Student Name: Benjaporn Wongmayura Student ID: s3961136

Data Preparation.

Before cleaning data

I have retrieved three datasets (Primary.csv, Secondary.csv, and Total school age.csv), and I checked the shape to see if it still has the same number of columns and rows. Then, I read the readme.txt file. Reading is essential to understanding the big picture of the dataset, what each column means, how the data is collected, and its limitation. Then, I copied the data to the new data frames before cleaning so that the changes we will make do not affect the original dataset.

Cleaning data

I have cleaned the data frames separately one by one since each data has different errors:

1. I deleted the unnecessary columns, e.g., Data source and Time Period, since I will not use these two columns in the analysis.
2. I checked for the redundant white space using the value_count function to count if the total items are the same number as retrieved since there was no white space error. There is no need to use the strip function to ensure no white space before and after the specific word.
3. We knew that the data types were in the wrong type. Therefore, I converted from object to float.
4. I checked for any NaN values and found plenty of NaN values. I decided to drop those missing values.
5. I spotted that there was an impossible value for some numeric data fields. However, after going through all the previous steps, the row with an impossible value was already removed.
6. I wrote the data frame to the new .csv file as cleaned_Primary.csv, cleaned_Secondary.csv, and cleaned_Total_scholl_age.csv. These three files are now ready for data exploration.
7. Finally, for the task 2.3, I joined the data frame for analysis.

In this dataset, there are no duplicated country, redundant white space. Below are the types of error found in the data preparation process:

Error 1: Wrong Data Type

After I retrieved the data from primary.csv, secondary.csv, Total school age.csv files, I roughly checked the data types of each column by using dtype function and found that all the numerical columns are in object type. As a result, I converted the data type to float and deleted the string of "%" in each value for all the columns. Then, I added the string of "%" to the header to make it more meaningful and better represent the column.

Error 2: Missing Value

There are many ways to deal with the missing data, such as dropping the null value, setting the value to null, and imputing with 0 or mean value etc. Before deciding which approach to choose, we must understand why and how the data is missing. In this case, the data were missing at random (MCAR). Hence, I chose to drop the missing data.

Error 3: Impossible Value

I have run the sanity checks for the numerical columns such as Total, Rural (Residence), Urban (Residence). In the Secondary.csv file, since we know the maximum percentage is 100%, Ukraine cannot have 179% of the total percentage of secondary students' internet access level.

Error 4: Different Country Code

I have found that Angola had a different code between primary.csv and secondary.csv files. In my analysis, I have not used this column directly. So, it will not affect my visualization part. However, we should fix this issue to make the data frame standardized, consistent, and efficient. I have fixed this issue by using replace function to replace "AGOA" with "AGO". After fixing this issue, we can use the country code to join multiple data frames for future analysis.

Data Exploration

Task 2.1

2.1.1) Nominal Variable: Region

Justification: I chose Region to represent nominal variable column because Region can be categorized

Figure2.1.1.1:

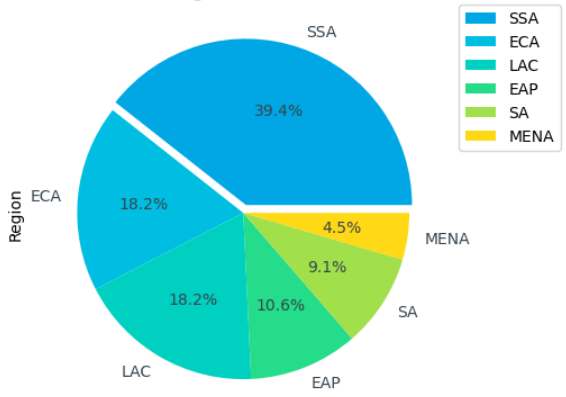


Figure2.1.1.2

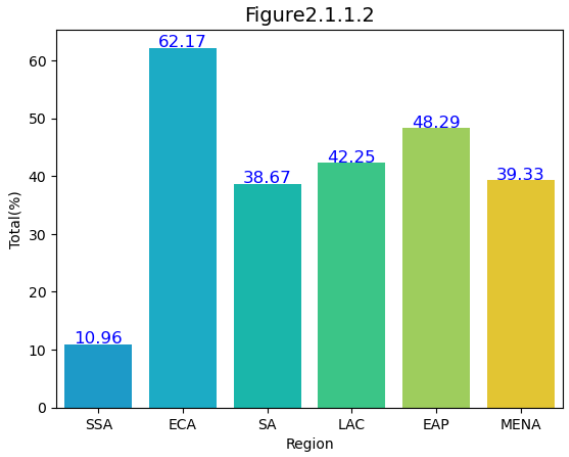


Figure 2.1.1.1: A pie chart of the total percentage of regions where primary school students can access the Internet.

