Business Challenge: EDA and SQL

Introduction

For this business challenge, I sought to use a substantial dataset that would truly test the limits of Python and SQL. After exploring various options, I selected the “Global Health Statistics” dataset from Kaggle because of its size and complexity. However, it is important to note that this dataset is fully synthetic, meaning any patterns, trends, or conclusions drawn apply only to this generated data rather than real-world conditions.

I. Database source

Data card was downloaded from
https://www.kaggle.com/datasets/malaiarasugraj/global-health-statistics

10.34740/kaggle/dsv/10028650

Provenance

The dataset is synthetic, and while it is based on real-world global health trends, no direct data from these sources has been used. Instead, values have been generated based on commonly available health statistics and epidemiological patterns.

Collection Methodology

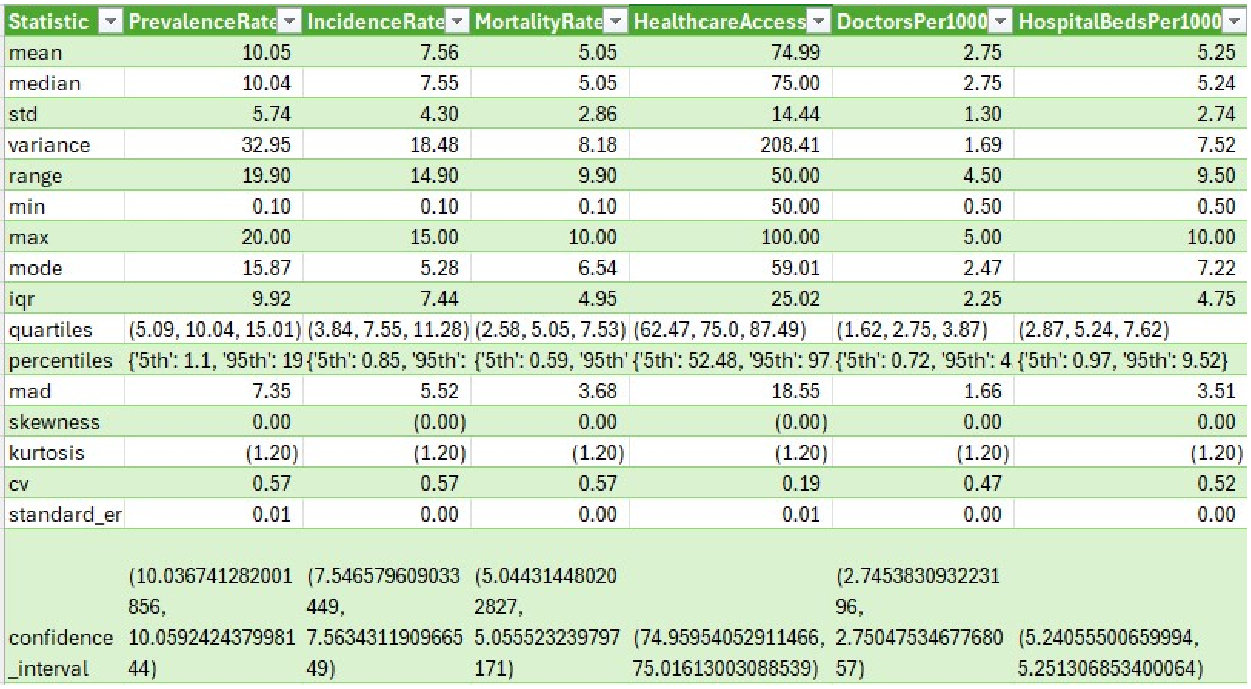
Data Generation, Randomization, Geographical Representation, Health Trends

Attribute	Description	Data Type
Country	The name of the country where the health data was recorded.	Categorical
Year	The year in which the data was collected.	Numerical - Discrete
Disease Name	The name of the disease or health condition tracked.	Categorical
Disease Category	The category of the disease (e.g., Infectious, Non-Communicable).	Categorical
Prevalence Rate (%)	The percentage of the population affected by the disease.	Numerical - Continuous
Incidence Rate (%)	The percentage of new or newly diagnosed cases.	Numerical - Continuous
Mortality Rate (%)	The percentage of the affected population that dies from the disease.	Numerical - Continuous
Age Group	The age range most affected by the disease.	Ordinal
Gender	The gender(s) affected by the disease (Male, Female, Both).	Categorical
Population Affected	The total number of individuals affected by the disease.	Numerical - Discrete
Healthcare Access (%)	The percentage of the population with access to healthcare.	Numerical - Continuous
Doctors per 1000	The number of doctors per 1000 people.	Numerical - Continuous

