



Investigating the Relationship Between Health Expenditure and Disease Prevalence

DSA 210 Introduction to Data Science (Spring 2024-2025)

Toprak Babacan 31156

Instructor: Özgür Asar

Assistant: Berke Odacı

29.05.2025

Table of Contents

<i>Motivation</i>	3
<i>Hypothesis Testing:</i>	4
<i>Model Training</i>	5
<i>Hyperparameter Tuning</i>	6
<i>Feature Importance Analysis</i>	6
<i>Conclusion</i>	7
<i>Shortcomings and Future Research</i>	8
<i>Use of ChatGPT During the Project</i>	8
<i>Appendix</i>	9
<i>References</i>	14

Motivation

The aim of this research is to investigate the question of whether government health spending does have a measurable impact on the prevalence of vaccine-preventable infectious diseases, for instance, measles and diphtheria. They can be prevented with vaccine yet still come with cases worldwide, particularly in developing nations. Through the examination of macroeconomic drivers such as health expenditure (%GDP), regional groups, and income levels, this project hopes to produce evidence-based findings that may inform public health decisions and resource allocation. In this project I aim to bridge the gap between epidemiological effect and economic policy through the application of interpretable machine learning techniques.

Data Source & Preprocessing

The data utilized were sourced from the World Bank and World Health Organization (WHO) country databases. The data include the following indicators:

- Health expenditure as percentage of GDP
- Classification at the regional and income group levels
- Reported cases annually for measles and diphtheria (2000–2020)

Data preprocessing was conducted using Python's pandas library. The datasets were merged on common keys—country name and year—to align the health expenditure data with disease case data. All year fields were cast to integer format to ensure proper joins. I filtered the data to include only the years 2000 to 2019 and retained only matching records across the three datasets (health, measles, and diphtheria). After merging, metadata such as region and income group classifications were joined to the main dataset. Missing values were removed using listwise deletion (`dropna()`), eliminating any rows with incomplete entries. No imputation or normalization of string fields (e.g., country names) was applied.

Lastly, data types were unified (e.g., year as integer, case numbers as float), and the dataset was saved for further analysis. At this point, the dataset contained numeric health and disease indicators alongside regional and economic classification columns.

For machine learning compatibility, I applied one-hot encoding to the categorical columns (Region, Income Group), creating binary indicators for each category. For handling the heavy right-skew of the distributions of Measles Cases and Diphtheria Cases, I performed a natural log transformation (along with adding a tiny constant to handle zeros) on Measles Cases and Diphtheria Cases, resulting in Log Measles Cases and Log Diphtheria Cases.

Data Analysis

Exploratory Data Analysis (EDA):

I graphed distributions with boxplots and histograms to confirm positive skewness of case numbers and variance imbalances between features. I employed correlation heatmaps and scatter plots to identify weak-to-moderate linear correlations. I also computed descriptive statistics and grouped summaries by region and income group to observe macro patterns. Other than one-hot encoding and log transformations, I verified feature scaling requirements (not required by tree models) and did proper typecasting. The final dataset comprised Health Expenditure per %GDP and 19 regional and income binary predictors.