This pie chart helps us answer which region the data owner collects most of the data from and which region the data are least collected from. We can see from the pie chart that most of the data are collected from countries in Sub-Saharan Africa (SSA) region, while the smallest number are from Middle East and North Africa (MENA).

Figure 2.1.1.2: A bar chart of average total percentage of primary student's internet access by region

Now that we know that this database is mainly collected from Sub-Saharan Africa countries. I am curious about how much each region contributed to its level of internet access of primary students. After I grouped each region and tried to find an average total percentage of internet access, I found out that ECA rank highest on the percentage of primary student's internet access level at 62.17%, followed by EAP at rate 48.29%. Whereas the least access internet of primary students came from SSA at 10.96%.

2.1.2) Ordinal Variable: Income Group

Justification: I chose Income Group to represent ordinal variable column because Income Group can be categorized and ranked.

Figure2.1.2.1:

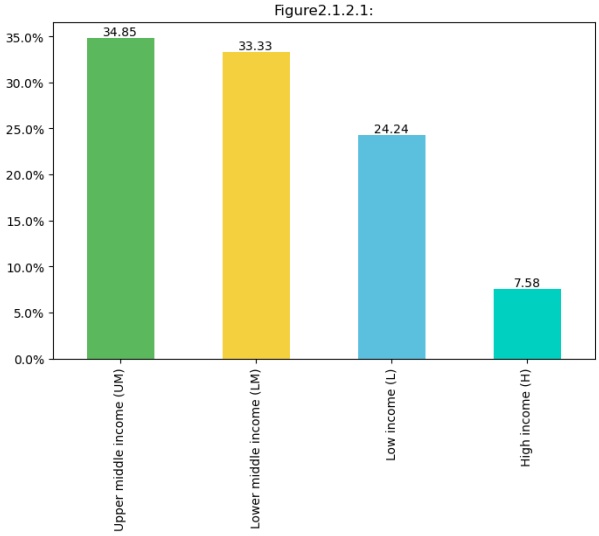


Figure2.1.1.2:

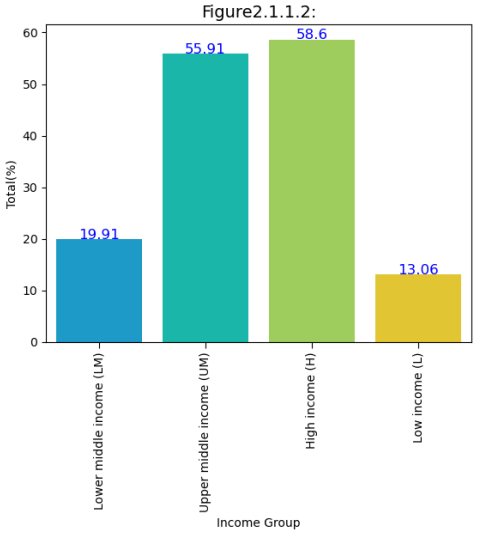


Figure 2.1.2.1: A bar chart of percentage of Income group where primary school students have the access to the internet.

The bar chart (Figure 2.1.2.1) illustrates the percentage of primary school students' access to the internet across various income groups. It indicates that most countries in the dataset fall into the upper middle-income category (34.85%), followed by lower middle-income, low-income, and high-income groups in descending order. This provides us with valuable insights into the income distribution of the countries analysed.

Figure 2.1.2.2: A bar chart of total percentage of primary school students having access to the internet by Income Group

My curiosity led me to investigate the contribution of each income group to the level of internet access for primary students. Upon grouping the data by income level and calculating the average percentage of internet access, I discovered that the high-income group had the highest percentage at 58.6%, followed by the upper middle-income group with 55.91%. Conversely, the lowest level of internet access for primary students was found in the low-income group, with a rate of only 13.06%.

