Student Name: Vivek Aggarwal Student ID: S4015465

Data Preparation

The following are the steps I followed to successfully achieve completion of task 1:

- I started by importing libraries like matplotlib, pandas, and NumPy. In order to make my code more readable, I disregarded warnings.
- I then loaded every dataset and presented each one individually. The data contained numerous mistakes and inaccurate numbers. However, in order to make the data for task 1 intelligible, we had to clean it by removing null values and inaccurate values.
- I discovered null values in the richest, poorest, urban, and rural columns and made the decision to remove those empty values from all four columns.
- I printed the data's shape to see if any null values had been removed.
- Then, as we are currently in the year 2023, I looked for any inaccurate values that might be present in the time period (survey year) column, such as 3019. I also threw them away.
- To decrease the amount of data, I eliminated unnecessary columns like the region where the data was collected, renamed the sub-region column to region, concatenated the nations and ISO3 columns into one column, and so on.
- After that, I updated the numerical columns that would be used for data exploration to have float data types.
- To display the columns in a better order, I also reindexed the columns. Finally, I named this file "cleaned_primary.csv" after it had been cleaned.

Error 1

ValueError- comes when cant compute datatype in the required way. As I switched data types, I was experiencing this kind of problem, which is pretty typical. I looked at its datatype, and it was displaying an object that was a string and could not be converted to a float due to the "%" symbol. I used the astype () function to convert the string into a float type in order to fix the mistake by replacing the "%" sign with a blank space.

Error 2

```
KeyError: "['Urban_percentage'] not in index"
```

There was an error because I used percentage instead of %, and this Urban_percentage was not defined

Error 3

```
SyntaxError: (unicode error) 'unicodeescape' codec can't decode bytes i
n position 2-3: truncated \UXXXXXXXX escape
```

I could not solve this error

Data Exploration

Task 2.1

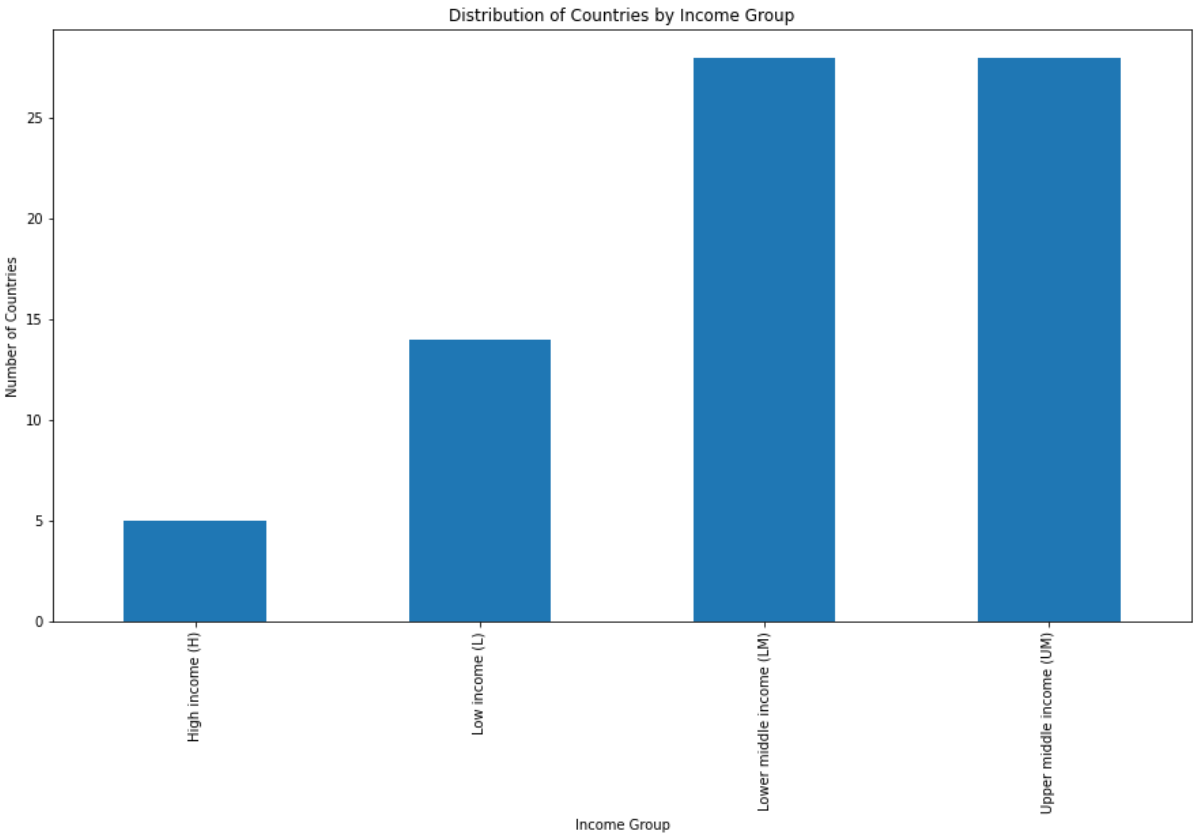
This code parses a dataset containing information about various countries. Their income groups, percentage of all children attending primary school, distribution of rural and urban populations. The first part of the code groups the data by income group and creates bar charts to show the distribution of countries within each group.

