Business Challenge: EDA and SQL

Dan N. Bigman-Montalvo

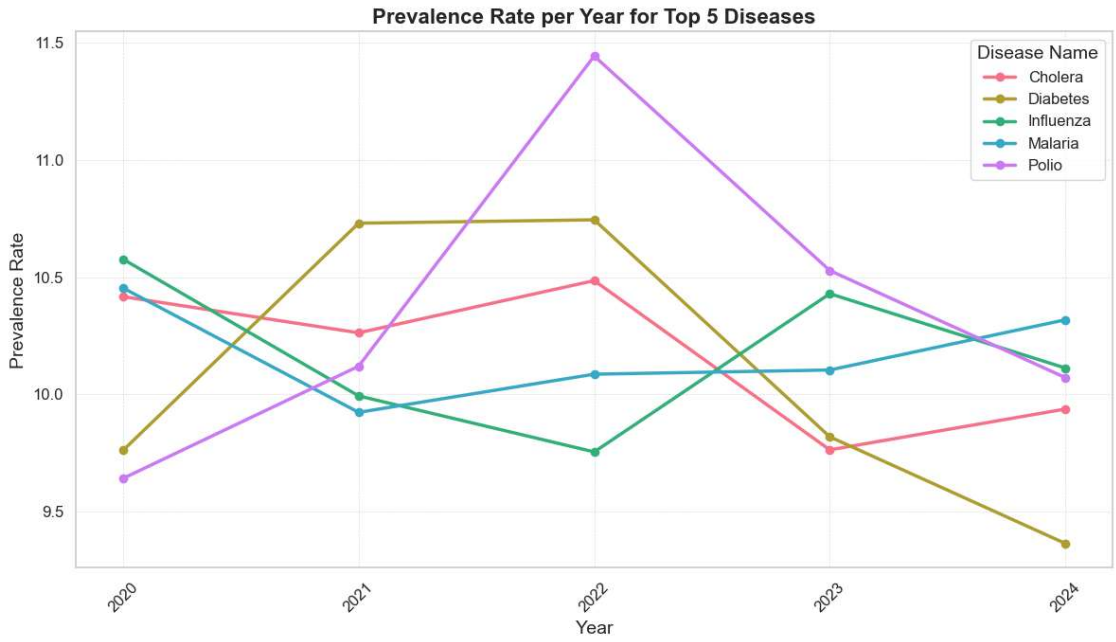
Ironhack DDEC Cohort

Global Health Statistics

- I selected the “Global Health Statistics” dataset from Kaggle
- This dataset is fully synthetic. Any patterns, trends, or conclusions drawn apply only to this generated data rather than real-world conditions.

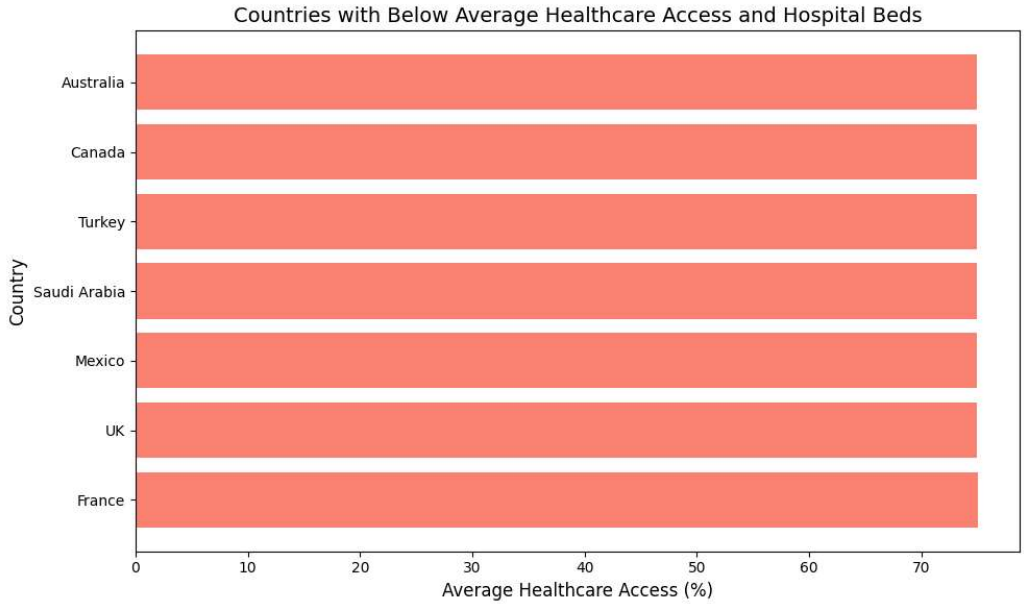
Which infectious diseases have the highest prevalence rates globally, and how have these rates changed over the past 5 years?