2.1.3) Numerical Variable: Total

Justification: I chose Region to represent numerical variable column because Total is a number column that can be measured

Figure 2.1.3: Histogram of Total percentage of primary students that have internet connection at home

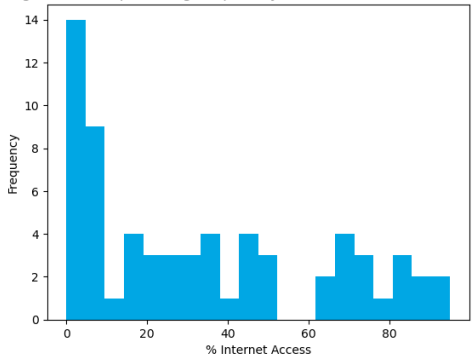


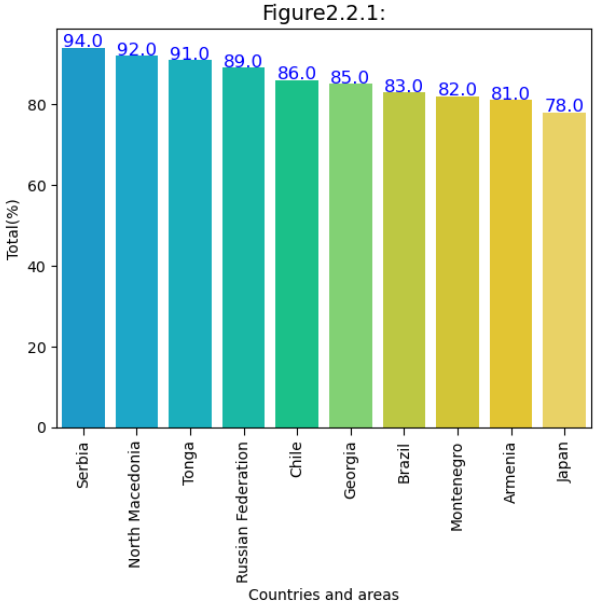
Figure 2.1.1.1: A pie chart of total percentage of regions where the primary school students have the access to the internet.

The histogram provides us with insight into the distribution of the total percentage of internet access level for primary students. Most countries in the dataset have a percentage between 0% and 10%, whereas the least common percentage falls between 50% and 60%. The data is positively skewed, with a right-leaning distribution.

Task 2.2

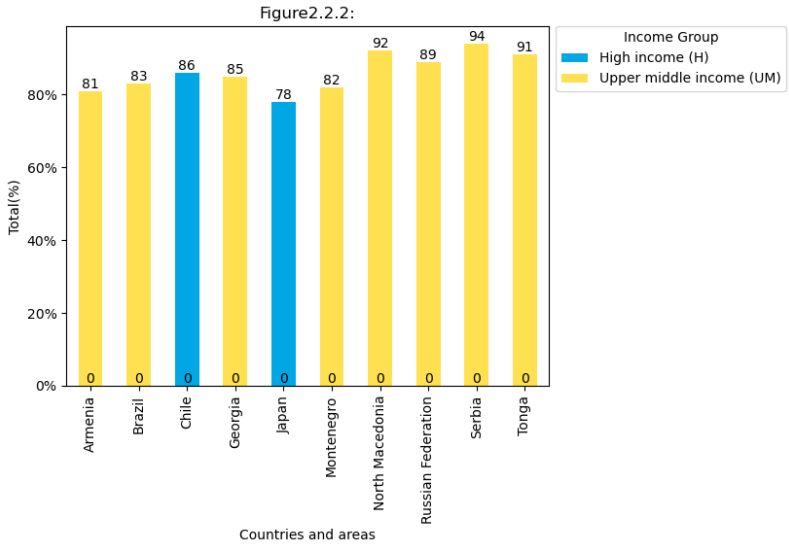
Top 10 countries with the highest percentage of total school age's internet access

Figure2.2.1:



Top 10 countries with the highest percentage of total school age's internet access by income group

Figure2.2.2:



Top 10 Countries with the highest total percentage of internet access of 3 -17 years old students (Rural residence VS Urban Residence)

Figure2.2.3:

