Assignment 1: Data Cleaning and Summarising

Thuy Linh Do (s3927777)

Data Preparation

Error Type 1: Missing Values

- **Description**: Missing values were detected in several columns, particularly those related to percentages of internet connectivity and wealth quintiles. In some cases, these missing values appeared as NaN (Not a Number), indicating gaps that could compromise the accuracy of the analysis if left unaddressed.
- **Approach**:
  - **Identification**: Upon loading and displaying the dataset, I observed that certain cells contained NaN, signaling missing data. To ensure that all missing values were identified, I employed the `isnull()` method, which checks each cell for missing values. This method was used in combination with `sum()` to count the number of missing values in each column, providing a comprehensive count of missing values in each column. This allowed me to pinpoint where the data was incomplete and required cleaning.
  - **Handling**:
    - Since the columns Residence (Rural), Residence (Urban), Wealth quintile (Poorest), and Wealth quintile (Richest) are critical for analysis, any rows with missing values in these columns were removed using the `dropna()` method. Specifically, the subset parameter was used to target these key columns, ensuring that only rows with complete data in these important areas were retained.
    - After removing the rows with missing values in the specified columns, I rechecked the dataset using `isna().sum()` to confirm that all missing values in the critical columns had been handled.
- **Justification**: By removing rows with missing values in crucial columns, I ensured that the dataset used for analysis was complete and reliable. This method prevented the analysis from being skewed by incomplete data, which could lead to inaccurate or biased results. The focus on key columns helped maintain the dataset's overall integrity while ensuring that the most relevant data was retained for accurate analysis.

Error Type 2: Duplicates

- **Description**: Duplicate records were found in the dataset, particularly within columns related to demographic information. Retaining these duplicates could lead to multiple counts of the same observation, distorting the results.
- **Approach**:
  - **Identification**:
    - To begin, I displayed the total number of rows in the dataset to establish a baseline for comparison before any cleaning was performed. This step ensured that any changes in the number of rows after removing duplicates could be easily tracked.
    - I then used the `duplicated()` method to identify how many rows were duplicates. This method flagged rows that were identical across all columns, allowing me to quantify the extent of duplication in the dataset.
  - **Handling**:
    - After identifying the number of duplicate rows, I removed them using the `drop_duplicates()` method. This action ensured that each observation in the dataset was unique.
    - To verify the effectiveness of the cleaning process, I displayed the number of rows again after the duplicates were removed. This comparison confirmed that the

duplicates had been successfully eliminated, without needing to manually inspect the data.

- **Justification**: Eliminating duplicate rows is crucial for maintaining the accuracy of statistical inferences, particularly in demographic analyses where each record should represent a unique individual or entity. By showing the number of rows before and after the cleaning process, I ensured that the dataset was accurately cleaned, providing a solid foundation for subsequent analysis.

Error Type 3: Outliers

- **Description**: Outliers were identified in columns related to the percentage of the population living in rural and urban areas and the wealth quintiles. More specifically, the Wealth quintile (Richest) column had a maximum value of 110%, which is unrealistic for a percentage and likely indicates an error in the data. Such outliers, if not addressed, could skew the analysis and lead to incorrect conclusions.
- **Approach**:
    - **Identification**: To identify these outliers, the percentage data in the relevant columns was first converted from string format to numeric format. This was necessary because percentage values stored as strings could not be directly evaluated or manipulated numerically. After conversion, a simple `max()` check was applied to each column to detect values greater than 100%, highlighting these as outliers.
    - **Handling**: Once the outliers were identified, they were corrected by capping the values at 100% with the lambda function. This ensures that all percentage data remains within the valid range of 0% to 100%, preserving the integrity of the dataset. This method was chosen because it effectively handles the outliers without removing any rows, thereby maintaining the completeness of the dataset. After applying these corrections, I checked to ensure that all values in these columns were within the valid range of 0% to 100%. The corrected dataset was then verified to confirm that no values exceeded the acceptable range.
- **Justification** The process of converting percentage data to numeric format was essential for accurate outlier detection and correction. By capping values at 100%, I ensured the integrity of the dataset while retaining all original records, thereby making the dataset suitable for reliable analysis.