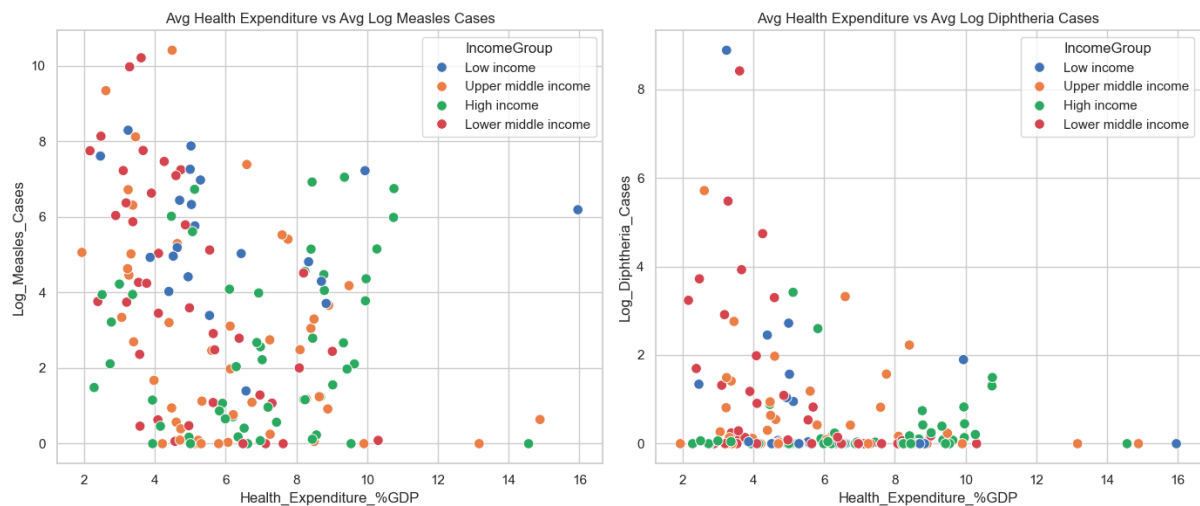


Fig. 1. Avg. Health Expenditure vs Diseases

Hypothesis Testing:

Using Pearson correlation, I correlated Health Expenditure per %GDP with log-transformed disease counts. Both diseases' p-values were significantly below 0.05 (measles: $p=0.0006$,

diphtheria: $p=0.0002$), i.e., statistically significant negative correlations. The null hypothesis of zero correlation was rejected for both.

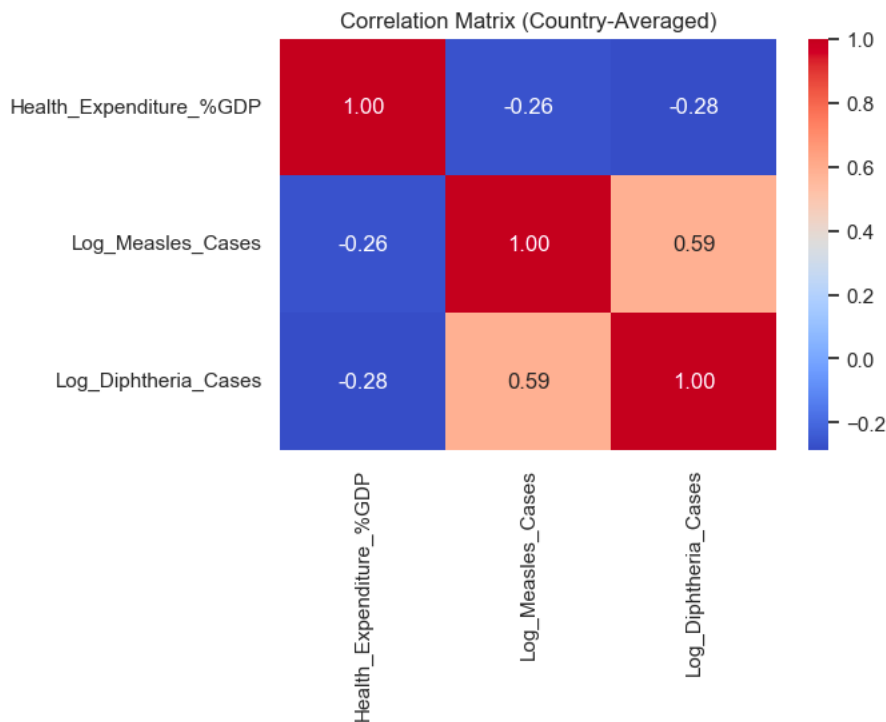


Fig. 2. Correlation Heatmap Between Health Spending and Disease Incidence

Model Training

I trained two different Random Forest Regressors using scikit-learn. Both models used the same set of features but with different target variables (Log Measles Cases and Log Diphtheria Cases).

The data was split into 80% train and 20% test sets using train test split. Default parameters were utilized to train baseline models with random state=42 for reproducibility. The models were evaluated through:

R² Score: Measure of how well the variance in the target is explained by the model.

MAE: Mean absolute magnitude of the errors made in predictions.

RMSE: Square root of mean squared errors, which penalizes larger deviations more.

Hyperparameter Tuning

For improved performance, I employed GridSearchCV with 5-fold cross-validation to find the best hyperparameters. The hyperparameter search space was as mentioned below:

n_estimators: [50, 100, 200]

max_depth: [5, 10, None]

min_samples_split: [2, 5, 10]

Minimum negative mean squared error was used to find the best models. These models were then used to predict on the test set, and performance metrics were redesigned.

Feature Importance Analysis

I have drawn bar plots for feature importances from the Random Forest models with the best tuning and presented them here. These indicate the relative importance each feature played in prediction. In both models, Health_Expenditure_%GDP dominated. Some regional features also had a high level of influence, followed by less influential income group features.