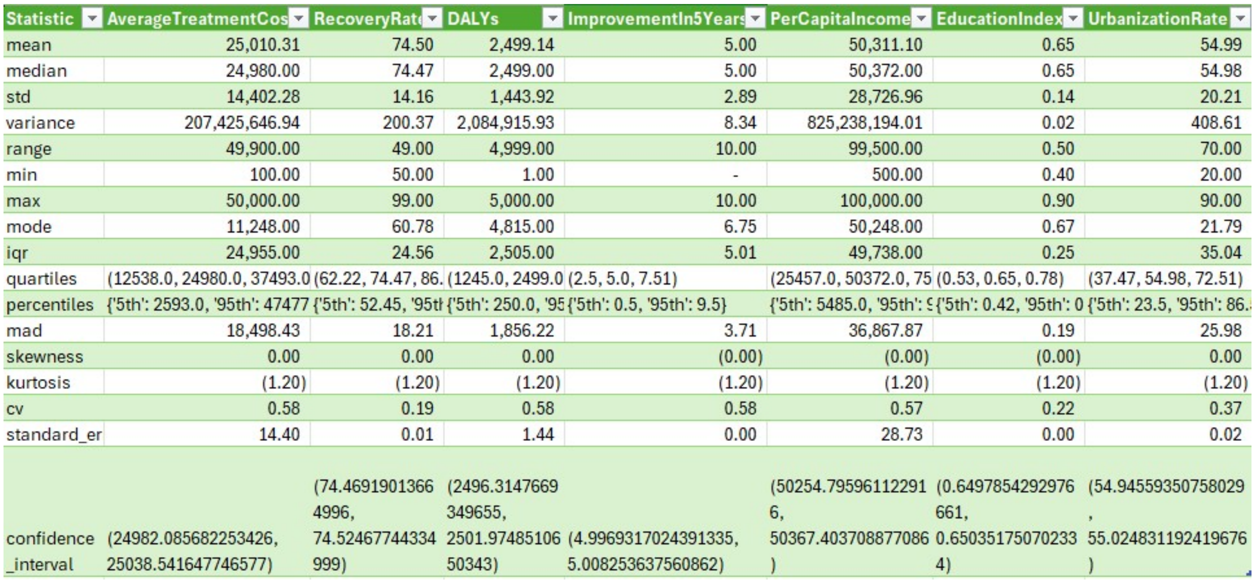
Hospital Beds per 1000	The number of hospital beds available per 1000 people.	Numerical - Continuous
Treatment Type	The primary treatment method for the disease (e.g., Medication, Surgery).	Categorical
Average Treatment Cost (USD)	The average cost of treating the disease in USD.	Numerical - Continuous
Availability of Vaccines/Treatment	Whether vaccines or treatments are available.	Categorical
Recovery Rate (%)	The percentage of people who recover from the disease.	Numerical - Continuous
DALYs	Disability-Adjusted Life Years, a measure of disease burden.	Numerical - Continuous
Improvement in 5 Years (%)	The improvement in disease outcomes over the last five years.	Numerical - Continuous
Per Capita Income (USD)	The average income per person in the country.	Numerical - Continuous
Education Index	The average level of education in the country.	Numerical - Continuous
Urbanization Rate (%)	The percentage of the population living in urban areas.	Numerical - Continuous

Distinct categorical values for DiseaseCategory	Count
Respiratory	90,588
Parasitic	91,178
Genetic	91,153
Autoimmune	91,153
Bacterial	90,509
Cardiovascular	90,968
Neurological	91,000
Chronic	90,445
Metabolic	91,332
Infectious	90,764
Viral	90,910
None	2

Null Values Per Column	Null Count
DoctorsPer1000	2
HospitalBedsPer1000	2
TreatmentType	2
AverageTreatmentCost	2
AvailabilityOfVaccinesTreatment	2



