Gross Domestic Product of the country in US dollars. |
| National Income per Capita | Average income per person in the country. |
| Net Income from Abroad | Net inflow or outflow of income from foreign transactions. |
| Agriculture Value Added (in USD) | Contribution of agriculture to GDP in US dollars. |
| Agriculture Value Added (%) | Percentage of GDP coming from agriculture. |
| Industry Value Added (%) | Percentage of GDP coming from industry. |
| Services Value Added (%) | Percentage of GDP coming from services. |
| Expense (% of GDP) | Government expenses as a percentage of GDP. |
| GDP per Capita | GDP divided by total population (economic output per person). |
| GDP Growth Rate | Percentage change in GDP from the previous year. |
| Foreign Direct Investment | Net inflow of foreign direct investments. |
| Exports of Goods and Services | Total exports as a percentage of GDP. |
| Imports of Goods and Services | Total imports as a percentage of GDP. |
| Male Life Expectancy | Average life expectancy of males. |
| Female Life Expectancy | Average life expectancy of females. |
| Inflation Rate | Percentage change in the general price level of goods and services |
| Year | The year of which the data is recorded |
| Country | The name of the country |

* 1. **Check duplicates**
* Luckily, there are no duplicate rows.
  1. **Data types of each column**
* All the features are in the correct type. However, feature **Year** is in incorrect type, therefore, we will retype this.
  1. **Missing values of each column**

|  |  |
| --- | --- |
| **Columns** | Number of missing values |
| Birth Rate | 171 |
| Death Rate | 199 |
| Compulsory Education Dur. | 462 |
| Employment in Industry (%) | 434 |
| Employment in Agriculture (%) | 410 |
| Employment in Services (%) | 410 |
| Female Employment in Agriculture (%) | 434 |
| Female Employment in Industry (%) | 410 |
| Female Employment in Services (%) | 410 |
| Unemployment (%) | 410 |
| GDP in USD | 76 |
| National Income per Capita | 1075 |
| Net Income from Abroad | 355 |
| Agriculture Value Added (in USD) | 321 |
| Agriculture Value Added (%) | 321 |
| Industry Value Added (%) | 284 |
| Services Value Added (%) | 318 |
| Expense (% of GDP) | 927 |
| GDP per Capita | 76 |
| GDP Growth Rate | 177 |
| Foreign Direct Investment | 438 |
| Exports of Goods and Services | 352 |
| Imports of Goods and Services | 352 |
| Male Life Expectancy | 239 |
| Female Life Expectancy | 239 |
| Inflation Rate | 498 |

* As can be seen that feature **National Income per Capita** has too many missing values therefore we will remove this feature.
* However, there are still missing values in most of columns, therefore, we would need to fill these values.
  + 1. **Filling missing values**
  + First, we drop countries that have too many null columns and countries removed are **Andorra, American Samoa, Faroe Islands, Gibraltar, Greenland, Isle of Man, Liechtenstein, Monaco, Northern Mariana Islands, Nauru, Korea, Dem. People's Rep., San Marino, Tuvalu, Kosovo.**
  + However, there are some % features that are missing. For example. in our dataset employment in sectors and value added and we had to deal this separately.
    - We first dealt with rows that had 2 non-missing features, we calculated the remaining column by taking 100% and minus 2 other columns' values.
    - Then we imputed missing value by mean of country, if mean of country is not available (NaN), we imputed by mean of region.
    1. **Standardization**
  + Finally, we standardized to ensure sum of three features equal to 100.
  1. **Data Distribution**
     1. **Numerical columns**

1. **Population Features**
   * + To visualize both female and male population also total population, we used stack plot since it could help us visualize the total population.

|  |  |
| --- | --- |
| A blue and pink graph  Description automatically generated | A graph of a number of people  Description automatically generated with medium confidence |

* + - Asia’s population is much higher than the EU’s, starting at 3.5 billion in 2000 vs the EU’s 700 million. By 2023, Asia grows to 4.5 billion while the EU stays at 700 million. Asia’s steady growth contrasts with the EU’s stable, slightly fluctuating numbers, reflecting Asia’s younger demographics and higher fertility rates vs the EU’s aging population and low fertility. EU’s gender is evenly split, while Asia has a slight male bias (51% male, 49% female), possibly due to regional social factors.

1. **Birth/Death Rate**
   * + For these features, we will calculate average rate by years and use line plots to visualize the trends.