DiseaseName	Avg_Prevalence
Polio	10.35
Cholera	10.18
Malaria	10.18
Influenza	10.16
Diabetes	10.10
Dengue	10.08
Zika	10.08
Measles	10.07
COVID-19	10.03
Leprosy	10.00

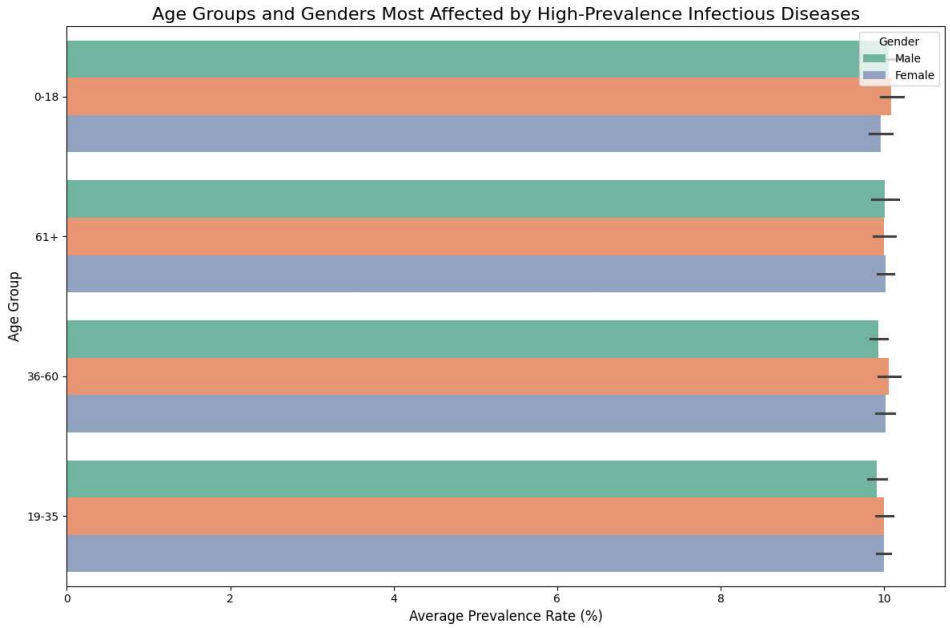


What countries have the worst access to healthcare?

Country	AvgHealthcareAccess	AvgHospitalBeds
Australia	74.880671	5.228248
Canada	74.899341	5.230515
Turkey	74.908387	5.243976
Saudi Arabia	74.933178	5.237754
Mexico	74.938198	5.24138
UK	74.966663	5.235328
France	74.977308	5.238784



Which age groups and genders are most affected by high-prevalence infectious diseases? Are there significant disparities?



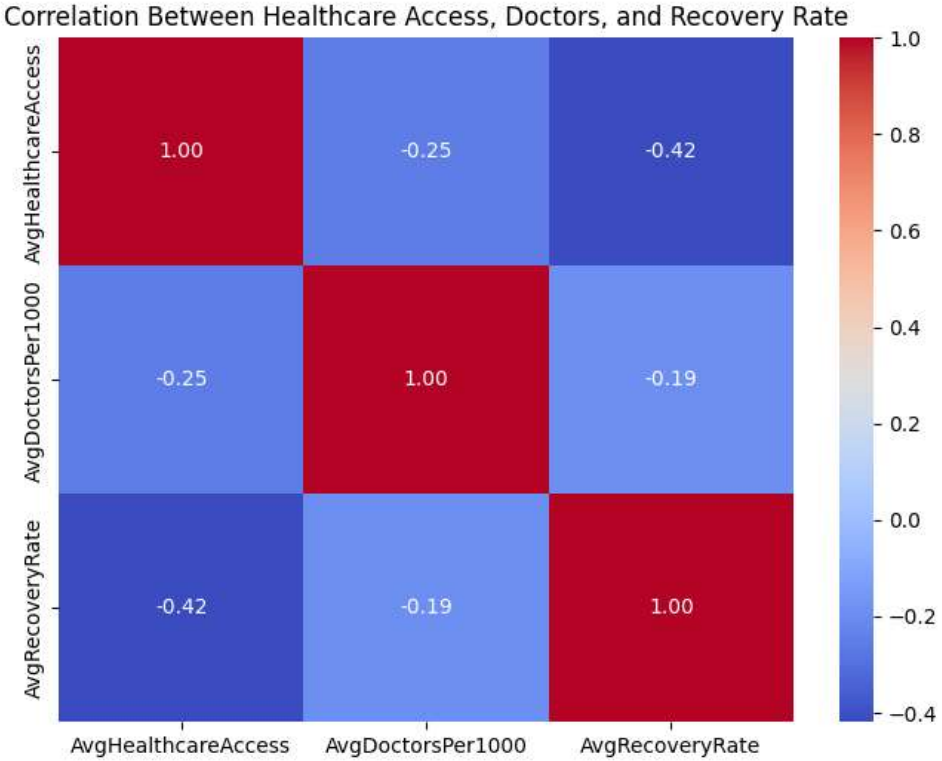
Prevalence Across Age Groups:
The prevalence rates are consistent across all age groups (0-18, 19-35, 36-60, 61+).No significant spikes or drops are observed, indicating that high-prevalence infectious diseases affect all age groups uniformly.