Error Type 4: Incorrect Data Entries in the 'Time period' Column

- **Description**: The 'Time period' column contained incorrect data entries such as "3562", "2099", etc. These values are not valid for representing periods, which should be within the range of realistic years, up to the current year (2024). Such erroneous data entries could compromise the accuracy of time-based analysis.
- **Approach**:
    - **Identification**: To identify these incorrect entries, I examined the unique values in the 'Time period' column using `unique()`. This code provided a list of all unique values in the column, allowing me to spot any out-of-range or unrealistic entries.
    - **Handling**: To correct the invalid entries, I implemented a function to clean the 'Time period' values. This function splits the 'Time period' values by the hyphen ('-') and ensures that each year within the period is less than or equal to 2024. Any year greater than 2024 is replaced with 2024. This function effectively addresses entries where years are beyond the current year, ensuring that all values are within the acceptable range. After applying this function to the dataset, I confirmed that the cleaning was effective by checking the unique values in the 'Time period' column again.
- **Justification**: The function to clean the 'Time period' column was necessary to correct unrealistic entries that could skew any analysis dependent on accurate time data. By setting all years beyond the current year to 2024, the data is brought within a plausible range, ensuring that future analyses will be based on valid and realistic periods.

Data Exploration

Task 2.1

2.1.1. Visualization of side-by-side boxplot

To understand the distribution of the total percentage of school-age children with internet access at home across different regions, I created a side-by-side boxplot. This boxplot allows for a visual comparison of the percentage distribution across all regions, helping to identify any significant disparities or outliers within each region.

The x-axis represents the different regions, and the y-axis shows the percentage of children with internet access. The boxplot provides a summary of the central tendency (median), spread (interquartile range), and potential outliers for each region. By examining the spread and the position of the median within each box, we can infer the variation and skewness of internet access across regions.

The boxplot is an effective way to visualize this data because it allows us to see how the distribution of internet access varies not just between individual data points, but across entire regions. This can help in identifying regions where access to the internet is more or less consistent, as well as regions with notable outliers.

2.1.2. Median Calculation by Region

After visualizing the data, I computed the median percentage of school-age children with internet access at home for each region. The median is a robust measure of central tendency, particularly useful when the data is skewed or contains outliers, as it provides a better representation of the typical value within each region compared to the mean.

The calculated medians offer insight into the typical internet access level within each region, allowing for a straightforward comparison. For instance, a region with a higher median value indicates that more than half of the children in that region have internet access at home, suggesting better overall connectivity.

2.1.3. Summary

- The boxplot revealed that some regions have a wider range of internet access percentages, indicating significant variability within those regions.
- The median calculations reveal stark disparities in internet access among school-age children across different regions. Europe and Central Asia (ECA) stand out with the highest median, suggesting better overall connectivity, while Sub-Saharan Africa (SSA) shows the most significant challenges with a median of just 7.0%. These findings point to the need for targeted efforts to improve internet infrastructure, particularly in regions like SSA and South Asia, where connectivity is most lacking.

These findings can be further explored to understand the underlying factors contributing to the disparities in internet access across regions, such as economic conditions, infrastructure, and policy differences.

Task 2.2: Wealth Quintile Analysis

2.2.1. Analysis

To explore the impact of wealth on internet access, I calculated the mean percentage of school-age children with internet access at home for the 'Poorest' and 'Richest' wealth quintiles across different countries. The mean for the 'Poorest' quintile is 18.59%, while for the 'Richest' quintile, it is 61.98%. This stark difference illustrates the disparity in internet access between the wealthiest and poorest populations.

After computing these means, I identified the top 10 countries with the highest percentages of school-age children with internet access at home for both wealth quintiles. The sorting of data in descending order allowed us to highlight countries where the percentage of children with access is particularly high within each wealth category.

2.2.2. Summary

- Mean Analysis: The average percentage of school-age children with internet access at home is significantly lower for the 'Poorest' quintile (18.59%) compared to the 'Richest' quintile (61.98%). This result clearly indicates a substantial digital divide based on wealth.

- Top 10 Countries for Wealth Quintile (Poorest):
  - The countries leading in providing internet access to their poorest populations include Somalia (100%), the Russian Federation (88%), and Brazil (84%). Notably, Somalia's 100% access rate is exceptional, though this could reflect specific data characteristics or reporting standards.
  - Other countries like Tonga, Chile, and Sri Lanka also perform relatively well, with percentages ranging from 71% to 83%, demonstrating successful efforts to bridge the digital divide for their poorest populations.
- Top 10 Countries for Wealth Quintile (Richest):
  - The top performers for the 'Richest' quintile include several countries with a 100% internet access rate, such as the Russian Federation, Serbia, and Somalia, indicating complete or near-complete access among the wealthiest segments.
  - The list also includes countries like Costa Rica and Armenia, which have nearly universal internet access for their richest populations, with rates around 99%.

