Question 3 + 4

720017170

Question 3 - See QMD for Code

(a) Import each dataset into memory as a separate data frame, keeping all countries as your sample.

```
plt.rcParams.update({'font.size': 14})

# Loading in the Data
Health_Data = pd.read_csv('Data/Health.csv', index_col=None)
Infant_Data = pd.read_csv("Data/Infant.csv",index_col=None)

# Replace .. with NA
Health_Data.replace("..", pd.NA, inplace=True)
Infant_Data.replace("..", pd.NA, inplace=True)

# Removing unnecessary columns
Health_Data = Health_Data.drop(columns=['Series Name', 'Series Code'])
Infant_Data = Infant_Data.drop(columns=['Series Name', 'Series Code'])

# Remove names in []
Health_Data.columns = Health_Data.columns.str.replace(r'\[.*\]', '', regex=True)
Infant_Data.columns = Infant_Data.columns.str.replace(r'\[.*\]', '', regex=True)

print(Infant_Data.head())
print(Health_Data.head())
```

```
     Country Name Country Code 2000  2001  2002  2003  2004  2005  2006   \
0     Afghanistan          AFG    92  89.3  86.6  83.7  80.9    78  75.1
1         Albania          ALB    24  22.9  21.6  20.4  19.1  17.8  16.5
2         Algeria          DZA  35.6  34.3    33  31.6  30.3    29  27.8
3  American Samoa          ASM  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
```

```
4         Andorra          AND   6.5   6.3     6   5.8   5.6   5.3   5.1

  2007   ... 2014  2015  2016  2017  2018  2019  2020  2021  2022  2023
0  72.3  ...  56.2  54.6    53  51.5  50.1  48.8  47.4  46.1  44.8  <NA>
1  15.3  ...   8.8   8.5   8.4   8.3   8.3   8.3   8.4   8.4   8.4  <NA>
2  26.6  ...    22  21.7  21.4    21  20.6  20.1  19.7  19.2  18.7  <NA>
3  <NA>  ...  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
4   4.9  ...   3.5   3.4   3.2   3.1     3   2.9   2.8   2.7   2.6  <NA>

[5 rows x 26 columns]
     Country Name Country Code          2000           2001         2002   \
0     Afghanistan          AFG          <NA>           <NA>  17.00758553
1         Albania          ALB     65.1501236    73.78884125  78.99478149
2         Algeria          DZA    62.11769485    67.33850098  66.94760132
3  American Samoa          ASM          <NA>           <NA>         <NA>
4         Andorra          AND  1287.00280762  1336.21142578  1486.171875

          2003          2004          2005          2006          2007   \
0    17.81492424   21.42946434   25.10707283   28.91982269   32.71720505
1   106.29218292  138.11340332  152.12762451  166.81382751  212.61096191
2    76.23547363   93.02433014  101.30373383  117.43313599  151.77920532
3          <NA>          <NA>          <NA>          <NA>          <NA>
4  1772.71337891  1990.0748291  2214.64697266  2139.27539063  2489.43115234

   ...          2014          2015          2016          2017   \
0  ...    60.18957901    60.05854034    61.48645782    66.90921783
1  ...   295.12359619   255.35635376   277.04321289    297.4619751
2  ...   361.15942383     292.275177   261.40023804   265.83843994
3  ...          <NA>          <NA>          <NA>          <NA>
4  ...  3089.84301758  2688.20629883  2755.44848633  2873.29614258

          2018          2019          2020          2021  2022  2023
0    71.33430481    74.23410797    80.28805542    81.31976318  <NA>  <NA>
1    351.3012085   367.75839233   396.88024902   464.74285889  <NA>  <NA>
2   266.46469116   235.99041748   206.03512573   204.56661987  <NA>  <NA>
3          <NA>          <NA>          <NA>          <NA>  <NA>  <NA>
4  3164.38842773  3026.59741211  3269.29736328  3505.99145508  <NA>  <NA>

[5 rows x 26 columns]
```

(b) If data are not already stored in this way, please reshape data so that they consist of a single line of data for each country and year.

```
# Pivoting the Data into a long format
Health_Data_long = pd.melt(Health_Data,
        id_vars=['Country Name', 'Country Code'],
        var_name='Year',
        value_name='Heathcare Expenditure (USD)')
Infant_Data_long = pd.melt(Infant_Data,
        id_vars=['Country Name', 'Country Code'],
        var_name='Year',
        value_name='Infant Mortality Rates (per 1,000 live births)')
```

	Country Name	Country Code	Year	Heathcare Expenditure (USD)
0	Afghanistan	AFG	2000	<NA>
1	Albania	ALB	2000	65.1501236
2	Algeria	DZA	2000	62.11769485
3	American Samoa	ASM	2000	<NA>
4	Andorra	AND	2000	1287.00280762

	Country Name	Country Code	Year	Infant Mortality Rates (per 1,000 live births)
0	Afghanistan	AFG	2000	92
1	Albania	ALB	2000	24
2	Algeria	DZA	2000	35.6
3	American Samoa	ASM	2000	<NA>
4	Andorra	AND	2000	6.5