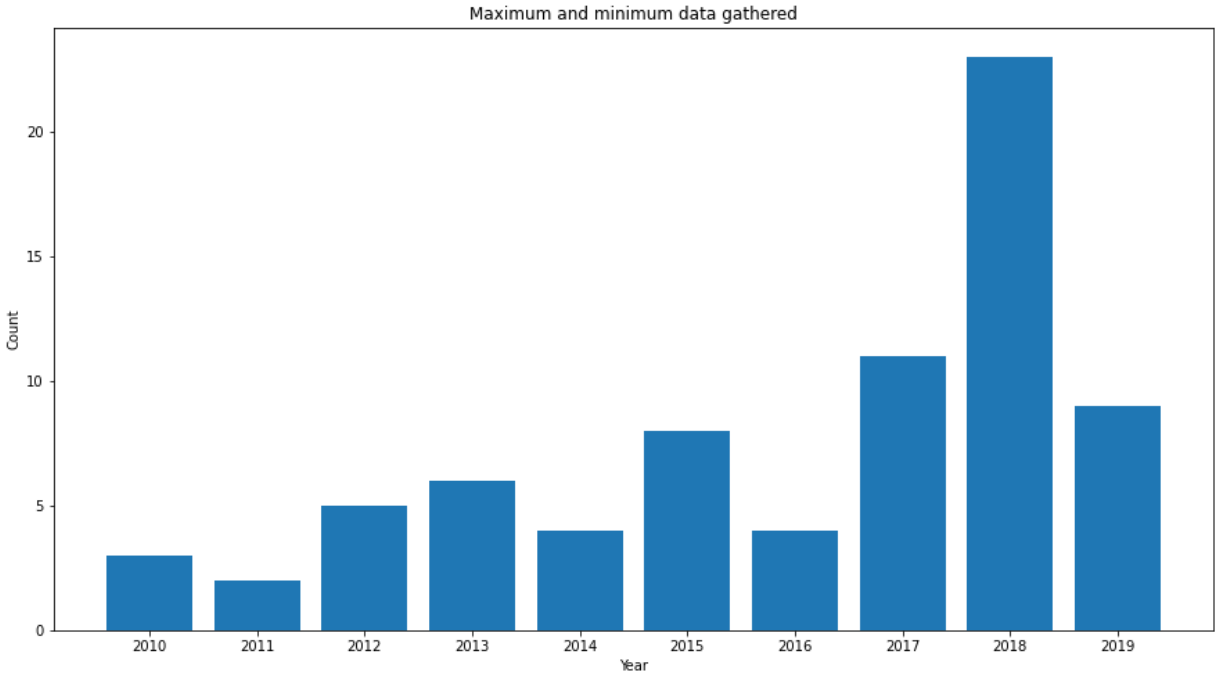
The following is the graph I printed in the notebook, and I picked these 2 attributes ,number of countries and income group, as the y and x axes respectively.