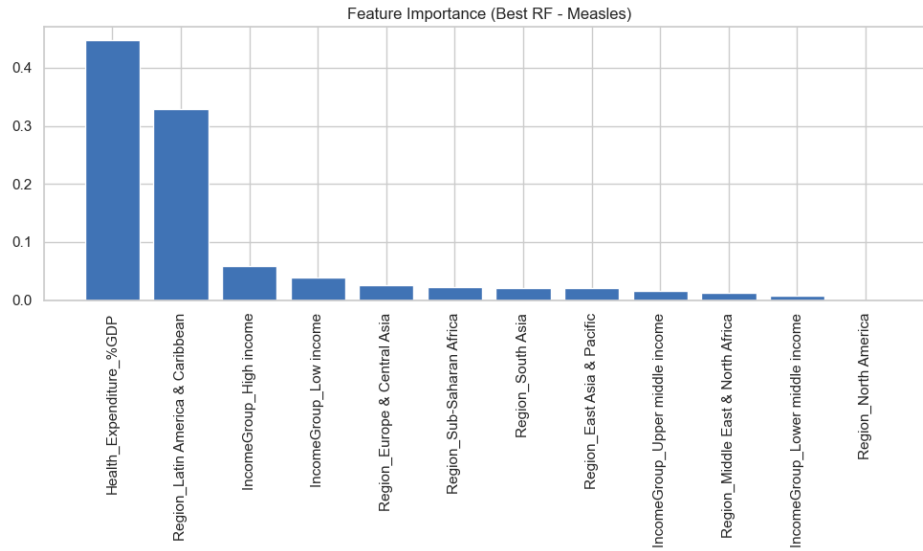


Fig. 3. Feature Importance of Measles

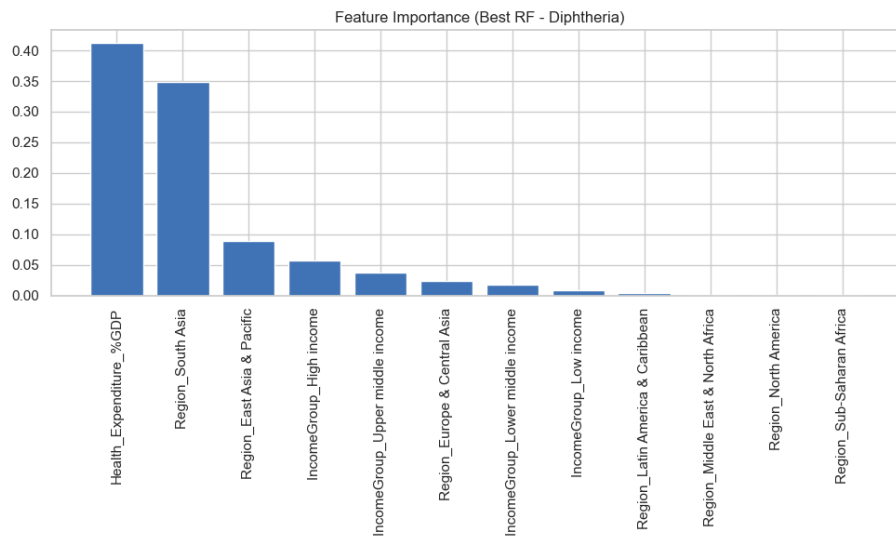


Fig. 4. Feature Importance of Diphtheria

Conclusion

- Health Expenditure per %GDP was the most powerful predictor in both models.
- For measles, the optimal Random Forest model provided $R^2 = 0.359$, $MAE = 1.832$, $RMSE = 2.545$.

- For diphtheria, the model provided $R^2 = 0.227$, $MAE = 0.75$, $RMSE = 1.243$.
- Hypothesis tests indicated:
Measles: $r = -0.264$, $p = 0.0006$
- Diphtheria: $r = -0.284$, $p = 0.0002$
Both results provided rejection of the null hypothesis, meaning significant negative correlations.
- Regional indicators (such as Latin America & Caribbean for measles, South Asia for diphtheria) emerged as strong predictors.
- Income group categories were weak predictors in both models.

Shortcomings and Future Research

Aggregated country-year data does not capture differences within countries or subnational heterogeneity.

Accuracy of reporting varies by country and adds noise to the target variables.

