#1 - Which country had the highest number of new leprosy cases in 2022? Why is it not a fair comparison to look at the raw number of cases when comparing prevalence of the disease across different countries?

The country with the highest number of new leprosy cases in 2022 was India. It is not a fair comparison to look at the raw number of cases when comparing the prevalence of the disease across different countries because a country with more people is much more likely to have a higher raw number of cases and vise versa.

#2 - Create a new variable in the dataset that provides the leprosy cases per 100K people in each country. Graph this new variable’s distribution and provide the relevant summary statistics inline within a short paragraph describing the distribution.

A graph of a number of cases

Description automatically generated

The minimum value is 0.00 cases, the Q1 number of cases is 0.00, the median number of cases per 100K is 0.12, the Q3 number of cases per 100K is 1.04, and the max number of cases per 100K is 29.60. The distribution is skewed right and has two upper outliers.

#3 - We want to compare cases per 100K across the different regions in this dataset.  In a single plot output, create a graph that shows the distribution of cases per 100K split by region. Output a nicely formatted table that provides the region name, number of countries in that region, and the median cases per 100K for each region.  Include a short paragraph summarizing differences in leprosy prevalence across regions.

A graph of a distribution of leprosy cases

Description automatically generated

The Americas, Europe, and the middle east generally have the least number of cases per 100k, with the African region having a larger and higher spread and more upper outliers, while APAC has the highest Q1 – Q3 range, demonstrating that they generally have the most cases of leprosy per 100k than any other region.

#4 - Investigate the relationship between cases per 100K and one of the other variables in the dataset (other than region) by making the appropriate bivariate graph and providing the relevant summary statistic inline within a short paragraph describing the relationship.

A graph with black dots

Description automatically generated

There is a weak, negative correlation between the GDP of a country in billions and leprosy cases per 100k people. This demonstrates that these two variables are not very strongly correlated. The correlation coefficient is -0.022.

#5 - Write a brief conclusion to your analysis summarizing what you found. Include a hyperlink to the website for the International Leprosy Association for readers wanting more information about this disease.

Although it would be thought that like most bacterial diseases, leprosy would be significantly more prevalent in poor economic regions such as South America, Africa, and the Middle East, it does not directly connect with economic status since those regions are similar in economic status as those with the highest number of leprosy cases. Countries with more people are more likely to have higher raw numbers of leprosy cases, notably India who has the highest population and the highest number of leprosy cases. Raw numbers of cases should not be used to calculate prevalence of leprosy in a region because it doesn’t account from the fact that countries with higher populations have more people who can be infected with the disease. You can read more about leprosy here: <https://leprosyhistory.org/>