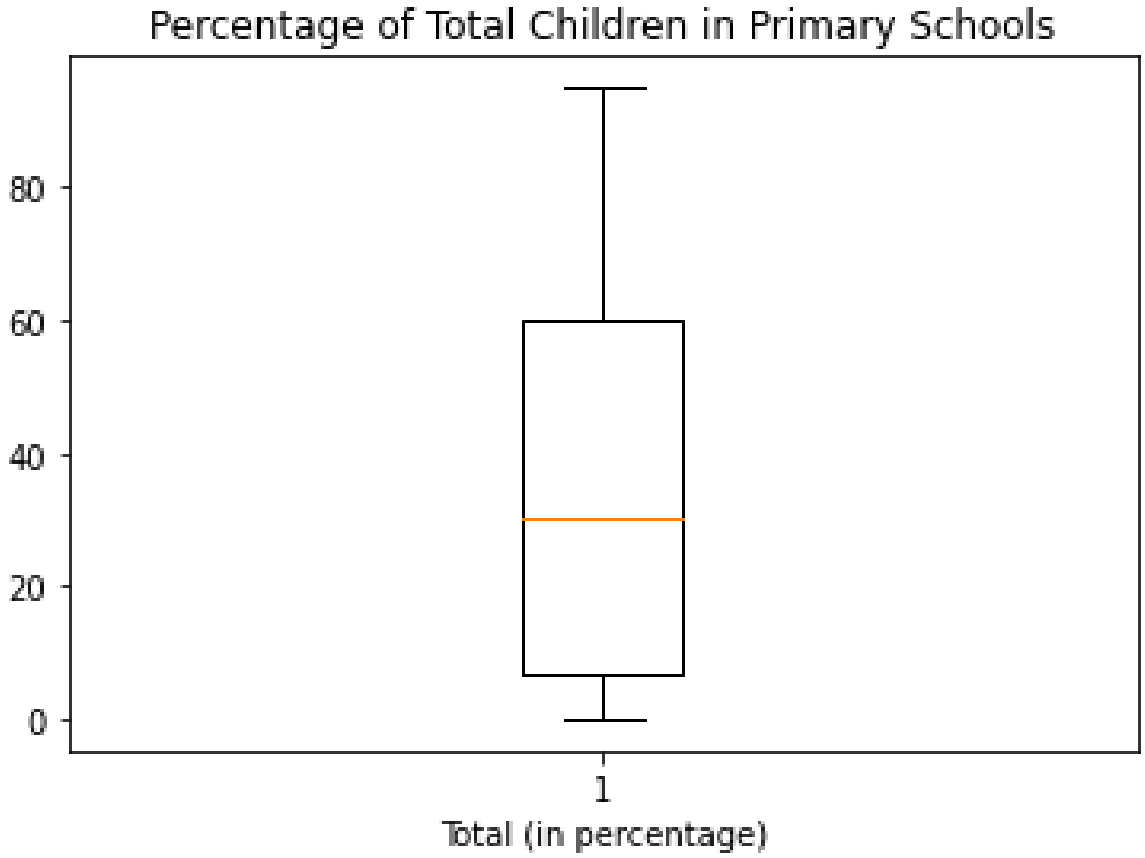
The second section of the code counts how many times each value appears in the Year column, which represents the year the data was collected. The number of countries in each year is then depicted in a bar chart.

The distribution of the Total_% column is displayed in a boxplot by the third section of the code.
The proportion of all children enrolled in primary school is shown here. The code's fourth section builds a scatter matrix to display the relationships between the various numerical columns in the data set.