A graph of a number of people with different colored lines

Description automatically generated

* + - The EU has nearly no population growth, with birth and death rates close and a slight post-2015 death rate rise (aging or COVID-19), while birth rates stay low, showing stagnation. Asia’s birth rate, starting high, dips until 2015 then rises, with a low, stable death rate reflecting a younger population and driving growth. Asia’s higher birth and lower death rate confirm its expansion, while the EU’s aging population ties to its rising death rate and stability.

1. **Compulsory Education Duration**
   * + For this feature, we use bar charts to perform comparison of EU and Asia over the time period.

A graph of blue and pink bars

Description automatically generated

* + - The EU’s education system, stable at 9–10 years, includes primary (6–7 years) and lower secondary (3–4 years) Asia’s 8–9 years compulsory education duration reflects progress, mixing advanced systems (Japan, South Korea: 9 years) with developing ones (India: 8 years), supported by policy and growth. The EU’s uniformity contrasts with Asia’s diversity.

1. **Employment Features**
   * + Since these features are in % and sum of them will be 100% therefore, we will use stacked bar chart to visualize the percentage of each accounting for by year.

A graph of different colored bars

Description automatically generated

* + - The EU’s economy depends heavily on services (70% to 80%) compared to Asia (30% to 50%), with Asia’s higher industry (20%–25% vs. EU’s 15%–25%) and agriculture shares (50% to 30% vs. EU’s 5% to 3%) reflecting its manufacturing and agrarian roots. The EU’s economy that focused on services, showed declining industry and agriculture, while Asia shifts from agriculture to services and industry. This ties to the EU’s stable, aging population and high GDP per capita, supported by 9–10 years of education, vs Asia’s growth and expanding economy.

1. **Gender Employment**
   * + Since there are only 2 genders over the years, we chose to use line graph for better visualization.

A graph with numbers and lines

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

* + - In the EU, men dominate agriculture (80–85%) and industry (70–75%), but services were dominated by women (50% to 60%). In Asia, female agriculture jobs drop (40% to 25%) due to mechanization or urban migration, while industry (30% to 40%) and services (30% to 50%) rise, reflecting economic growth, a younger population, and higher birth rates (17.5–20). The EU’s developed economy favors women in services, while Asia’s developing economy shows broader gender shifts, tied to population growth and education duration time.

1. **Value Added Features**
   * + Since these features are in % and sum of them will be 100% therefore, we will use stacked bar chart to visualize the percentage of each accounting for by year.

A graph of different colored bars

Description automatically generated with medium confidence

* + - The EU’s economy, with 80% in services, thrives on knowledge and high female employment, supported by strong education, driving high GDP per capita, while industry and agriculture decrease. Asia’s services rise (30% to 50%) and agriculture drops (40% to 20%), as industry peaks then dips (30% to 25%), showing a key sector. The EU’s focus on services contrasts with Asia’s developing mix of sectors, which match their demographic and employment patterns.

1. **GDP Features**
   * + For GDP per Capita, we chose box plot provide high-level information immediately, offering general information about a group of data’s symmetry, skew, variance, and outliers.

A graph of a graph of the gdp per capita

Description automatically generatedA graph with a bar and a line

Description automatically generated with medium confidence

* + - Asia’s GDP per capita varies widely, with a median of $10,000 USD and outliers above $80,000, showing stark differences like between Japan and other poorer countries. Europe’s is more uniform, with a median of $40,000 USD, mostly between $20,000 and $60,000, reflecting stable and developed economies, while Asia’s wider range highlights its economic diversity and development gaps.
    - To visualize the trend of GDP over the years, line chart is the best option here.

|  |  |
| --- | --- |
| A graph showing the growth of the gdp  Description automatically generated | A graph with a line graph and red and blue lines  Description automatically generated |
| A graph of growth rate  Description automatically generated | |

* + - As Asia’s total GDP, despite lower per capita, has outpaced Europe since 2010, growing steadily at 4%–6%. Maybe Europe spends more on defense, while both saw high 2020–2023 spending due to COVID-19 healthcare. Both faced the 2009–2010 Financial Crisis and 2020 pandemic, but Asia’s faster growth shows greater economic dynamism than the EU’s stable trajectory.

1. **Net Income from Abroad and FDI**
   * + For these features, we chose line graph to plot the trends over time.

|  |  |
| --- | --- |
|  |  |

* + - The EU’s net income fluctuated quite strongly, with sharp dips and spikes from the 2009 financial crisis and 2020 COVID-19, often negative or fluctuating, while Asia’s net income stayed positive and stable until a sharp decline post-2015, possibly due to economic crises, trade disruptions, or policy shifts. The EU’s FDI is highly unstable, peaking in 2004–2008 and dropping in 2016–2023, whereas Asia’s FDI shows consistent, sustained growth.

