Data preparation and preprocessing*

AUTHOR: TUSIIME GEORGE TREVOUR
COURSE: Software Engineering
REG NO.:21/U/07417/EVE
STD NO.: 2100707417

I. Introduction

The dataset at hand is a comprehensive collection of indicators related to public health and access to basic services across various countries and years. It encompasses a diverse range of metrics that shed light on key aspects of health, sanitation, and water services, allowing us to gain insights that go beyond basic observations. we can uncover relationships, patterns, and interactions among the indicators that might not be immediately apparent. This deeper understanding can lead to more informed decision-making, better policy formulation, and targeted interventions to address specific challenges within the realms of public health, sanitation, and water services.

II. Introducing the contents of the dataset

- Key Variables and Indicators:
- *A. Country Name and Code:*

  The dataset includes identification information for each country, enabling us to contextualize the indicators based on geographic and political divisions.
- *B. Year:*

  The year associated with the data provides a temporal dimension to the dataset, allowing us to track trends and changes over time.
- *C. Health Indicators:*

  The dataset encompasses indicators related to malaria, including 'Incidence of malaria' and 'Malaria cases reported.' These indicators offer insights into the prevalence and impact of malaria within different populations.
- *D. Sanitation and Water Services:*

  The dataset features indicators that gauge the extent of access to basic services. 'People using safely managed drinking water services' and 'People using safely managed sanitation services' offer a glimpse into the provision of clean water and proper sanitation facilities.
- *E. Population Distribution:*

  Indicators like 'Rural population (% of total population)' and 'Urban population (% of total population)' shed light on how populations are distributed between rural and urban areas in different countries.
- *F. Basic Services Access:*

  The dataset includes metrics on access to basic services, such as 'People using at least basic drinking water services' and 'People using at least basic sanitation services.' These indicators highlight the progress made in ensuring fundamental amenities for the population.
- *G. Geographic Coordinates:*

  'Latitude' and 'Longitude' provide geographical coordinates for each country, enabling spatial analysis and visualization of the data.
- Insights and Implications:
  Through the exploration of this dataset, we can uncover potential relationships between health, sanitation, water services, and population dynamics. Correlations between variables can provide valuable insights into the effectiveness of interventions and policies aimed at improving public health and access to essential services. Furthermore, this dataset serves as a foundation for evidence-based decision-making and targeted resource allocation to address critical challenges faced by communities across the globe.

III. Data Cleaning Activities

1. Determining the Shape of the Dataset

The shape of the dataset was determined, revealing the number of rows and columns. This overview provides an initial understanding of the dataset's dimensions which also helps to determine the efficient way to display it and keep track of our data during manipulation.

2. Getting Data Info

A summary of the dataset's basic information was obtained, including the count of non-null values, datatypes, and memory usage. that aids in understanding the data's characteristics.

3. Converting Year Datatype to Datetime

The 'Year' column was converted from integer to datetime datatype. This transformation facilitates chronological analysis and time-based visualizations and also easies the task of comparision base on year we achieved this by using the to_datetime() function.

Identify applicable funding agency here. If none, delete this.

4. Determining Datatypes

The datatype of each column was examined to ensure consistency and accuracy. Inconsistent datatypes can affect analysis and modeling outcomes.We achieved this using the datatype() function

5. Renaming Column Names

Column names were renamed for clarity and conciseness.we transformed long or unclear names into more descriptive labels to enhance readability and understanding. we achieved this using the rename() function

6. Checking for Duplicates

We checked for duplicate values in our dataset with the aim of identifing and removing them from the dataset to ensure data accuracy and consistency. these could arise from data collection errors or processing issues.fortunately there were none. We achieved this by using the duplicated() function the compares the rows to see if one is the duplicate of the other.

7. Data Normalization

We checked for values that were out of range based o the fact that %_of_rural_population when added to the %_of_urban_population the sum should be 100% but we found incidences where the sum exceeded the expected value which posed the need to normalise our data so as to remove the inconsistencies hence we created a new column to cater for the sum of %_of_rural_population and %_of_urban_population that we named %_of_total_population. we performed normalization on %_of_rural_population and %_of_urban_population to bring data values within a common scale. .

8. Handling Null Values

We addressed missing values using appropriate strategies to minimize their impact on analysis. Different indicators required distinct treatments:

we filled cells where values are ”Missing At Random”, which means that their values could be determined using values from other columns.

The algorithm is based on four equations which represent a mathematical relationship between some of the columns as shown below:

$$
\begin{aligned}
a.\ \text{\%\_using\_safe\_sanity\_services} &= (\text{\%\_of\_rural\_using\_s} \\
&\quad + (\text{\%\_of\_urban\_usin}
\end{aligned}
$$

$$
\begin{aligned}
b.\ \text{\%\_using\_atleast\_basic\_drinking\_water\_services} &= (\text{\%\_of\_rural\_using\_a} \\
&\quad \times \text{\%\_of\_rural\_popul} \\
&\quad \times \text{\%\_of\_urban\_popu}
\end{aligned}
$$

$$
\begin{aligned}
c.\ \text{\%\_using\_atleast\_basic\_sanity\_services} &= (\text{\%\_of\_rural\_using\_a} \\
&\quad \times \text{\%\_of\_rural\_popul} \\
&\quad \times \text{\%\_of\_urban\_popu}
\end{aligned}
$$

$$
\begin{aligned}
d.\ \text{\%\_using\_safe\_drinking\_water\_services} &= (\text{\%\_of\_rural\_using\_s} \\
&\quad \times \text{\%\_of\_rural\_popul} \\
&\quad \times \text{\%\_of\_urban\_popu}
\end{aligned}
$$

Eritrea is the only country with null values in the ’%_of_rural_population’ and ’%_of_urban_population’ columns, which can be calculated using values provided in the columns ’%_using_atleast_basic_sanity_services’, ’%_of_rural_using_atleast_basic_sanity_services’, ’%_of_urban_using_atleast_basic_sanity_services’, ’%_using_atleast_basic_drinking_water_services’, ’%_of_rural_using_atleast_basic_drinking_water_services’, and ’%_of_urban_using_atleast_basic_drinking_water_services’. And we latter normalised to ensure that the values we calculated were not out of range.

9. Getting the Percentage of Null Values

We calculated the percentage of null values in each column to assess data completeness and identify potential data quality issues.

IV. Conclusion:

The data cleaning process for the ”MalariaAfricaDataset.csv” dataset encompassed a series of crucial steps. By checking for duplicates, handling null values, renaming column names, determining dataset shape and datatypes, calculating the percentage of null values, obtaining data information, converting year datatype, and performing data normalization, the dataset was refined for subsequent analysis. The cleaned dataset now stands as a reliable foundation for exploring correlations, trends, and insights related to public health and malaria indicators across different countries and years.