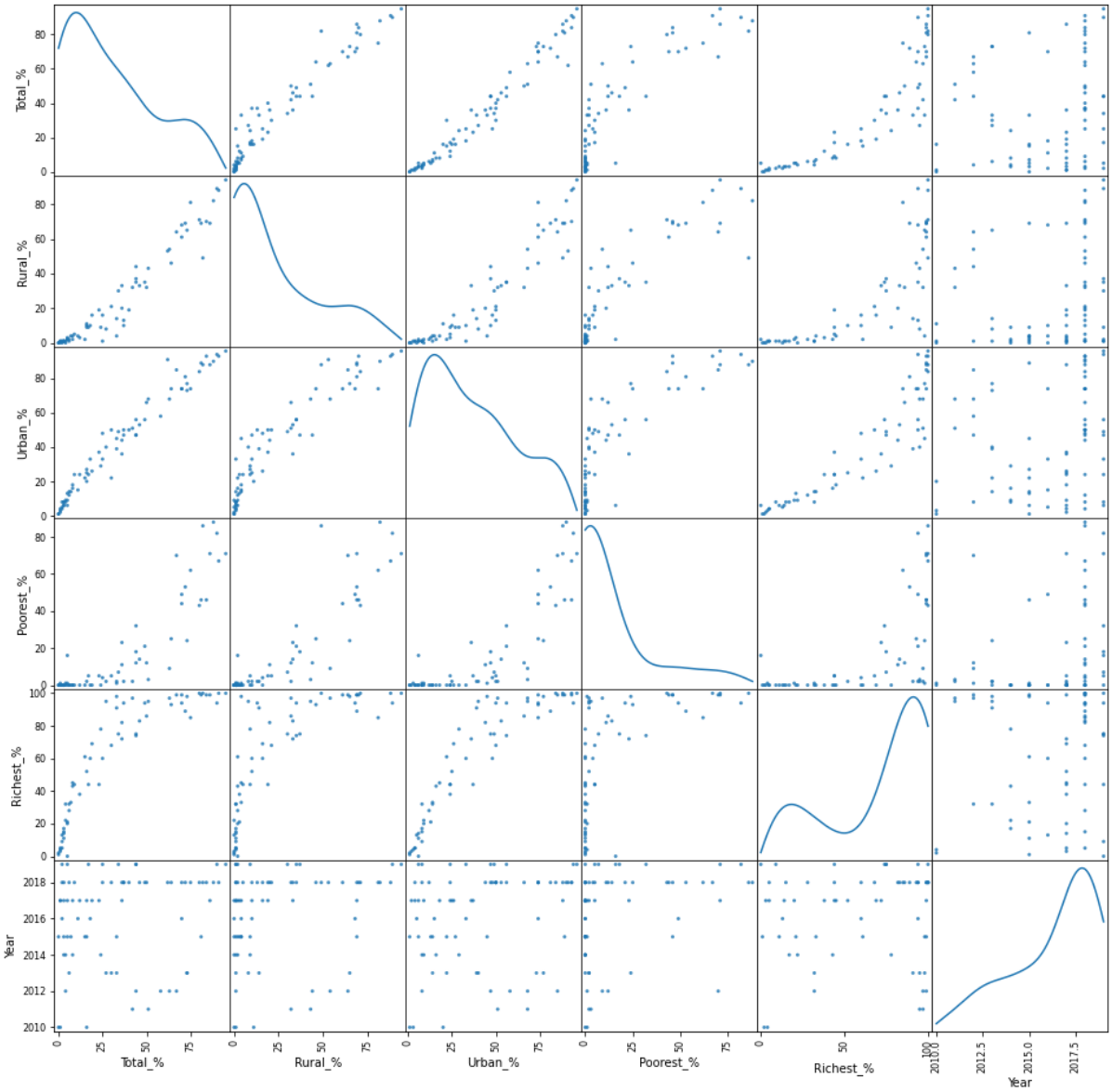
All the graphs for task 2.1 are given below:




Maximum and minimum data gathered
Count
Year
0
5
10
15
20
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019


Percentage of Total Children in Primary Schools
0
20
40
60
80
1
Total (in percentage)

The code offers a straightforward exploratory examination of the dataset, offering information on the distribution of nations by income level, the frequency of data collection, and connections between different dataset elements.



