Assignment 1: Data Cleaning and Summarising

Sudarshanan Gujuluwa Sundaram Santharam (s4085030)

This report details the pre-processing and exploration of a dataset related to percentage of children in a school attendance age that have internet connection at home. The aim is to identify errors, clean, process, and analyse the data for meaningful insights.

Data Preparation

Error Type 1: Remove leading and trailing spaces

- Initially, leading and trailing spaces was cleaned by stripping to make sure the dataset doesn’t contain this error.This ensures data consistency and accuracy.

Error Type 2: Removal of duplicate rows

- Duplicate rows based on unique combinations of key columns (subset=['ISO3', 'Countries and areas', 'Region', 'Sub-region']) was identified and removed in the dataset. This reduces the redundancy and improves the reliability of data. The ‘duplicated’ and ‘drop_duplicates’ methods of dataframe object were used to achieve this task.

Error Type 3: Outlier Detection and Correction

- **Income Group Outliers**

  Outliers in the Income Group column were identified based on frequency counts. Inconsistent values, such as "Lower middle income (LLM)" and "Lower middle income (LMM)", were replaced with "Lower middle income (LM)" to attain uniformity.

- **Percentage Validation (0 <= 'Valid Percentage' <= 100)**

  The dataset's percentage values were checked to make sure they fell between 0% and 100%. Any values that fell outside of this range were replaced with the null values. These null values will be handled later.

- **Detect and Fix Incorrect Extreme Percentage Values using Box Plot**

  Box plot was utilized to detect the outliners in the ‘Total’ column grouped by ‘Region’ column. It was evident from the box plot that few extreme percentage values such as 100% and 0% were incorrect. The below rules were applied to fix the issue:

  - If Total is 100%, then both Rural and Urban percentages should also required to be 100%.
  - If both Rural and Urban percentages were 100%, then Total should be 100%.

  Again, boxplot is used to check if the issue is fixed (Refer Fig 1). Further, boxplot for other columns containing percentage values were analysed to detect any outliers but none were discovered.

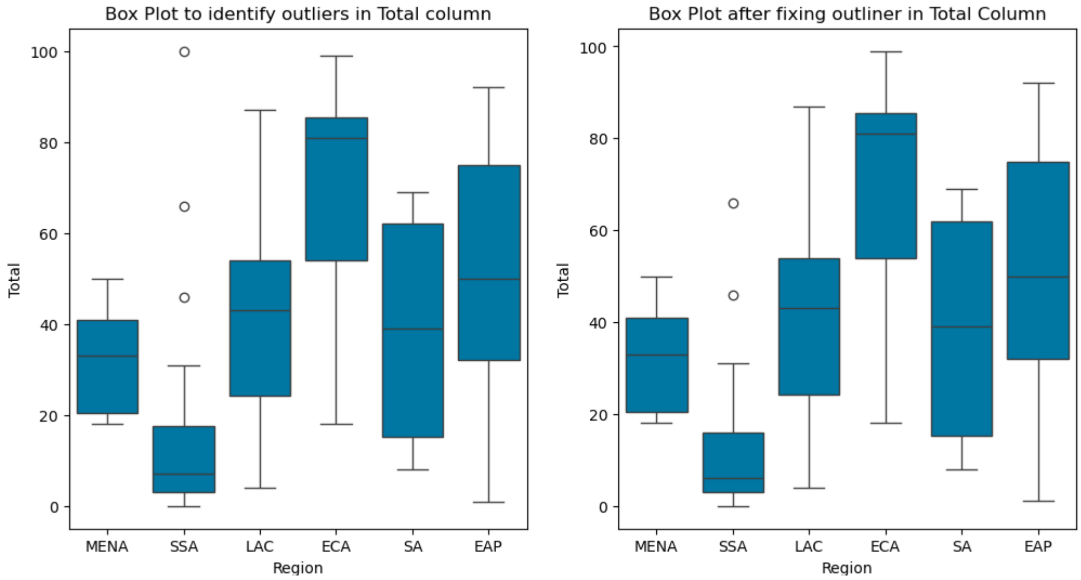


Fig 1

Error Type 4: Split the Time Period Column and fix the invalid years

- The Time Period column was divided into Start Year and End Year columns which helps in later exploration.
- If Time period value contains one year instead of two, then fill both "Start year" and "End year" was filled with the same year.
- Time period column was dropped to reduce the data redundancy.
- In order to maintain the accuracy of the data, entries that did not fall within the 1900–present year range were eliminated in both ‘Start year’ and ‘End year’ column. Hear the year range(1900-present) were decided based on the box plot for start and end year column