Gender Differences:
Prevalence rates for males and females are nearly identical across all age groups. There are no notable gender disparities, suggesting both genders are similarly affected by high-prevalence infectious diseases.

Error Bars:
Minimal variation in the error bars further confirms the consistency of average prevalence rates across age groups and genders.

Equal Distribution: The data shows that infectious diseases in the dataset are distributed equally across all age groups and genders.

Is there a correlation between healthcare access, the number of doctors per 1000 people, and the recovery rate for specific diseases?

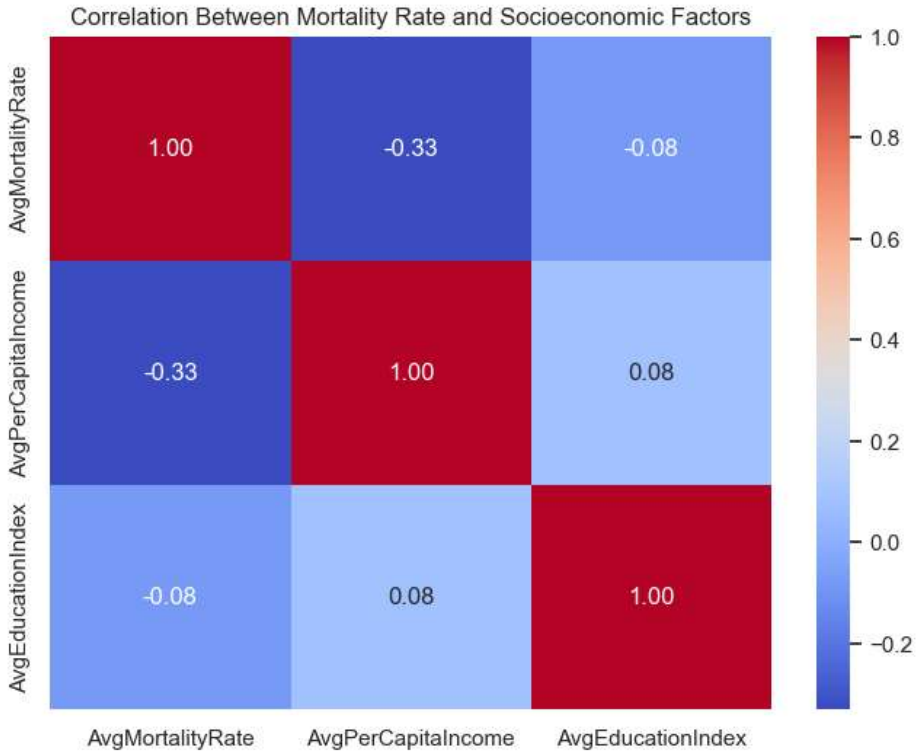


AvgHealthcareAccess and AvgDoctorsPer1000: **Correlation coefficient: -0.25**
Indicates a weak negative correlation. This suggests that as healthcare access improves, the number of doctors per 1000 people decreases slightly. This could reflect errors in measuring healthcare access versus medical workforce distribution.

AvgHealthcareAccess and AvgRecoveryRate: **Correlation coefficient: -0.42**
Indicates a moderate negative correlation. Surprisingly, better healthcare access correlates with lower recovery rates. This counterintuitive result may point to data limitation in measuring healthcare outcomes.