The model explained only a fraction of variance in prevalence of disease, suggesting omitted variables such as immunization rate, healthcare facilities, and health awareness in the population.

Other variables such as vaccination rate of children, ratio of doctors to population, and hospital bed availability may be explored in future studies.

Experimenting with various algorithms such as Gradient Boosting, LightGBM, or Neural Networks can yield higher predictive accuracy.

In correlating economic indicators with disease outcomes, our results support the assertion that spending on healthcare has a quantifiable impact on public health in low- and middle-income regions.


Use of ChatGPT During the Project

I have engaged in several conversations with ChatGPT in the process of preparing both the data analysis and the machine learning segments (submissions part 2 and 3) for my project.

To avoid being academically dishonest, I will share the conversations (inputs and outputs throughout the process) in the appendix part. It **does not include** the back and forth conversations while I was trying to get the desired output from GPT

Appendix

API SH	Görselleştirme görüntülenemedi
Metadata Country API SH	Görselleştirme görüntülenemedi
Metadata Indicator API SH	Görselleştirme görüntülenemedi
Kızamık	Görselleştirme görüntülenemedi
Difteri	Görselleştirme görüntülenemedi
Difteri	Görselleştirme görüntülenemedi

tüm datasetlerim  Yapmam gereken gereksinim bu.

I want to filter 2000 to 2019. I want to analyse between this time slot



Perfect — to analyze trends between 2000 and 2019, here's the modified Python code to filter the health expenditure and reshape it properly for merging with yearly measles and diphtheria data (assuming similar structure over years for those two as well):

Great — now let's implement **Step 2: Data Cleaning & Transformation** in code.

Your dataset is already filtered for 2000–2019, merged, and cleaned. Now we'll:

- Apply **log transformation** to `Measles_Cases` and `Diphtheria_Cases` (to reduce skew)
- (Optional) Create a **mean value per country** if you want a single value per country
- Handle any lingering **non-numeric values**

✓ Python Code for Step 2: Data Transformation

```
python  
import numpy as np

# == Replace non-numeric commas in 'Measles_Cases' (e.g. "14 156") and convert to int ==
merged["Measles_Cases"] = merged["Measles_Cases"].astype(str).str.replace(",", "").str.replace(" ", "")
merged["Diphtheria_Cases"] = merged["Diphtheria_Cases"].astype(float)

# == Apply log(1 + x) transformation to cases ==
```

```

# == Apply log(1 + x) transformation to cases ==
merged["Log_Measles_Cases"] = np.log1p(merged["Measles_Cases"])
merged["Log_Diphtheria_Cases"] = np.log1p(merged["Diphtheria_Cases"])






# == OPTIONAL: If you want to aggregate to average per country over the years ==
mean_df = merged.groupby("Country Name").agg({
    # "Health_Expenditure_%GDP": "mean",
    # "Measles_Cases": "mean",
    # "Diphtheria_Cases": "mean",
    # "Log_Measles_Cases": "mean",
    # "Log_Diphtheria_Cases": "mean",
    # "Region": "first",
    # "IncomeGroup": "first"
}).reset_index()

# == Show transformed data ==
merged.head()

```

Step 4: Exploratory Data Analysis (EDA)

Let me know and I'll give you full Python code to:

1. Plot histograms 
2. Plot scatterplots 
3. Plot box plots by income group 
4. Get descriptive statistics 
5. Correlation heatmap 

```

import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import pearsonr

# == Scatter Plots ==
plt.figure(figsize=(14, 6))

plt.subplot(1, 2, 1)
sns.scatterplot(data=mean_df, x="Health_Expenditure_%GDP", y="Log_Measles_Cases", hue="IncomeGroup")
plt.title("Avg Health Expenditure vs Avg Log Measles Cases")

plt.subplot(1, 2, 2)
sns.scatterplot(data=mean_df, x="Health_Expenditure_%GDP", y="Log_Diphtheria_Cases", hue="IncomeGroup")
plt.title("Avg Health Expenditure vs Avg Log Diphtheria Cases")

plt.tight_layout()
plt.show()

# == Correlation Coefficients ==
corr_measles, pval_measles = pearsonr(mean_df["Health_Expenditure_%GDP"], mean_df["Log_Measles_Cases"])
corr_diphtheria, pval_diphtheria = pearsonr(mean_df["Health_Expenditure_%GDP"], mean_df["Log_Diphtheria_Cases"])

```



These are the codes and results of my data analysis project. For the next part the instruction just says: apply machine learning methods to the dataset. How should we continue? I also sent you the readme file for you to understand the project.