Error Type 5: Handle Missing values

- Observation containing three or more missing values were dropped.
- Remaining missing values are filled with the mean of each column calculated within each region. This strategy preserved the data’s context while filling in gaps, resulting in a more comprehensice dataset for analysis.
- The missing values were addressed at the end to ensure that all null values introduced during the previous error corrections were properly handled.

Error Type 6: Data type convertion for percentage columns (String to integer)

- Columns with percentage values are stored as String rather than numerical values. These column values are converted to integer for calculations during data exploration stage. For statistical calculations, such as figuring out the mean and median, this conversion was necessary.

Data Exploration

Task 2.1

This boxplot shows how different geographic regions school age children's internet connectivy varies.

Connectivity in East Asia and the Pacific (EAP) ranges approximately from 0% to 80%. The median is about 40%, which shows that eventhough a lot of children have some level of internet connectivity, there is a lot of variation in this region.

Generally, connectivity in Europe and Central Asia (ECA) is high, with the majority of data points lying between 60% and 80%. Then median in this region (70%) shows the robust and steady internet access

Sub-Saharan Africa (SSA) region has a median connectivity rate of 10%, which is generally poor. The majority of children in this region have extremely poor internet connectivity

There are huge differences in internet access worldwide, as illustrated by the box plot. Sub-Saharan Africa has the lowest median connectivity, whereas Europe and Central Asia have the highest.

Task 2.2

The richest quintile percentage is much higher than the poorest quintile percentage, which clearly indicates the digital divide between them.

Countries like Serbia, North Macedonia, Chile, and Sri Lanka appear in both lists, suggesting these countries have significant representation in both the poorest and richest quintiles.

Top 10 countries with the highest percentages for Wealth quintile (Poorest)

1. Russian Federation
2. Brazil
3. Tonga
4. Chile
5. Sri Lanka
6. North Macedonia
7. Serbia
8. Japan
9. Kyrgyzstan
10. Montenegro

Top 10 countries with the highest percentages for Wealth quintile (Richest)

1. Bulgaria
2. Barbados
3. Serbia
4. North Macedonia
5. Chile
6. Armenia
7. Costa Rica
8. Colombia
9. Sri Lanka
10. Georgia

Task 2.3

Boxplot, lineplot and histogram were used to visualize the distribution of Total internet connectivity percentages across different regions.

The **median value** for urban residences is 25.5%, compared to just 7.0% for rural residences. This median comparison emphasizes the gap, as the midpoint of internet access in urban residences is substantially higher, indicating more consistent access in cities compared to rural regions.

The **mean percentage** of school-age children with internet access at home is significantly higher in urban residences (29.29%) compared to rural residences (12.25%). This indicates that, on average, children in urban residences have more access to home internet connectivity than those in rural residences.

Both urban and rural data show **positive skewness**, meaning that in both cases, there are some regions with higher connectivity that stretch the distribution towards the higher end.

Urban residences typically have more diversified internet connection than rural residences, which suffer with lower levels of internet connectivity. These plot effectively shows the digital divide between Urban and rural internet connectivity.

