Data Preparation

Effective data preparation enhances data analysis efficiency, minimizes errors and inaccuracies during data processing, and ensures that processed data is more accessible to users. It is a crucial step prior to modelling because feeding the model with poor-quality data results in unreliable predictions. The goal is to eliminate the following types of errors from each attribute:

- Data entry mistakes.
- Unnecessary white spaces.
- Invalid values.
- Missing values.
- Outliers.

There is no universal methodology for cleaning every attribute in the same way, as it depends on various factors such as attribute type, data recording methods, restrictions, and references. These factors, among others, will be illustrated below. However, for all attributes, two essential methods can be employed: 'value_counts()' and 'info()'. 'value_counts()' can typically identify data entry mistakes, unnecessary white spaces, and invalid values. In contrast, the info method can detect the number of missing values. Outliers can be detected using a boxplot and the interquartile range (IQR).

ISO3

For this attribute, 'value_counts()' identified redundant white spaces. According to the accompanying README file, ISO3 codes must consist of three characters. A for loop with a conditional statement checking the length of each observation suffices in this case.

Countries and Areas

'value_counts()' also detected redundant white spaces for this attribute. The README file indicates that countries and areas should follow the correct format found in the State of the World's Children Statistical Annex 2017. By loading the dataset and employing a for loop with conditional statements to check for discrepancies, any deviations from the standard format are considered errors.

Region

'value_counts()' found no errors for this attribute. To further verify this, manual overrules were used. The README file specifies that the region should be one of the following:

- EAP
- ECA
- EECA
- ESA
- LAC
- MENA

- NA
- SA
- SSA
- WCA

A for loop with a conditional statement was employed to ensure each observation belongs to this list.

Sub Region

The same method applied to the region was used for the sub-region, revealing five errors through manual overrules.

Income Group

The region and sub-region methods were also applied to the income group. The README file states that the income group should belong to one of the following categories:

- Low income (L)
- Lower middle income (LM)
- Upper middle income (UM)
- High income (H)

Manual overrules identified one error.

Total, Rural, Urban, Poorest, and Richest Attributes

For any percentage-type attribute in the dataset, two user-defined functions were utilized: ‘to_float()’ and ‘impossible_values()’. The ‘to_float()’ function converts the attribute type to float, while the ‘impossible_values()’ function checks for invalid values or data entry mistakes. If any are found, the ‘replace()’ method corrects the error. If no errors are detected, outliers are checked using a boxplot. Outliers identified in the boxplot are located using the IQR and replaced with either the mean or median, depending on the attribute's distribution. Skewed distributions use the median, while others use the mean. The .info function at the beginning of the file provides information on the number of missing values. If any are found, ‘isnull()’ is used, and missing values are replaced in the same manner as outliers.

Data Source

For this attribute, ‘value_counts()’ identified white space errors, with one instance found. Some data source names contain non-ASCII characters, which must be altered to ensure easy access for users. A quick Google search provided the correct naming for the survey, which was then replaced using the ‘replace()’ function.

Time Period

For this attribute, ‘value_counts()’ was used to assess the format of the time. According to the README file, the format should only include years. As a result, the ‘apply()’ function was employed to remove the months from any observations. Additionally, the README file specifies that the timeframe starts from 2010. ‘value_counts()’ revealed no entries prior to 2010. Logical manual overrules were applied to check for any entries beyond 2023, which were then replaced with the median.

Final step

Finally, it is crucial to scan the exported file for any missed errors. If any is found, the code need to be changed to fix the errors correctly.

Data Exploration

Data exploration is a critical step in any data science project, as it facilitates a deeper understanding of a dataset, making data navigation and utilization more straightforward. The more thoroughly an analyst or data scientist comprehends the data they are working with, the more accurate their analysis or model will be. An effective approach to data exploration is to ask thought-provoking questions. The deeper the questions, the more insights can be derived. The ‘group_by()’ function is likely the most crucial function for identifying relationships between attributes and addressing intriguing questions.

Task 2.1

Region (Nominal Attribute)

A combination of ‘value_counts()’ and ‘plot()’ enables the creation of a pie chart to visualize regional distribution. To gain further insights, the effect of the region on income groups was investigated. By posing this question, the distribution of income groups across each region became evident. Another question explored whether the region influenced the total number of children with internet connections at home. This inquiry allowed the relationships between region, income group, and total to be plotted in a single chart. Regions with High income (H) and Upper middle income (UM) appeared to have the highest total of children with internet connections at home.

Income Group (Ordinal Attribute)

Similar to the region, a pie chart displayed the distribution of income groups. To determine which type of residence is most likely to have an internet connection based on income group, a bar chart was created to visualize the relationship between income group and residence type. The same approach was taken with the wealth quintile.

Total (Numerical Attribute)

A bar chart depicted the distribution of the attribute. An engaging question to consider is whether the number of children with internet connections at home increases over time. A line plot was used to identify any trends.

Task 2.2 – 2.3

The ‘group_by()’ function can be used to address Tasks 2.2 and 2.3.

References

COSC2670 Practical Data Science slides by Dr. Yongli Ren.

A. Boschetti and L. Massaron, Python Data Science Essentials.

Murtaza Haider, Getting Started with Data Science: Making Sense of Data with Analytics.

D. Cielen and A. Meysman and M. Ali, Introducing Data Science.