Statistic	PrevalenceRate	IncidenceRate	MortalityRate	HealthcareAccess	DoctorsPer1000	HospitalBedsPer1000
mean	10.05	7.56	5.05	74.99	2.75	5.25
median	10.04	7.55	5.05	75.00	2.75	5.24
std	5.74	4.30	2.86	14.44	1.30	2.74
variance	32.95	18.48	8.18	208.41	1.69	7.52
range	19.90	14.90	9.90	50.00	4.50	9.50
min	0.10	0.10	0.10	50.00	0.50	0.50
max	20.00	15.00	10.00	100.00	5.00	10.00
mode	15.87	5.28	6.54	59.01	2.47	7.22
iqr	9.92	7.44	4.95	25.02	2.25	4.75
quartiles	(5.09, 10.04, 15.01)	(3.84, 7.55, 11.28)	(2.58, 5.05, 7.53)	(62.47, 75.0, 87.49)	(1.62, 2.75, 3.87)	(2.87, 5.24, 7.62)
percentiles	{'5th': 1.1, '95th': 19	{'5th': 0.85, '95th':	{'5th': 0.59, '95th'	{'5th': 52.48, '95th': 97	{'5th': 0.72, '95th': 4	{'5th': 0.97, '95th': 9.52}
mad	7.35	5.52	3.68	18.55	1.66	3.51
skewness	0.00	(0.00)	0.00	(0.00)	0.00	0.00
kurtosis	(1.20)	(1.20)	(1.20)	(1.20)	(1.20)	(1.20)
cv	0.57	0.57	0.57	0.19	0.47	0.52
standard_er	0.01	0.00	0.00	0.01	0.00	0.00
confidence_interval	(10.036741282001856, 10.059242437998144)	(7.546579609033449, 7.563431190966549)	(5.044314480202827, 5.055523239797171)	(74.95954052911466, 75.01613003088539)	(2.745383093223196, 2.750475346776805 7)	(5.24055500659994, 5.251306853400064)



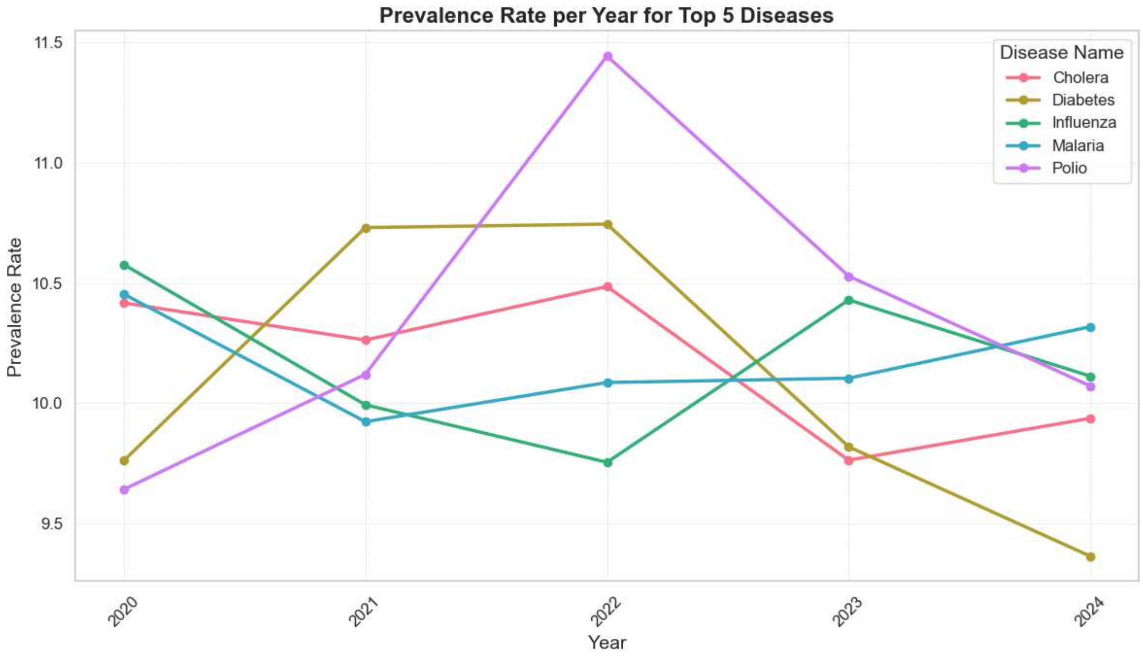
Statistic	AverageTreatmentCos	RecoveryRat	DALYs	ImprovementIn5Years	PerCapitaIncome	EducationIndex	UrbanizationRate
mean	25,010.31	74.50	2,499.14	5.00	50,311.10	0.65	54.99
median	24,980.00	74.47	2,499.00	5.00	50,372.00	0.65	54.98
std	14,402.28	14.16	1,443.92	2.89	28,726.96	0.14	20.21
variance	207,425,646.94	200.37	2,084,915.93	8.34	825,238,194.01	0.02	408.61
range	49,900.00	49.00	4,999.00	10.00	99,500.00	0.50	70.00
min	100.00	50.00	1.00	-	500.00	0.40	20.00
max	50,000.00	99.00	5,000.00	10.00	100,000.00	0.90	90.00
mode	11,248.00	60.78	4,815.00	6.75	50,248.00	0.67	21.79
iqr	24,955.00	24.56	2,505.00	5.01	49,738.00	0.25	35.04
quartiles	(12538.0, 24980.0, 37493.0	(62.22, 74.47, 86.	(1245.0, 2499.0	(2.5, 5.0, 7.51)	(25457.0, 50372.0, 75	(0.53, 0.65, 0.78)	(37.47, 54.98, 72.51)
percentiles	{'5th': 2593.0, '95th': 47477	{'5th': 52.45, '95th	{'5th': 250.0, '95	{'5th': 0.5, '95th': 9.5}	{'5th': 5485.0, '95th': 9	{'5th': 0.42, '95th': 0	{'5th': 23.5, '95th': 86.
mad	18,498.43	18.21	1,856.22	3.71	36,867.87	0.19	25.98
skewness	0.00	0.00	0.00	(0.00)	(0.00)	(0.00)	0.00
kurtosis	(1.20)	(1.20)	(1.20)	(1.20)	(1.20)	(1.20)	(1.20)
cv	0.58	0.19	0.58	0.58	0.57	0.22	0.37
standard_er	14.40	0.01	1.44	0.00	28.73	0.00	0.02
confidence_interval	(24982.085682253426, 25038.541647746577)	(74.46919013664996, 74.52467744334999)	(2496.3147669349655, 2501.9748510650343)	(4.9969317024391335, 5.008253637560862)	(50254.795961122916, 50367.403708877086)	(0.6497854292976661, 0.650351750702334)	(54.94559350758029, 55.024831192419676)