AvgDoctorsPer1000 and AvgRecoveryRate: **Correlation coefficient: -0.19**
This reflects a weak negative correlation, suggesting that increasing doctors per 1000 people does not strongly correlate with higher recovery rates. It may hint at external influences like disease severity or broader systemic health challenges.

Which diseases have the highest mortality rates, and how do socioeconomic factors influence these rates?



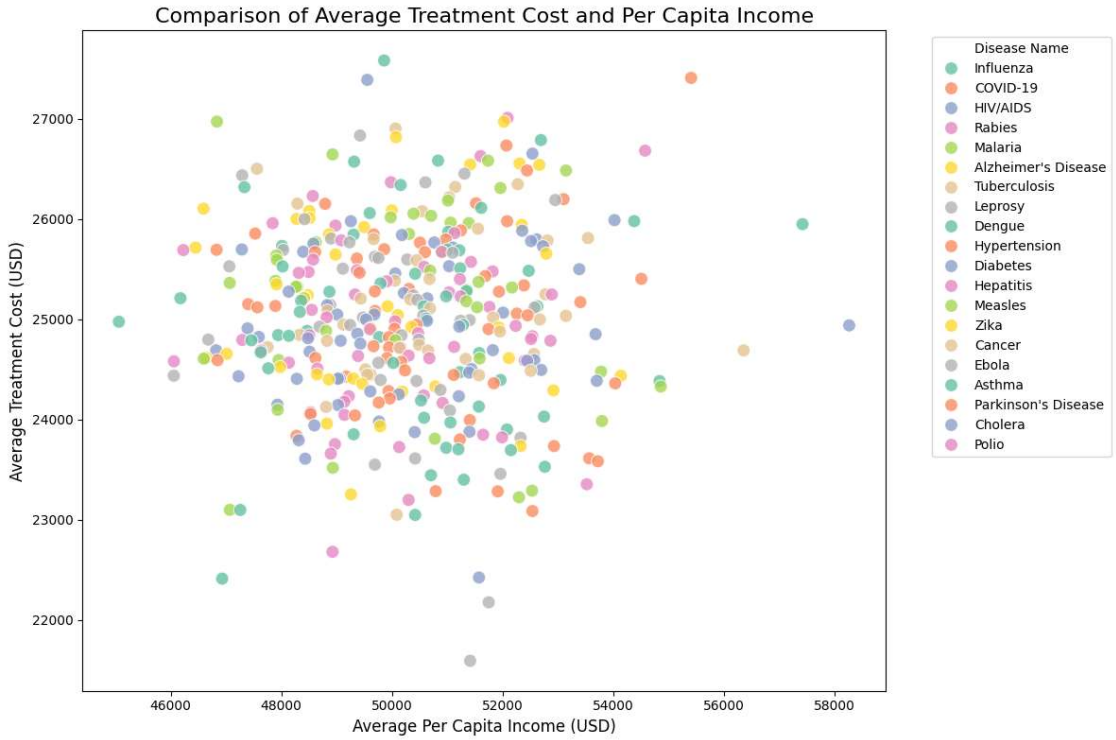
AvgMortalityRate and AvgPerCapitaIncome: Correlation coefficient: **-0.33**
This indicates a weak to moderate negative correlation, suggesting that higher per capita income is associated with lower mortality rates. This aligns with the expectation that increased income levels provide better access to healthcare and improved living conditions.

AvgMortalityRate and AvgEducationIndex: Correlation coefficient: **-0.08**
The very weak negative correlation implies that education levels have a minimal direct impact on mortality rates in this dataset. This could indicate that other factors, such as healthcare access or disease prevalence, might be more influential.

AvgPerCapitaIncome and AvgEducationIndex: Correlation coefficient: **0.08**
A weak positive correlation indicates a slight association between higher education levels and increased per capita income.

The most notable relationship is the weak to moderate negative correlation between **AvgMortalityRate** and **AvgPerCapitaIncome**, highlighting the role of income in reducing mortality rates. **AvgEducationIndex** has minimal impact on income and mortality, suggesting other factors might play a more significant role in the dataset.

What is the average treatment cost (USD) for the most common infectious diseases, and how does it compare to the per capita income in different countries?



