Assignment 1: Data Cleaning and Summarising

Deepthi Suresh (S3991481)

Data Preparation

The data set consists of data regarding the digital connectivity information of children in a school attendance age that have internet connection at home. The data consists of information on 90 different countries and 11 variables. The following errors were found and corrected.

Error Type 1: Duplication

In order to identify if there were any duplicate rows in the dataset, count of the subset of "Countries and areas", "Income Group" and "Time period" columns were taken using the value_counts() function. This returned an output showing Togo and Guatemala having duplicate entries. Using the data.duplicated() function we were able to locate the 3 rows with the exact same data. The drop_duplicates() function was used to remove these duplicates and keep only 1 relevant entry.

Error Type 2: Data Entry Errors - Typos

On running a value_counts() function on the Income group, it was found that there were data entry errors. According to the brief, there are only 4 types of Income group - low, lower-middle, upper-middle, and high, however the dataset consists of 6 types - Upper middle income (UM), Lower middle income (LM), Low income (L), High income (H), Lower middle income (LLM) and Lower middle income (LMM). Clearly Lower middle income (LLM) and Lower middle income (LMM) are data entry errors that need to be corrected. In order to correct these loc() function was used and these rows were assigned to the correct category of Lower middle income (LM).

Error Type 3: Impossible values and Inconsistencies

On running a value_counts() function on the Time period column, two errors were found. The first error was the inconsistency in entering the Years. Some time periods were entered as a single year whereas others were entered as a range of years. In order to make the data consistent, the starting year of the time period range was assumed to be the year on which the survey was conducted. To rectify this, a function was defined to split the time periods and take in only the starting year as the data for the country.

The value_counts() shows that there are impossible values of '2076' and '3562'. Since there is no way of ascertaining the year in which these data were collected, it is best to remove these entries.

Error Type 4: Missing Values

On the initial check, there were 23 missing values across Residence (Rural), Residence (Urban), Wealth quintile (Poorest) and Wealth quintile (Richest). Since these are the 4 important numeric columns, a check for NAs across all 4 columns were conducted by using the isna() function. There were 4 such rows, and these were dropped from the dataset by using the drop() function as they do not provide most of the information required for the analysis.

There were 4 remaining missing values, which were imputed with the mean of each column to not lose any good data. The remaining dataset consisted of 81 rows as opposed to the 78 if we dropped all rows with missing values.

Error Type 2: Impossible values in percentages

The total percentage that a variable can take up is 100%, so a sanity check was conducted among the numeric columns to see if this rule was violated. This resulted in one row having 110% in the "Wealth quintile (Richest)". The value of the column was manually edited to 100% by using the loc() function. The percentage of students in the richest quantile can never be 110%.

Data Exploration

Task 2.1

In order to create a boxplot maplotlib was imported along with pandas library. A side-by-side boxplot was created using the boxplot() function with "Total" as column and Region as the "By" input. Appropriate titles and labels were also produced. The Median (of the total percentage) for each Region was computed using the groupby() and median() functions

Findings:

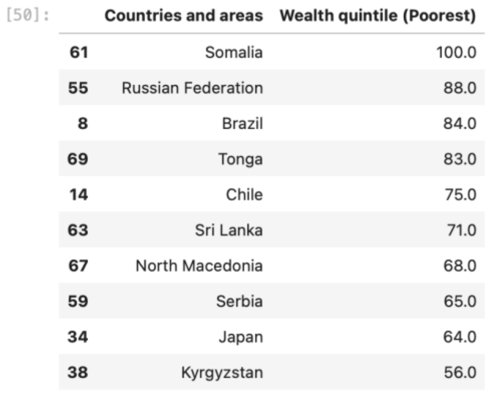
ECA had the highest median of 77.5% and SSA had the lowest median of 7% indicating a significant gap between the percentage of students with digital connectivity in these two regions. The median also indicates that half of the countries in these region falls below this percentages. One interesting observation that can be noted is that all the other regions have a median within the range of 33% and 47%.ECA has the highest average number of students at 66.5% who has internet at home whereas SSA has the lowest average of 15% amongst all the regions.SA has the highest interquartile range of 59% and SSA has the lowest of 13%. SSA has countries with both lowest and the highest number of students with digital connectivity with a 0% and 100%, however the country under SSA with a 100% connectivity is clearly an outlier. The boxplot shows that SSA has 3 outliers. The region with maximum connectivity is ECA at a 94%. In LAC and MENA 75% of the countries has less than 58% and 41%

Task 2.2

In order to calculate the mean for poorest and richest quintiles, a subset of the data was created and the mean() function was used on it. These were then sorted in Descending order using the sort_values() functions.