Business Challenge: EDA and SQL

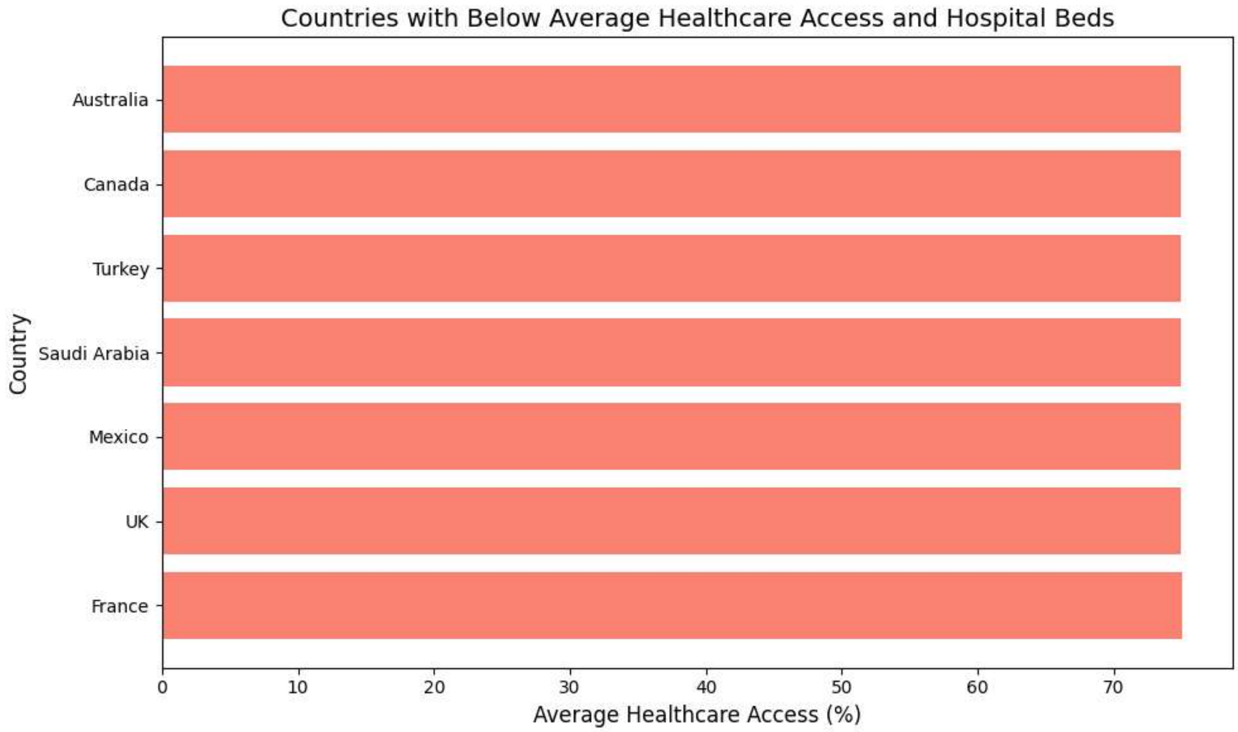
1. Which infectious diseases have the highest prevalence rates globally, and how have these rates changed over the past 5 years?