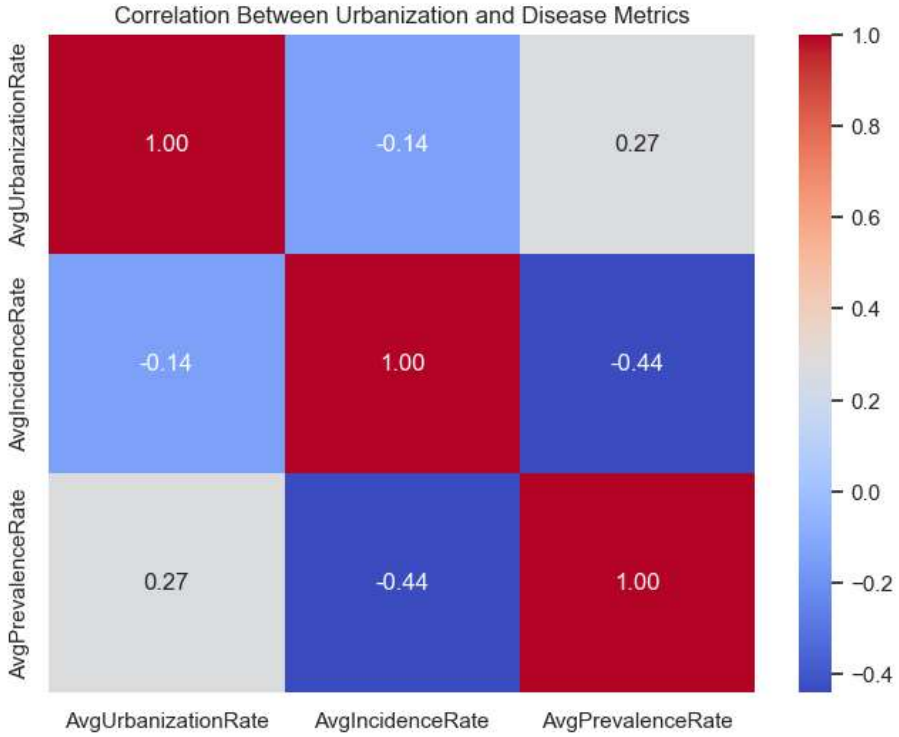
Distribution of Points:
Each point represents a country-disease combination. The horizontal axis shows per capita income, and the vertical axis represents treatment costs.

No Clear Correlation: No strong linear relationship exists between income and treatment costs; the costs remain relatively consistent across different income levels.

Clustered Costs: Most treatment costs range from $22,000 to $27,000 USD.

Higher Income Outliers: Countries with per capita incomes above $55,000 show more significant variability in treatment costs, possibly due to differences in healthcare infrastructure or treatment approaches.

Does the urbanization rate affect the incidence and prevalence rates of infectious diseases? Are urban areas more vulnerable to certain outbreaks?

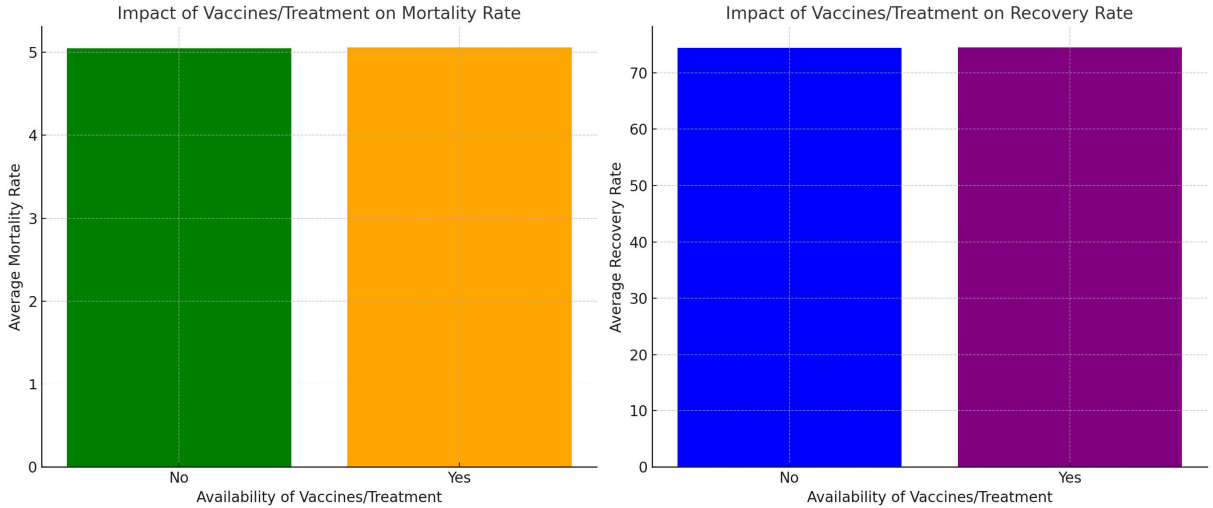


Urbanization demonstrates a mixed impact on infectious disease dynamics. The correlation coefficient between urbanization and incidence rates (-0.14) suggests a weak negative relationship, implying that urbanization slightly decreases new infections.