Task 2.2

This code examines a dataset that contains data about several nations. Their socioeconomic status, the proportion of kids in primary school, and the distribution of people living in rural and urban areas.

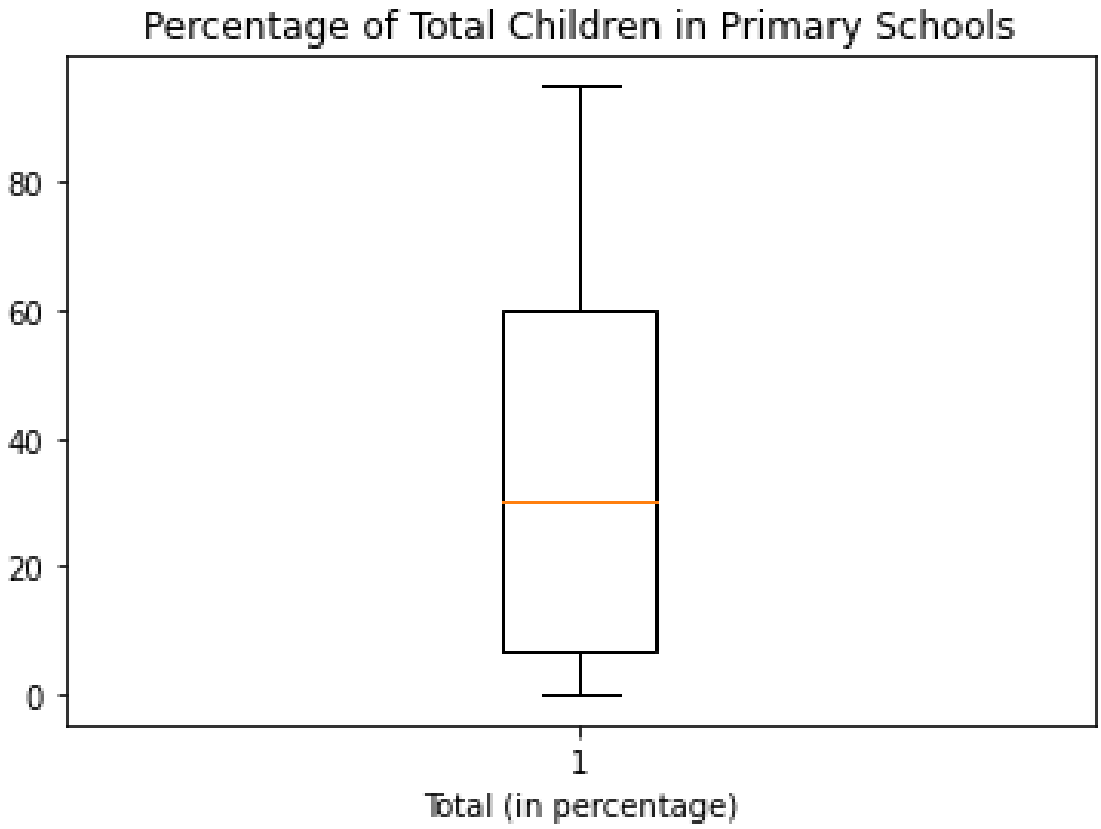
The first section of the code shows the dataset's top 5 rows and renames several columns to make exploring easier.