Top Infectious Diseases:

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

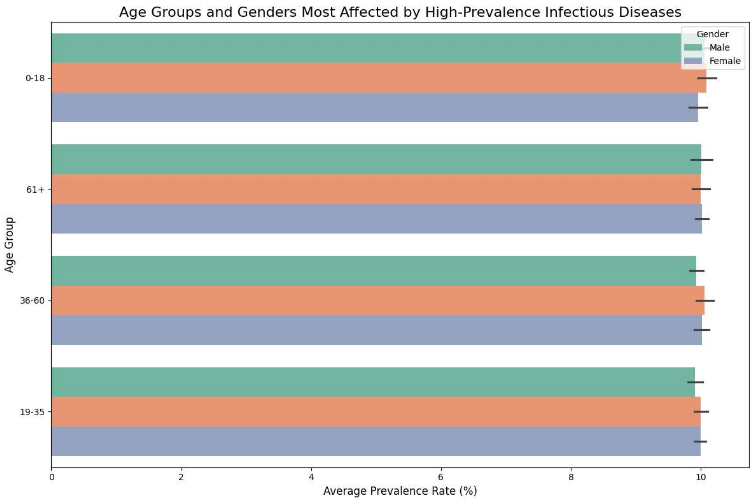


2. What countries have the worst access to healthcare?



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

3. 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.

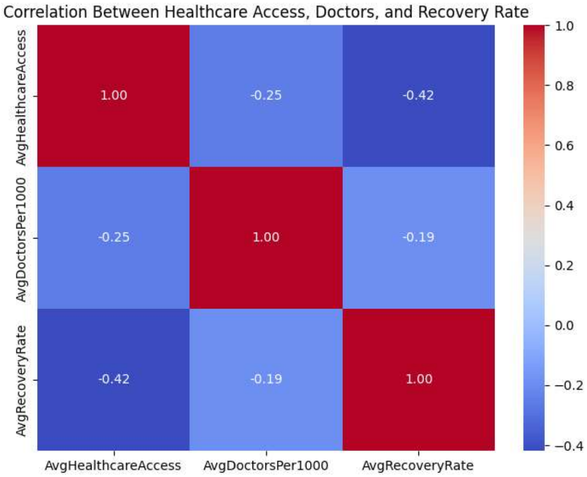
Error Bars:

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.

4. Healthcare System Correlation:

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

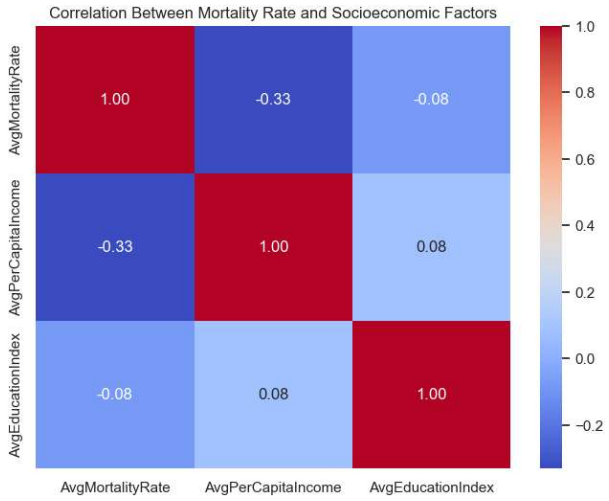
Indicates a moderate negative correlation. Surprisingly, better healthcare access correlates with lower recovery rates. This counterintuitive result may point to data limitation in measuring healthcare outcomes.