Thanks for sharing the full project screenshots. Based on your completed EDA and hypothesis testing, the

✓ 2. Suggested Models

Depending on the task you choose above:

For Classification (`IncomeGroup`)

- Logistic Regression (baseline)
- Random Forest Classifier
- XGBoost Classifier (strong performance, handles non-linearity)
- SVM (if features are well-separated)
- KNN (for visual, intuitive explanation)

For Regression (`Log_Measles_Cases` , `Log_Diphtheria_Cases`)

- Linear Regression
- Random Forest Regressor
- XGBoost Regressor
- Ridge/Lasso Regression (for interpretability)

✓ 3. Preprocessing Tips

- One-hot encode categorical variables: `Region` , `IncomeGroup` (if not predicted).
- Standardize features (especially for models like SVM or Logistic Regression).
- Train/test split (e.g., 80/20).
- Use stratified splitting if doing classification.

✓ 4. Model Evaluation

Use the following metrics:

Classification:

- Accuracy, Precision, Recall, F1
- Confusion Matrix
- Classification Report



✓ Recommended Models

Start simple → increase complexity:

Model	Why Use It?
Linear Regression	Baseline; easy to interpret effect sizes
Random Forest Regressor	Captures nonlinearities, good performance
XGBoost Regressor	State-of-the-art; handles noise/outliers well
Ridge/Lasso	Useful if you want coefficients + regularization

ama yani ben accurate model yapmak istiyorum heh, hep meanmerged mü kullandık biz? o data visualization kolay olsun diye değil miydi? merged'de bir sürü data var 2500 satır falan

Evet, şu ana kadar hep `mean_df` (164 satırlık, ülke bazında ortalamalar) ile çalıştık. Bu, görselleştirme ve genel korelasyonları incelemek için mantıklıydı.

Ancak:

- `merged_df` aslında **2647 satır** içeriyor ve her yıl için ülke verilerini barındırıyor.
- Yani: **modelleme için daha fazla veri** → **potansiyel olarak daha iyi tahmin gücü** demek.

Özetle:

- Evet, `mean_df` görselleştirme ve korelasyon analizi için uygundu.
- Ama sen **daha accurate bir model** istiyorsan, artık `merged_df` ile yeniden eğitim yapmak mantıklı. Bu daha fazla örnek (data point) sağlar, modeli genelleştirme açısından iyileştirebilir.

Hazırsan, `merged_df` üzerinden modellemeye geçelim ↓ ? Baştan feature engineering (encoding, split, model evaluation)...

✓ Observations about `merged_df`:

- It covers **yearly data from 2000–2019**, across many countries.
- It includes:
 - `Health_Expenditure_%GDP`
 - Disease counts and log-scaled versions
 - `Region` and `IncomeGroup` (categorical)
- No missing values — that's great.

✓ Next steps (clean, scalable modeling):

Let's:

1. Encode `Region` and `IncomeGroup`.
2. Train/test split.
3. Train `RandomForestRegressor` or `XGBoost` (we'll compare).
4. Evaluate R^2 , MAE, RMSE.
5. Plot feature importances.



Next step:

I'll now:

- Split into train/test
 - Train a model (RandomForest first)
 - Evaluate with R^2 , MAE, RMSE
- Shall I continue with that? [y-]

Perfect — we've now prepped your full dataset:

✓ Dataset Overview:

- **2,647 rows, 10 features** after one-hot encoding.
- Features include:
 - `Health_Expenditure_%GDP`
 - Region dummies (6 regions, one dropped)
 - Income group dummies (3 groups, one dropped)
- Targets:
 - `Log_Measles_Cases`
 - `Log_Diphtheria_Cases`

tamamdır, o zaman baştan başlayalım, encoding, Step 2: Define Features & Targets, Train/Test Split Step 3: Train Random Forest Regressor for Measles and Diphtheria geklinde bana kodları ver sonra da hyperparameter tuning ve feature importance bakarız

Harika! Aşağıda sırayla ilerleyen kodları bulacaksın. Datasetin `merged_health_measles_diphtheria_2000_2019.csv` olduğunu ve `pandas` ile yüklediğini varsayıyorum (`merged` adında bir DataFrame).