This effect is minimal and likely influenced by other factors. On the other hand, urbanization and prevalence rates show a weak positive correlation (0.27), indicating that urban areas might sustain a higher prevalence of diseases over time due to factors like chronic infections or persistent environmental conditions.

Urban areas do not appear significantly more vulnerable to the initial spread of outbreaks, as suggested by the weak negative correlation with incidence rates. However, the weak positive correlation with prevalence rates highlights urban vulnerabilities in maintaining ongoing cases.

How does the availability of vaccines or treatments impact the mortality and recovery rates for specific infectious diseases?

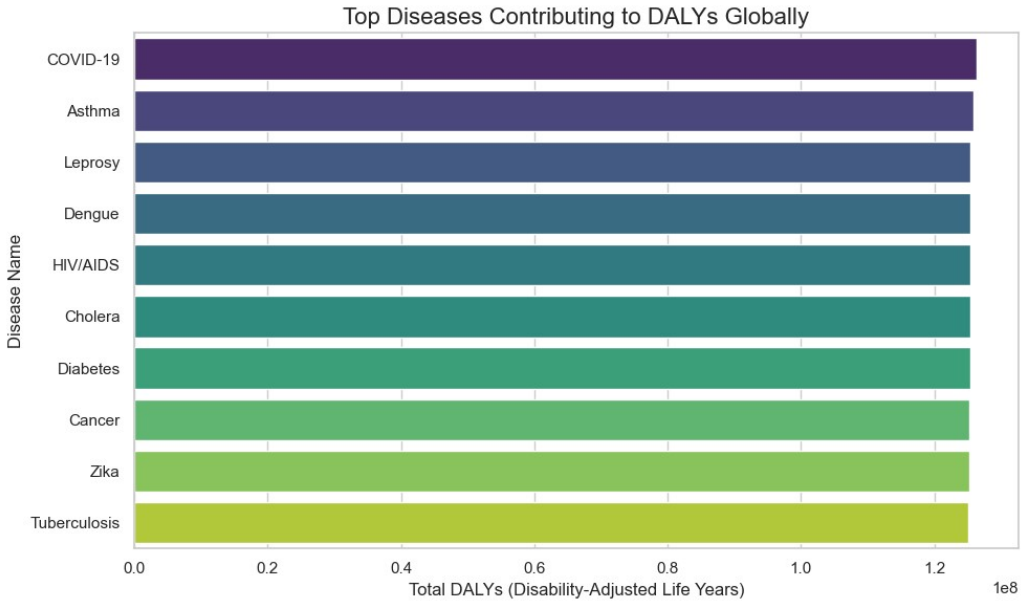


The analysis reveals that diseases with no available vaccines/treatments have an average mortality rate of 5.047, while those with available vaccines/treatments have a slightly higher rate of 5.059. This negligible difference suggests that vaccines and treatments might not directly impact mortality in this dataset and that other factors, such as disease severity or healthcare access, could play a more significant role.

For recovery rates, diseases with no vaccines/treatments show an average recovery rate of 74.37%, compared to 74.50% for those with vaccines/treatments. While the improvement is modest, it indicates a potential marginal benefit from the availability of treatments or vaccines.

The data suggests that the presence of vaccines or treatments alone may not be sufficient to drive significant differences in outcomes, as healthcare infrastructure, early diagnosis, and treatment implementation are likely critical. Additionally, the dataset may include diseases with inherently high baseline mortality or recovery rates, which could overshadow the observable impact of vaccines and treatments.

Which diseases contribute the most to DALYs (Disability-Adjusted Life Years)?



COVID-19 ranks highest, reflecting its global health impact over recent years.

Chronic and infectious diseases such as **Asthma**, **Leprosy**, and **HIV/AIDS** also have significant contributions, emphasizing their long-term impact on health.

Cholera and **Dengue** highlight the persistent burden of preventable and treatable diseases, especially in regions with inadequate healthcare infrastructure.

Which countries have the highest disease burden?