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.

5. Mortality Factors:

Which diseases have the highest mortality rates, and how do socioeconomic factors (e.g., per capita income, education index) 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.

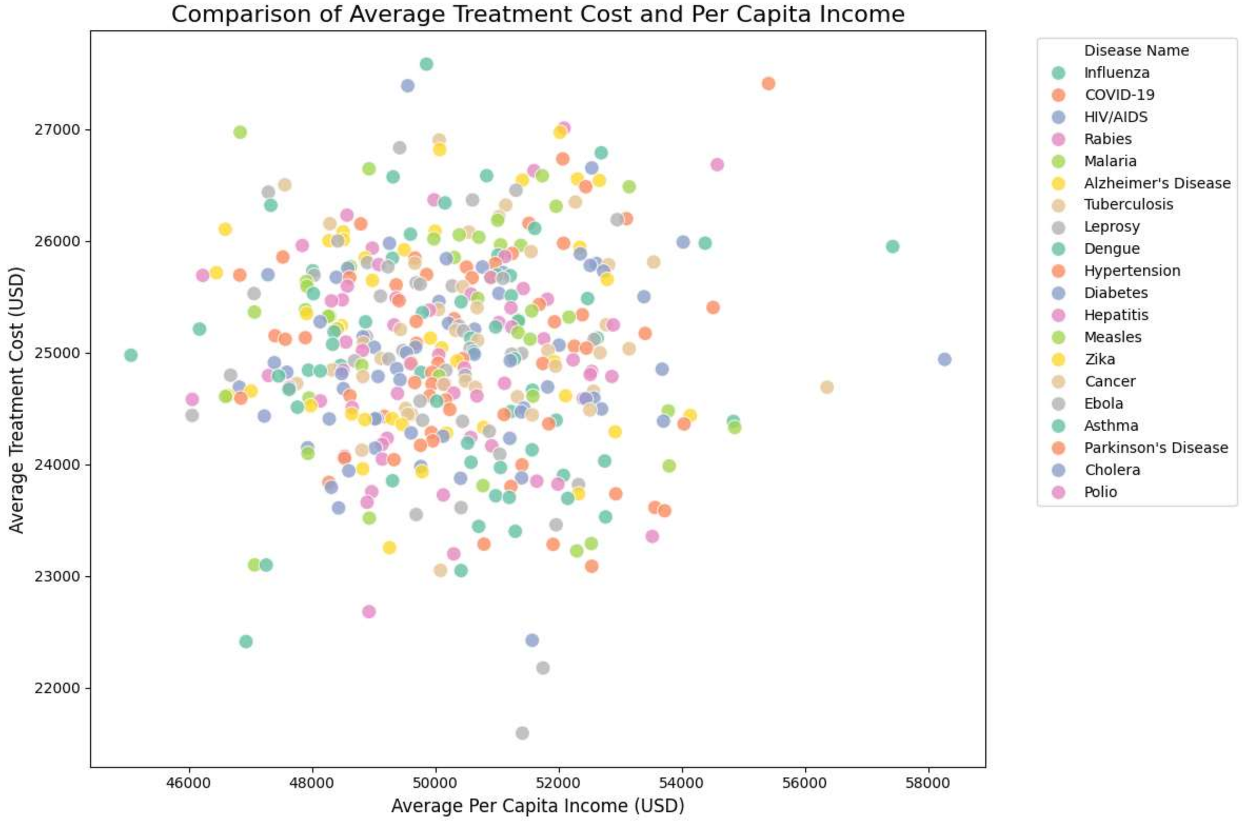
AvgPerCapitaIncome and AvgEducationIndex: Correlation coefficient: 0.08

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.

6. Economic Burden:

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.

Key Insights:

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.

Disease-Specific Trends:

Diseases like **COVID-19**, **HIV/AIDS**, and **Asthma** exhibit broad variation in treatment costs across countries. Diseases like **Polio** and **Measles** have more tightly grouped costs, suggesting standardized treatment pricing.