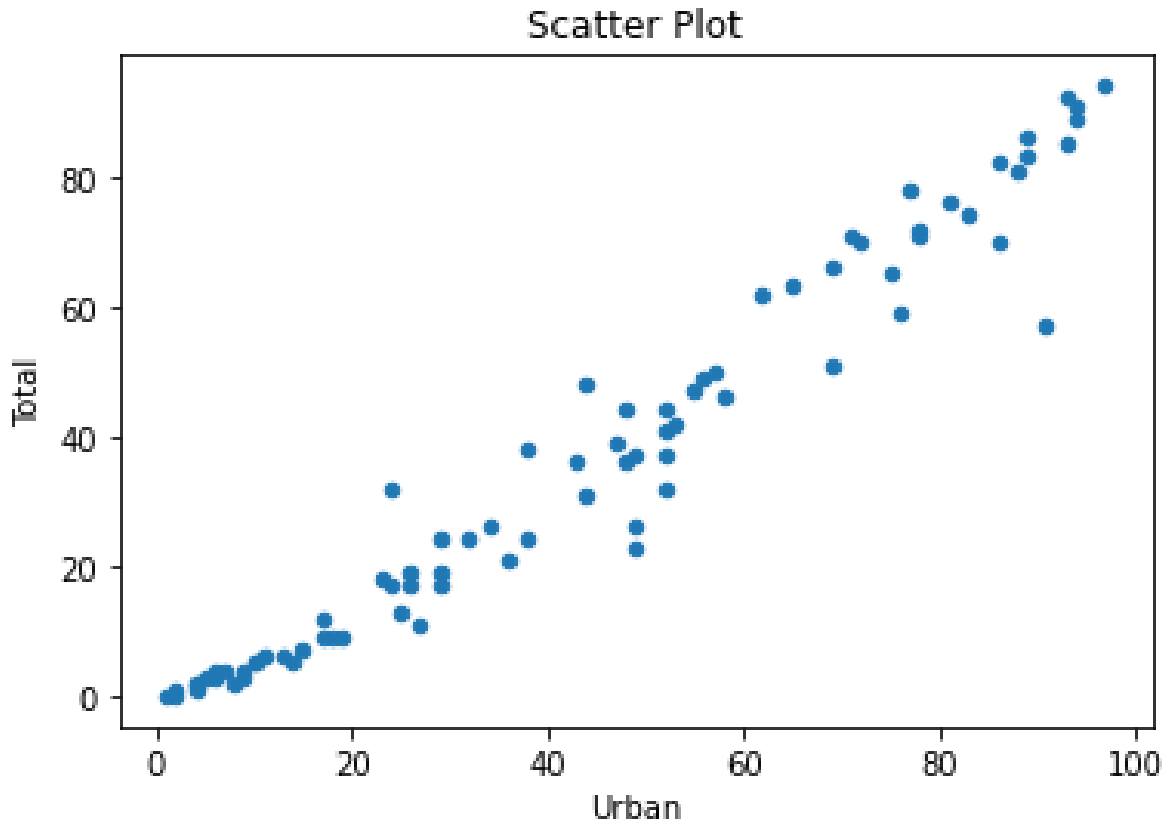
Some columns' data types in the second section of the code were initially represented as strings and have since been changed to floats. Next, use the describe () method to output summary statistics for the dataset's numeric columns.