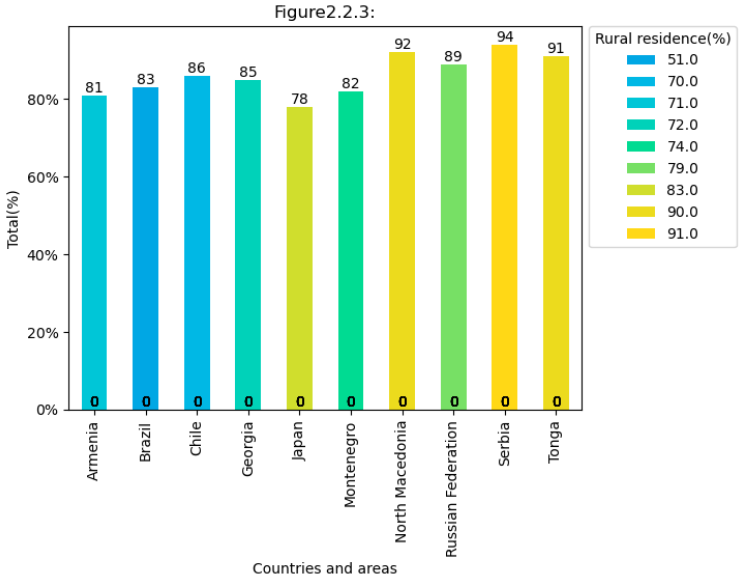
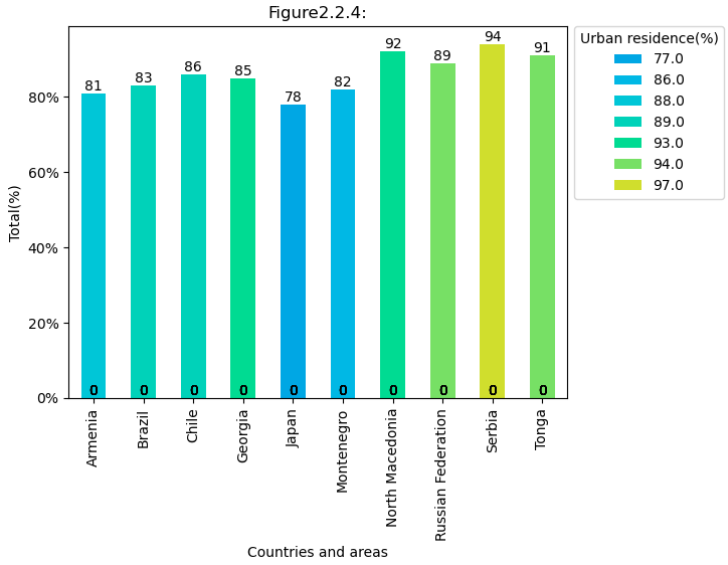
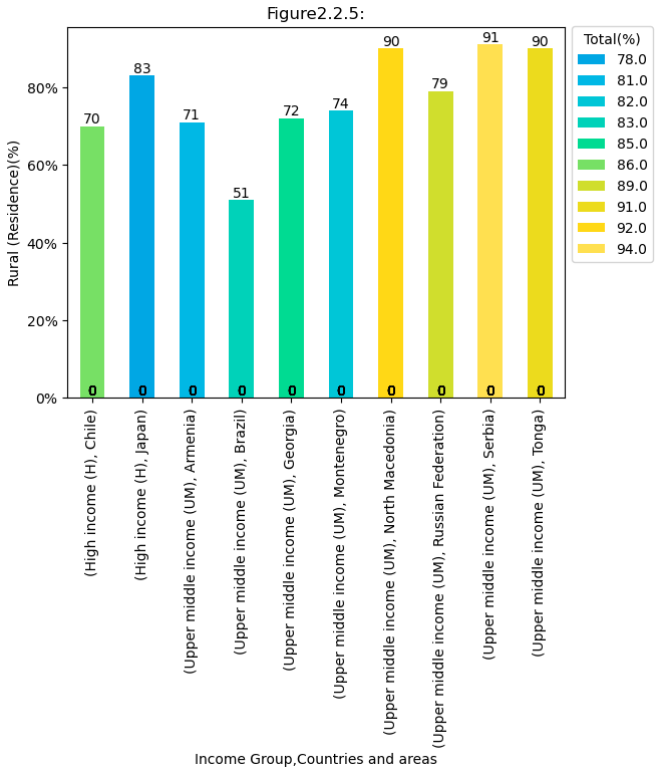


Figure2.2.4:



Top 10 Countries with the highest total percentage of internet access of 3 -17 years old students by countries and income group (Rural residence VS Urban Residence)



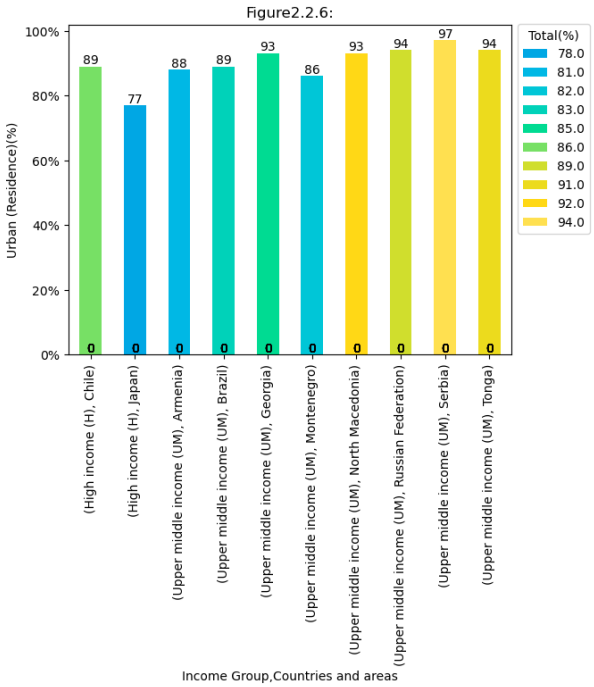


Figure 2.2.1: Top 10 countries with the highest percentage of total school age's internet access

The country with the highest percentage of internet access for students between the ages of 3-17 is Serbia, with a rate of 94%. North Macedonia and Tonga follow with the second and third-highest rates, respectively.