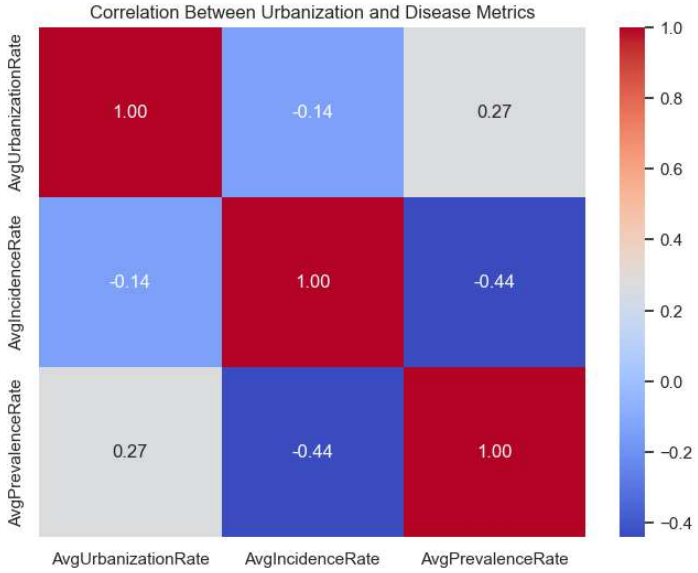
Global Consistency: Many diseases exhibit globally consistent treatment costs, likely driven by standardized healthcare policies.

Income-Independent Costs: Treatment costs appear to depend more on disease-specific factors than on local economic conditions.

Variability in High-Income Countries: Greater treatment cost variation in wealthy nations may reflect differences in healthcare access and approaches.

7. Urbanization & Disease Spread:

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



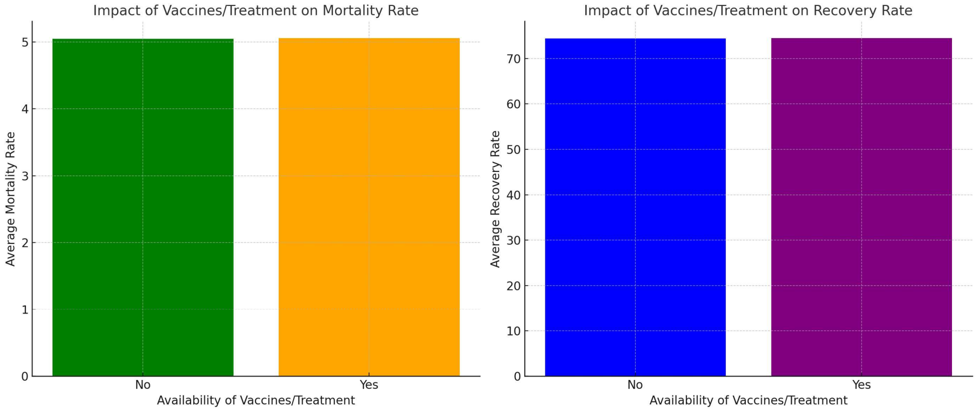
Urbanization demonstrates a mixed impact on infectious disease dynamics. The correlation coefficient between urbanization and incidence rates (-0.14) suggests a weak negative relationship, implying that urbanization slightly decreases new infections. However, this effect is minimal and likely influenced 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. Factors contributing to this include crowded living conditions that hinder containment efforts, pollution, and lifestyle-related risks, and insufficient healthcare infrastructure in underdeveloped urban areas, leading to prolonged disease durations.

In summary, urbanization has a negligible effect on new infections but contributes to the persistence of diseases, emphasizing the need for targeted strategies to address urban-specific challenges. Investments in healthcare infrastructure improved living conditions, and access to healthcare in urban settings can mitigate long-term disease prevalence while maintaining preparedness for new outbreaks.