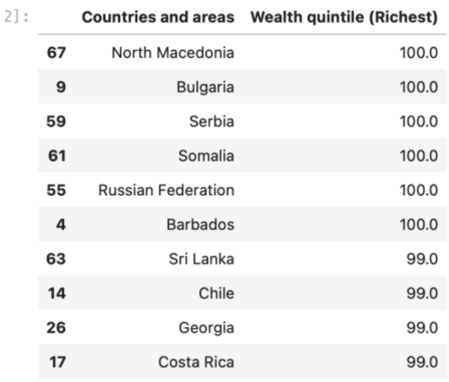
Finding:

The richest wealth quintiles have a mean percentage of 61.06 % of school-age children who have internet connection at home, whereas in the poorest quintiles the number drops to 18.24%, a significant difference. The top 10 countries with the highest mean percentage in the poor and rich quintiles are as follows:



[50]:

	Countries and areas	Wealth quintile (Poorest)
61	Somalia	100.0
55	Russian Federation	88.0
8	Brazil	84.0
69	Tonga	83.0
14	Chile	75.0
63	Sri Lanka	71.0
67	North Macedonia	68.0
59	Serbia	65.0
34	Japan	64.0
38	Kyrgyzstan	56.0



2]:

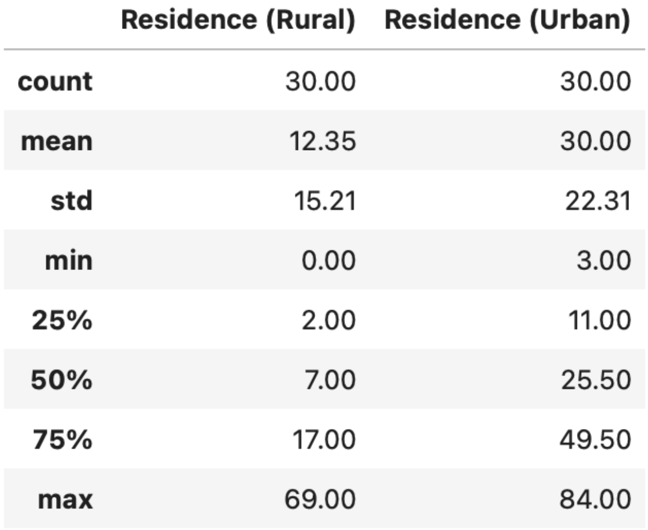
	Countries and areas	Wealth quintile (Richest)
67	North Macedonia	100.0
9	Bulgaria	100.0
59	Serbia	100.0
61	Somalia	100.0
55	Russian Federation	100.0
4	Barbados	100.0
63	Sri Lanka	99.0
14	Chile	99.0
26	Georgia	99.0
17	Costa Rica	99.0

Task 2.3

In order to compare the percentages of different categories of Residence (Rural versus Urban) within the Lower middle income (LM) group, a subset of the data was created called the LM_data. The statistics for the same was created using the describe() function.

Findings:

In the lower middle-Income group, there are 30 countries each under Rural and Urban residential areas. As expected, the digital connectivity of students in the urban residential area is higher at 30% compared to the rural areas that is at a 12.35%. There are countries in the rural area with 0% connectivity which can be quite alarming, whereas the same number in the urban area is 3%. However, looking at the maximum connectivity values, the country with the maximum connectivity in the Urban area is at an 84% whereas at the rural areas it is 69%. 75% of the countries in the rural areas lies at a 17% digital connectivity whereas in the urban areas it is 49.50%. The urban areas have a larger standard deviation of 22.31% compared to the rural areas indicating a higher variability. The median for urban residence is significantly higher at a 25.50% compared to the rural areas. Overall, the digital connectivity is higher in the urban areas compared to the rural areas within the lower middle-Income group.



	Residence (Rural)	Residence (Urban)
count	30.00	30.00
mean	12.35	30.00
std	15.21	22.31
min	0.00	3.00
25%	2.00	11.00
50%	7.00	25.50
75%	17.00	49.50
max	69.00	84.00

Use of AI Tools

AI tools like Val and ChatGPT can significantly simplify tasks and clarify objectives. They are invaluable for understanding complex concepts and their interactive nature allows for addressing follow-up questions effectively. AI can provide code snippets and logic for data cleaning and manipulation, making these operations more manageable.

Additionally, AI tools aggregate various problem-solving suggestions that might be hard to find otherwise. They help identify syntax errors in code, explain why certain pieces of code don't work, and suggest improvements to enhance program efficiency. With numerous functions available in programming, it's challenging to know them all. AI can assist in discovering different functions and providing information about them, thus enhancing learning. Moreover, AI is excellent for interpreting visualizations and statistics, helping to identify trends and patterns that might be otherwise overlooked.

I have used AI tools to deepen my understanding of technical concepts, especially given my non-technical background. AI has been helpful in breaking down code, answering follow-up questions not covered by Python documentation, and streamlining the search for solutions. Instead of visiting multiple websites, I use AI to pinpoint issues and then conduct targeted research. AI has also been crucial in correcting syntax errors in my code.

References

- GeeksforGeeks. (2024, March 21). *How to fill NAN values with mean in Pandas?* GeeksforGeeks. https://www.geeksforgeeks.org/how-to-fill-nan-values-with-mean-in-pandas/
- *pandas.DataFrame.sort_values — pandas 2.2.2 documentation*. (n.d.). https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sort_values.html