Figure 2.2.2: Top 10 countries with the highest percentage of total school age's internet access by income group

We can see from the chart that most of the countries in the top 10 countries with highest internet access level in 3-17 years old students are from upper middle-income countries. Only few are from high-income countries. This suggests that there is a correlation between income level and internet access level, with higher income levels indicating better internet access.

Figure 2.2.3: Top 10 Countries with the highest total percentage of internet access of 3 -17 years old students (Rural residence)

We can see that the country with the highest total percentage of internet access at rate 94% has the percentage of children (live in rural area) in a school attendance age that have internet connection at home at 91%. Whereas Japan, which has the lowest total percentage of internet access at rate 78%, has the percentage of children (live in rural area) in a school attendance age that have internet connection at home at 83%. The range of internet access level in rural area are considerably wide, ranging from 51% to 94%.

Figure 2.2.4: Top 10 Countries with the highest total percentage of internet access of 3 -17 years old students (Urban Residence)

We can see that the country with the highest total percentage of internet access at rate 94% has the percentage of children (live in urban area) in a school attendance age that have internet connection at home at 97%. Whereas Japan, which has the lowest total percentage of internet access at rate 78%, has the percentage of children (live in urban area) in a school attendance age that have internet connection at home at 77%. The range of internet access level in rural area are narrow, ranging from 77% to 97%. We can see that the level of internet access in the urban areas is surprisingly similar.

Figure 2.2.5: Top 10 Countries with the highest total percentage of internet access of 3 -17 years old students by countries and income group (Rural residence)

Figure 2.2.6: Top 10 Countries with the highest total percentage of internet access of 3 -17 years old students by countries and income group (Urban Residence)

Task 2.3

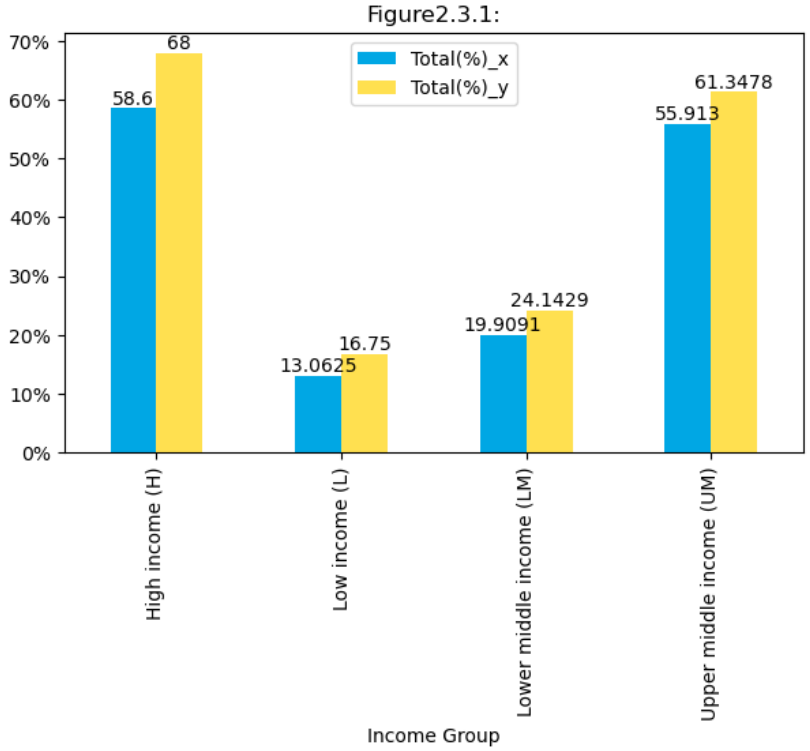


Figure 2.3.1: A bar chart comparing between average total percentage of primary student's internet access and secondary student's internet access from Lower Middle-Income background. *Please note that: (total(%)_x = "total(%)_primary, total(%)_x = "total(%)_secondary)

Secondary students from Lower middle income (LM) (24.14%) have higher average total percentage of internet access level than the Primary students' average total percentage (19.9%).