8. Vaccine/Treatment Availability:

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



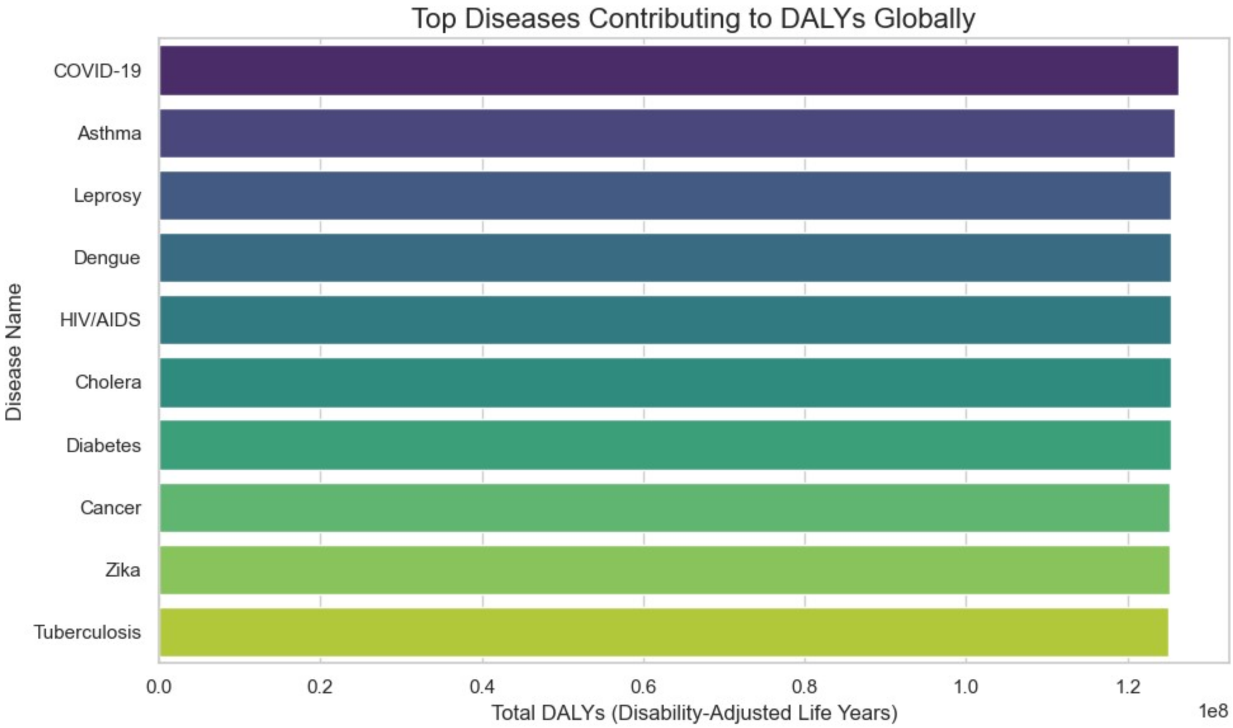
The analysis reveals that diseases with no available vaccines/treatments have an average mortality rate of 5.047, while those with available vaccines/treatments have a slightly higher rate of 5.059. This negligible difference suggests that vaccines and treatments might not directly impact mortality in 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.

9. **Disability and Life Impact:**

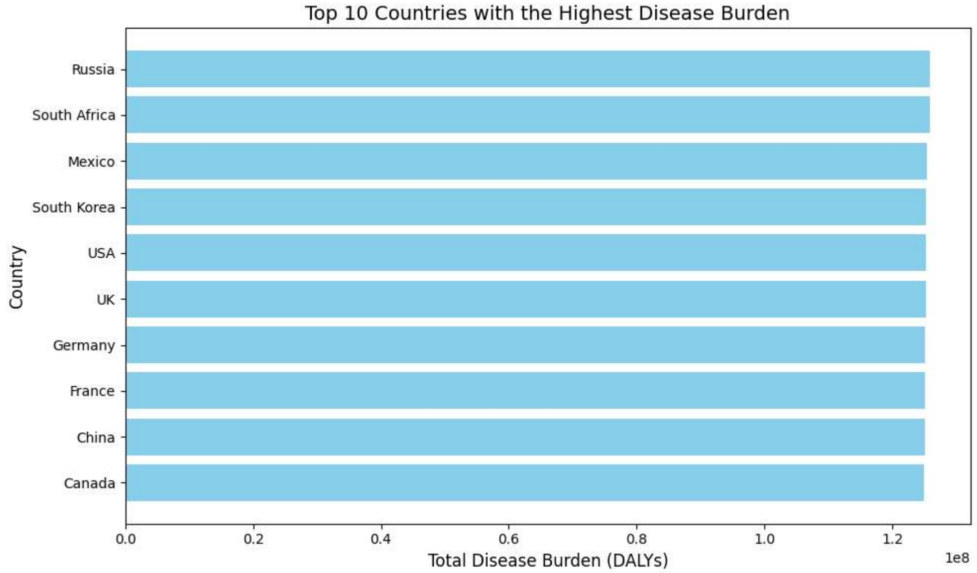
Which diseases contribute the most to DALYs (Disability-Adjusted Life Years)?



Insights:

- **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.

10. What are the countries with the highest disease burden?


Top 10 Countries with the Highest Disease Burden
Country
Russia
South Africa
Mexico
South Korea
USA
UK
Germany
France
China
Canada
1e8
Total Disease Burden (DALYs)