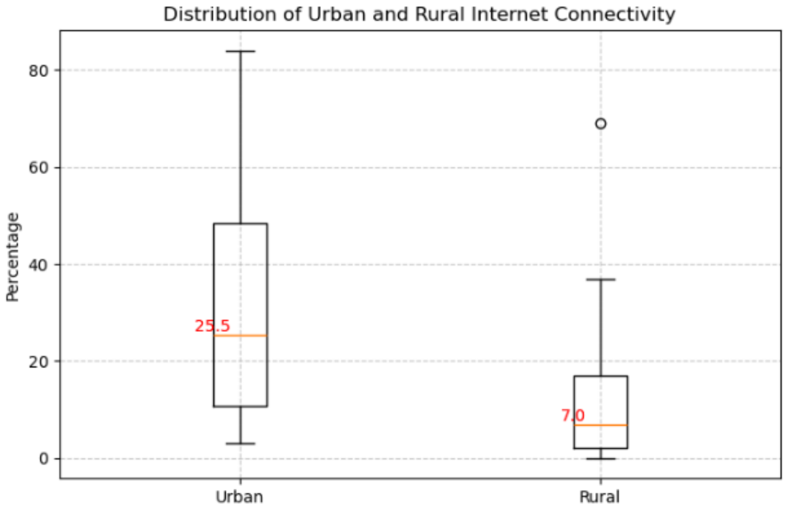
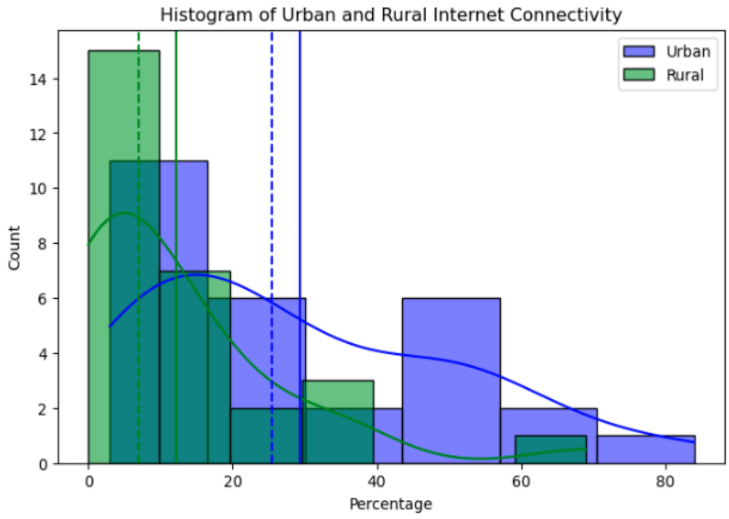


Fig 2

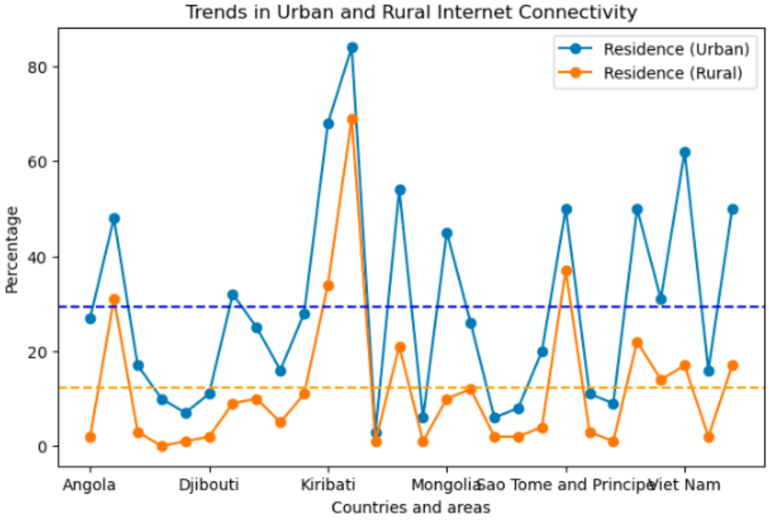


- - - Median Line

---- Mean Line

```
Skewness for rural: 2.09
Skewness for urban: 0.76
```

Fig 3



Use of AI Tools

AI Tool has been used to research and learn Python libraries for ploting various graphs and visualizing data. This helped to implement the necessary graphs using libraries like pandas, matplotlib, and seaborn for better analysis.

References

pandas Documentation: Used to explore various APIs and methods for handling DataFrames, as well as for data manipulation and cleaning.

https://pandas.pydata.org/docs/

Skewness - Wikipedia: Used to deepen my understanding of skewness and its significance in data analysis.

https://en.wikipedia.org/wiki/Skewness