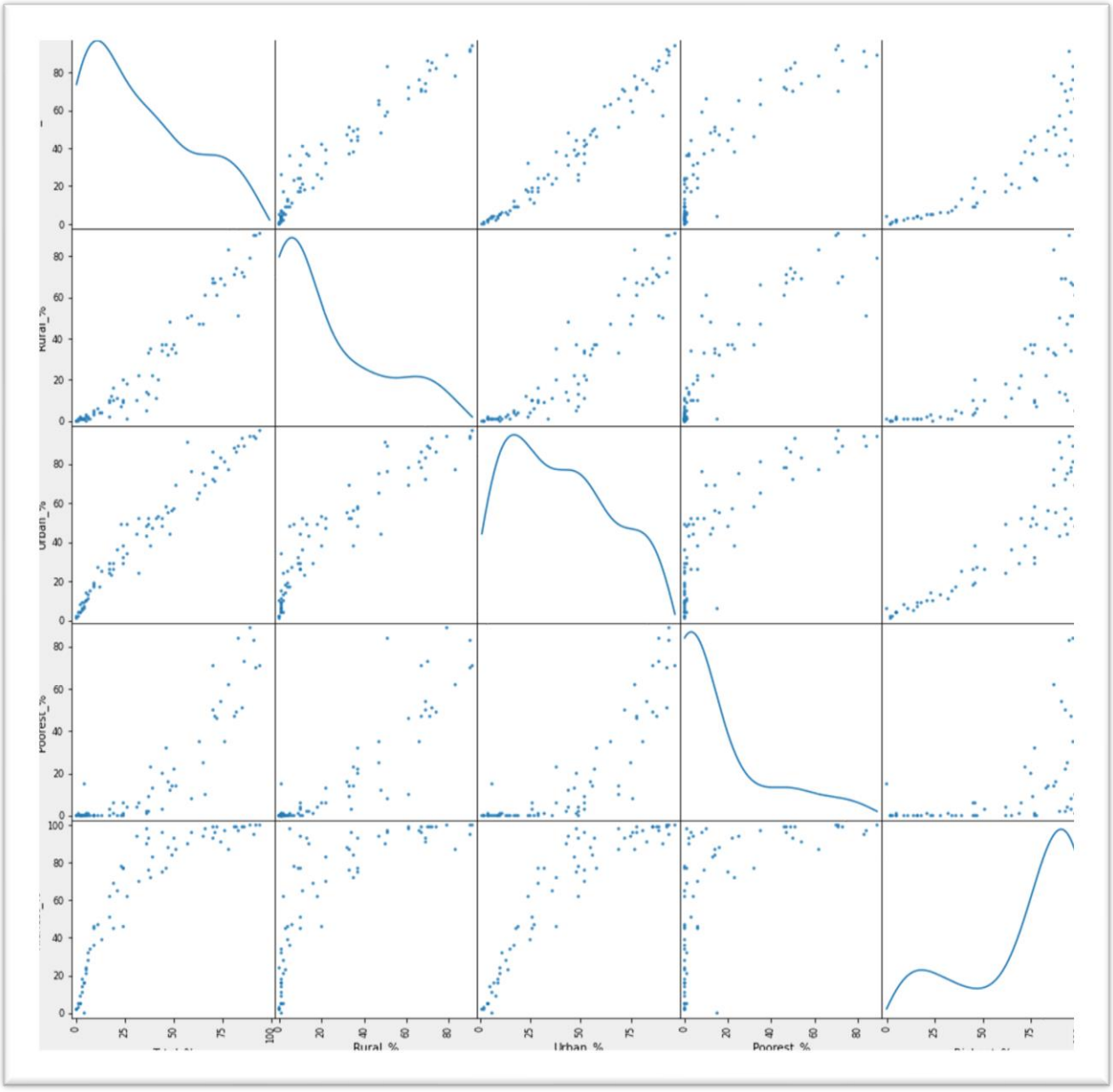
The distribution of the Total_% column is displayed in a boxplot by the third section of the code.
The proportion of all children is shown below after the explanation of rest of task 2.2.

The fourth part of the code creates a scatterplot showing the relationship between the Urban_% and Total_% columns. In the fifth part of the code, a scatter matrix is created to show how the various numerical columns in the data set correlate with one another.

The Total_% column is used in the sixth section to sort the DataFrame data using the sort_values () method in decreasing order. The head () method of the Country column is then used to choose the top 10 nations with the greatest values in the Total_% column. The ten nations with the greatest percentage of children enrolled in primary education are listed as a consequence.







Task 2.3

This task was the easiest of them all and didn't require us to make or plot any graphs. This code uses the Income Group column to filter the data from the Primary and Secondary data frames to only include low- and middle-income nations. Next, determine the typical proportion of elementary and high school children who use the internet at home by using the Total_% column in each filtered data frame. Print the final outcome.