These findings emphasize the critical role that wealth plays in determining internet access, with the wealthiest populations enjoying far greater connectivity. The identification of leading countries in both categories provides valuable insights into which nations are succeeding in closing the digital divide and which may need targeted interventions to ensure equitable access to digital resources.

Task 2.3: Internet Access in Lower Middle Income Group (Rural vs Urban)

To thoroughly analyze internet access disparities between rural and urban areas within the Lower middle income (LM) group, I used three statistical measures: mean, median, and standard deviation. These measures provide a comprehensive view of the central tendency, typical experience, and variability in internet access across different regions.

2.3.1. Comparison by Mean Percentage

Mean is a key measure that provides an overall average, summarizing the typical level of internet access within each residence type (rural or urban). The calculation of mean percentages helps to highlight general trends in how internet access is distributed among school-age children in LM countries.

- Findings:
  - Rural Areas: The mean percentage of internet access is approximately 12.5%, indicating limited access for children in these areas.
  - Urban Areas: The mean percentage is approximately 28%, reflecting better access but still far from universal.
- Analysis: The gap between rural (12.5%) and urban (28%) areas underscores a significant digital divide, with urban children being more than twice as likely to have internet access. Despite better access in urban areas, the relatively low mean indicates that many children still lack connectivity, impacting their education and development.

2.3.2. Comparison by Median Percentage

Median provides the middle value in the dataset, offering insight into what a "typical" level of internet access might look like for rural and urban areas. The median is particularly useful in cases where the data might be skewed or contain outliers, as it is less affected by extreme values than the mean.

- Findings:
  - Rural Areas: The median is approximately 4.5%, suggesting that more than half of rural areas have minimal or no internet access.
  - Urban Areas: The median is approximately 23.5%, indicating that while urban areas fare better, significant gaps remain.
- Analysis: The 19% gap between rural (4.5%) and urban (23.5%) medians highlights the severe lack of access in rural areas, further emphasizing the digital divide and the need for targeted improvements.

2.3.3. Comparison by Median Percentage

Standard deviation measures the amount of variation or dispersion in the percentages of internet access. A higher standard deviation indicates that the data points (in this case, internet access percentages) are spread

out over a wider range of values, while a lower standard deviation suggests that the values are closer to the mean.

- Findings:
  - Rural Areas: The standard deviation is 16%, indicating limited variation, with consistently low access across rural areas.
  - Urban Areas: The standard deviation is 23%, suggesting greater variability, with some urban areas having significantly better access than others.
- Analysis: The higher variability in urban areas reflects differences in infrastructure and resources, while the lower variability in rural areas points to uniform challenges in connectivity.

2.3.4. Summary

The analysis of internet access in the Lower middle income (LM) group, separated by rural and urban residence types, reveals significant disparities:

Urban areas have notably higher mean and median percentages of internet access compared to rural areas, indicating better overall connectivity. However, the variability (as shown by the standard deviation) is also higher in urban areas, suggesting that while some urban regions are well-connected, others may still struggle with limited access.

Rural areas, on the other hand, show consistently low levels of internet access, with little variation, reflecting widespread challenges in providing connectivity.

These findings highlight the need for targeted interventions, particularly in rural areas, to address the digital divide and ensure that all children, regardless of their location, have equal opportunities to access the internet. Addressing these disparities is crucial for promoting educational equity and supporting the development of all regions within the LM group.

Use of AI Tools

In completing Tasks 1 and 2, I utilized ChatGPT in the following ways:

- **Coding Guidance**: I received help in structuring the code for handling data issues in Task 1, including tips on efficient methods to handle missing values, duplicates, and outliers. Since I am not good at coding, I made many mistakes and errors during the implementation. ChatGPT helps me with debugging the code and running the code successfully.
- **Conceptual Understanding**: ChatGPT assisted in clarifying statistical concepts like mean, median, and standard deviation, which were crucial for the analysis in Task 2. For instance, these tools can clarify the significance of statistical measures and how they relate to the data being analyzed.
- **Efficiency and Time-Saving:** Using AI tools can significantly reduce the time required to complete tasks by automating repetitive aspects of coding, providing instant feedback on potential errors, and offering quick explanations of concepts. This allows more time to be spent on analyzing the results and drawing meaningful conclusions.

By leveraging AI tools, I was able to approach the tasks with greater confidence, ensuring that both the technical and explanatory components of the assignment were handled effectively.