✓ Step 1: Encoding Categorical Variables

```
python
# One-hot encoding for Region and IncomeGroup
df_encoded = pd.get_dummies(merged, columns=['Region', 'IncomeGroup'], drop_first=True)
```

✅ Step 1: Encoding Categorical Variables

```
python 📄 Kopyala 🗑️ Düzenle  
  
# One-hot encoding for Region and IncomeGroup  
df_encoded = pd.get_dummies(merged, columns=['Region', 'IncomeGroup'], drop_first=True)
```

✅ Step 2: Define Features & Targets, Train/Test Split

```
python 📄 Kopyala 🗑️ Düzenle  
  
from sklearn.model_selection import train_test_split  
  
# Drop unnecessary columns  
X = df_encoded.drop(columns=[  
    "Country Name", "Country Code",  
    "Measles_Cases", "Diphtheria_Cases",  
    "Log_Measles_Cases", "Log_Diphtheria_Cases"  
)  
↓
```

```
📄 Kopyala 🗑️ Düzenle  
  
# Fit models  
rf_measles.fit(X_train_m, y_train_m)  
rf_diphtheria.fit(X_train_d, y_train_d)  
  
# Predictions  
y_pred_m = rf_measles.predict(X_test_m)  
y_pred_d = rf_diphtheria.predict(X_test_d)  
  
# Evaluation  
print("=== Measles Evaluation ===")  
print("R² Score :", round(r2_score(y_test_m, y_pred_m), 3))  
print("MAE      :", round(mean_absolute_error(y_test_m, y_pred_m), 3))  
print("RMSE      :", round(mean_squared_error(y_test_m, y_pred_m, squared=False), 3))  
  
print("\n=== Diphtheria Evaluation ===")  
print("R² Score :", round(r2_score(y_test_d, y_pred_d), 3))  
print("MAE      :", round(mean_absolute_error(y_test_d, y_pred_d), 3))  
print("RMSE      :", round(mean_squared_error(y_test_d, y_pred_d, squared=False), 3))
```

✅ Step 4: Hyperparameter Tuning with GridSearchCV (for Measles)

```
python 📄 Kopyala 🗑️ Düzenle  
  
from sklearn.model_selection import GridSearchCV  
  
# Define hyperparameter grid  
param_grid = {  
    'n_estimators': [50, 100, 200],  
    'max_depth': [5, 10, None],  
    'min_samples_split': [2, 5, 10]  
}  
  
# GridSearchCV setup  
grid_search = GridSearchCV(  
    RandomForestRegressor(random_state=42),  
    param_grid,  
    cv=5,  
    scoring='neg_mean_squared_error',  
    n_jobs=-1  
)  
↓
```

```
importances = desc_1.feature_importances_  
features = X_train_m.columns  
indices = np.argsort(importances)[::-1]  
  
# Plot  
plt.figure(figsize=(10, 6))  
plt.title("Feature Importance (Best RF - Measles)")  
plt.bar(range(len(importances)), importances[indices])  
plt.xticks(range(len(importances)), features[indices], rotation=90)  
plt.tight_layout()  
plt.show()
```

İstersen aynı `GridSearchCV` adımlarını **diphtheria** için de hemen ekleyebilirim. Ayrıca istersen XGBoost veya başka bir modelle karşılaştırmalı sonuçlar da alabiliriz. Hazırsan devam edelim.

```
print("Pearson Correlation (Health Exp vs Log Measles Cases):", round(corr_measles, 3))  
print("P-value:", round(pval_measles, 4))  
  
print("Pearson Correlation (Health Exp vs Log Diphtheria Cases):", round(corr_diphtheria, 3))  
print("P-value:", round(pval_diphtheria, 4))  
  
# === Heatmap for Correlation Matrix ===  
plt.figure(figsize=(6, 4))  
sns.heatmap(mean_df[[  
    "Health_Expenditure_%GDP",  
    "Log_Measles_Cases",  
    "Log_Diphtheria_Cases"  
]].corr(), annot=True, cmap="coolwarm", fmt=".2f")  
plt.title("Correlation Matrix (Country-Averaged)")  
plt.show()
```

References

World Health Organization (WHO). (2024). *Measles and Diphtheria - Immunization Data*. Retrieved from <https://www.who.int/data/gho/data/themes/immunization>

World Bank. (2024). *Health expenditure, total (% of GDP)*. Retrieved from <https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS>

World Bank. (2024). *World Bank Country and Lending Groups*. Retrieved from <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519>