1. **Import and Export**
   * + Mirror bar chart is a great way to reflect for displaying import and export side by side in a format that's both aesthetically pleasing and straightforward to compare.

|  |  |
| --- | --- |
|  |  |

* + - Europe’s export and import values are stable but were heavily impacted by the 2009 financial crisis and 2020 COVID-19 pandemic, with imports exceeding exports since 2010, showing a trade deficit. Asia’s export and import values are also stable from 2000 to 2023, but imports consistently surpassed exports, indicating a persistent trade deficit over the period.

1. **Inflation Rate**

|  |  |
| --- | --- |
|  |  |

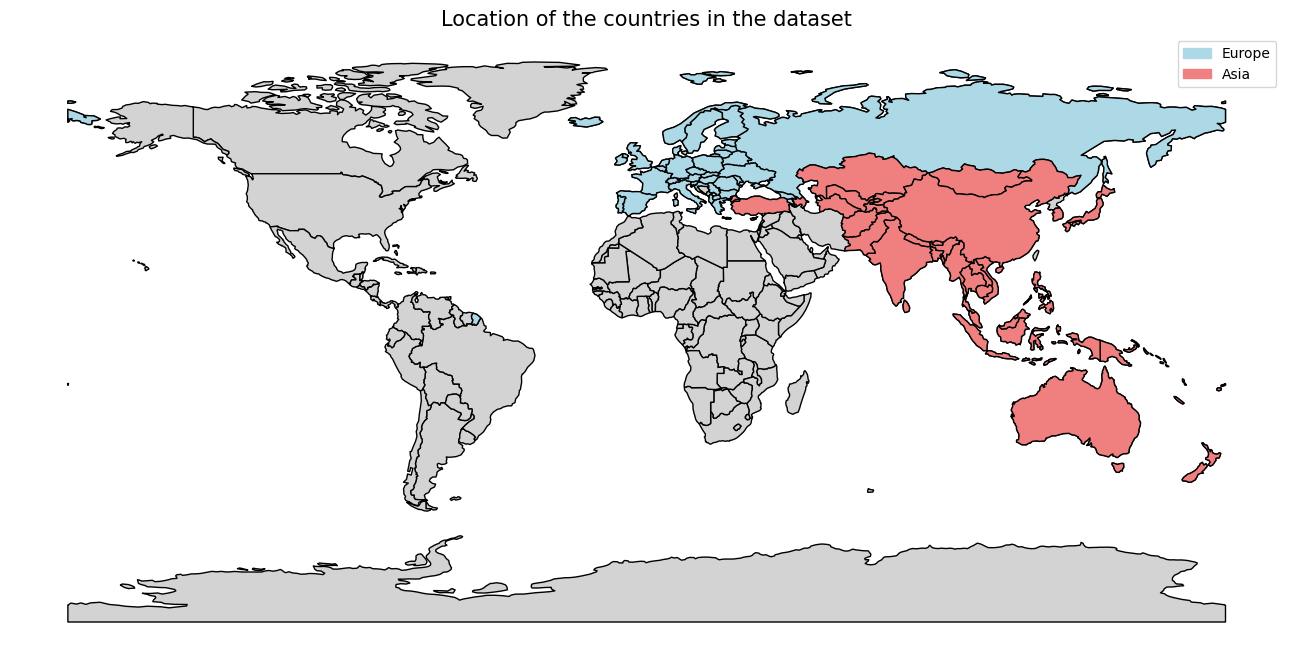
* + - In Europe, Belarus, Ukraine, and Serbia faced the highest inflation, with Eastern Europe (e.g., Ukraine, Russia) generally seeing higher rates than Western nations (e.g., UK, France) due to conflicts, and crises. In Asia, Türkiye, Myanmar, and Sri Lanka experience high inflation, caused by political instability, and crises, common among developing economies with debt burdens.

1. **Life Expectancy**

|  |  |
| --- | --- |
|  |  |

* + - Europe’s Switzerland and Iceland had the highest life expectancy with great healthcare, economies, and lifestyles. In Asia, Hong Kong, Macau, and Japan are top countries with great healthcare, living standards, joined by high-income nations like South Korea, Singapore, and New Zealand. No Southeast Asian countries rank high, showing poorer nations lag. Strong economies and healthcare boost life expectancy in both regions.
    1. **Categorical Columns**

1. **Countries**
   * + To visualize the location of countries in our dataset, map was the best option here.



* 1. **Abnormal Values and Outliers**
* To check abnormality and outliers of time-based data. I will use z-score to visualize the distribution of each feature by years. Since the demographic and socio-economic datasets seem to be quite skewed, we will choose threshold 2.
* **Population Features:** Asia’s total populations’ outliers might include countries with extreme population growth like India and China, whereas in the EU smaller nations like Luxembourg could be outliers. Male bias can be seen in Asia, while EU is balanced due to fewer outliers.
* **Birth/Death Rate:** High-fertility countries like Afghanistan could be outliers, while in the EU, countries with very low birth rates like Italy might stand out. In terms of Death Rate, EU outliers might include countries like Bulgaria with higher death rates due to aging populations. In Asia, conflict zones like Afghanistan might show spikes during crises.
* **Compulsory Education Dur.:** The EU’s 9–10 years is uniform, while Asia’s 8–9 years varies. In Asia, countries like India (8 years) or rural areas with less might be outliers compared to Japan (9 years). The EU likely has fewer outliers due to standardized policies.
* **Employment in Industry/Agriculture/Services**: In Asia, heavily agrarian countries like Nepal (over 60% in agriculture) could be outliers. In the EU, service-dominated economies (80%) might flag countries like Greece (still around 10% in agriculture) as outliers.
* **GDP Features**: China’s GDP is an outlier in Asia compared to smaller economies like Samoa. In Europe, smaller nations like Malta (GDP around $18 billion) might be outliers against Germany. In terms of GDP per Capita, High-income Asian countries like Singapore and low-income ones like Afghanistan are outliers. In Europe, outliers might include Luxembourg. For GDP Growth Rate: Outliers in Asia include negative growth in Afghanistan (-0.18), while in the EU, countries like Ukraine (post-2020) might show sharp declines.
* **Foreign Direct Investment**: The EU’s FDI is unstable (peaking 2004–2008, dropping 2016–2023), while Asia’s is consistent. EU outliers might include sharp drops (e.g., post-Brexit UK), and in Asia, countries like Myanmar (post-2015 decline) could be flagged.
* **Exports/Imports of Goods and Services**: Both regions show trade deficits, but crisis years (2009, 2020) might flag outliers.
* **Male/Female Life Expectancy**: High life expectancy in Japan (over 85 years) and Hong Kong could be outliers in Asia, while in Europe, lower life expectancy in Eastern Europe (e.g., Ukraine, around 65 years) might be flagged.
* **Inflation Rate**: In Asia, countries like Myanmar and Türkiye also show high inflation due to instability, while in Europe, Eastern nations like Belarus and Serbia have higher rates than the West.
  1. **Correlation between variables**
     1. **European Countries**

A colorful squares with black text

Description automatically generated with medium confidence

* + In Europe, a strong negative correlation (-0.84) between Employment in Agriculture and Services reflects a shift to service-based economies, mirrored by a -0.64 correlation between Agriculture and Services Value Added. Higher GDP per capita correlates with increased government spending (0.81) and lower unemployment (-0.42). Unemployment and Female Unemployment are highly correlated (0.98), as is service sector employment (0.85), showing gender inequality. Education duration positively correlates with life expectancy (0.23-0.24), while agricultural employment negatively impacts it (-0.55 males, -0.54 females). Exports and Imports are strongly correlated (0.90), and larger economies (GDP in USD) see higher net income from abroad (0.79), indicating integrated trade and financial inflows.
    1. **Asian Countries**

A colorful chart with numbers

Description automatically generated with medium confidence

* + In Asia, a strong negative correlation (-0.75) between Employment in Agriculture and Services shows a shift to service-based economies, less intense than Europe’s (-0.84), with a -0.62 correlation between Agriculture and Services Value Added. Higher GDP per capita links to more government spending (0.87) and lower unemployment (-0.45), stronger than Europe. Unemployment and Female Unemployment correlate highly (0.95), as does agricultural employment (0.95), showing gender parity. Education duration and life expectancy correlate positively (0.25), slightly weaker than Europe, but agricultural employment strongly reduces life expectancy (-0.75), more than Europe (-0.55/-0.54). Exports and Imports correlate strongly (0.95), and larger economies (GDP in USD) have high net income from abroad (0.87), both slightly stronger than Europe (0.90, 0.79).

1. **Questions**
   1. **Question 1**
2. **What is question?**

How unemployment affects GDP over the years between 2 Asia and Europe?

1. **How to answer**
   * Overall trend and broad changes in GDP and unemployment between Asia and Europe over time.
     + To easily observe the trend of unemployment and GDP features over time, we will choose to use line chart.
   * Correlation between average unemployment rate and GDP between Asia and Europe across all years.