(c) Calculate the total number of countries observed in each data frame Calculate the total number of years observed in each data frame.

```
# Counts the number of unique contries
num_countries_FDI = Health_Data_long["Country Name"].nunique()

# Outputs the number of unique countries using an f string
print(f"Total number of unique countries in Health_Data: {num_countries_FDI}")

num_years_FDI = Health_Data_long["Year"].nunique() # Counts the number of unique years
print(f"Total number of unique years observed in Health_Data: {num_years_FDI}")
```

```
num_countries_GDP = Infant_Data_long["Country Name"].nunique()
print(f"Total number of unique countries in Infant_Data: {num_countries_GDP}")

num_years_GDP = Infant_Data_long["Year"].nunique()
print(f"Total number of unique years observed in Infant_Data: {num_years_GDP}")
```

```
Total number of unique countries in Health_Data: 217
Total number of unique years observed in Health_Data: 24
Total number of unique countries in Infant_Data: 217
Total number of unique years observed in Infant_Data: 24
```

(d) Calculate the number of observations for which data is missing

```
# Sums the number of missing values in each dataset
missing_values_Health = Health_Data_long.isna().sum().sum()
print(f"Total missing observations in Health_Data: {missing_values_Health}")

missing_values_Infant = Infant_Data_long.isna().sum().sum()
print(f"Total missing observations in Infant_Data: {missing_values_Infant}")

# Create a dataframe showing the number of missing values for each dataset
missing_values_table = pd.DataFrame({
    'Dataset': ['Health_Data', 'Infant_Data'],
    'Total Missing Observations': [missing_values_Health, missing_values_Infant]
})
```

```
Total missing observations in Health_Data: 1099
Total missing observations in Infant_Data: 700
```

(e) Join the two files by country and year so that you have single dataframe containing both variables. Explain clearly what type of join this is, and carefully check that the number of observations resulting from the join makes sense.

```
# Merge the data on Country Name, Country code and Year
merged_data = pd.merge(Health_Data_long, Infant_Data_long, on=['Country Name',
'Country Code', 'Year'])
print(merged_data.head())

# Print the number of rows in the DataFrame
num_rows = merged_data.shape[0]
print(f"Number of rows in the DataFrame: {num_rows}")
```

```
     Country Name Country Code   Year Heathcare Expenditure (USD)  \
0     Afghanistan          AFG  2000                         <NA>
1         Albania          ALB  2000                   65.1501236
2         Algeria          DZA  2000                  62.11769485
3  American Samoa          ASM  2000                         <NA>
4         Andorra          AND  2000                1287.00280762

  Infant Mortality Rates (per 1,000 live births)
0                                             92
1                                             24
2                                           35.6
3                                           <NA>
4                                            6.5
Number of rows in the DataFrame: 5208
```

The join completed in the above code chunk is an inner join and only keeps rows that exist in both Health_Data_long and Infant_Data_long. If a country-year exists in one dataset but not the other, it will be dropped.

Question 4 - Investigating the Relationship Between Current Healthcare Expenditure per capita and Infant Mortality Rates from 2000 - 2022

Missing Data - Table 1

	Dataset	Total Missing Observations
0	Health_Data	1099
1	Infant_Data	700

Both datasets contained a considerable amount of missing data, illustrated in Table 1. Potentially due to countries not collecting the data or collecting the data at different year intervals. Missing data can have a large impact on data analysis if not handled properly and can lead to skewed or incorrect conclusions. The year 2023 contained no data; therefore, this column was dropped. To deal with the other missing data, I decided to drop all rows containing missing data, sometimes, this could result in a significant reduction of sample size; however, in this case, 1099 observations were removed (21.1% of the dataset) and only 27 countries were dropped, indicating this was an effective method to handling missing data as there were still 4109 observations. An alternative approach would've been mean, multiple or regression imputation if dropping rows with missing data caused a significant decrease in sample size.

Summary Statistics - Table 2

Variable	N	Mean	Median	SD	Min	Max
Heathcare Expenditure (USD)	4109.0	956.0	256.7	1685.7	4.0	12473.8
Infant Mortality Rates (per 1,000 live births)	4109.0	26.9	17.7	25.0	1.4	138.3

Table 2 displays the summary statistics for healthcare expenditure (USD) and infant mortality rates (per 1,000 live births) across 4109 observations, revealing significant differences between countries.

Healthcare expenditure per capita showed a mean of $956.0 but a far lower median of $256.7, indicating a negatively skewed distribution where few countries spend significantly more. The large standard deviation ($1,685.7) and range ($4.0–$12,473.8) highlight large global and temporal differences in healthcare investment.

Infant mortality rates show similar variation, with a mean of 26.9 deaths per 1,000 live births and a median of 17.7. The high standard deviation (25.0) and range (1.4–138.3) suggest major differences in healthcare quality and access.

Overall, the data indicates global differences in healthcare funding and outcomes, suggesting that higher healthcare expenditure may be linked to lower infant mortality.

Distribution Analysis

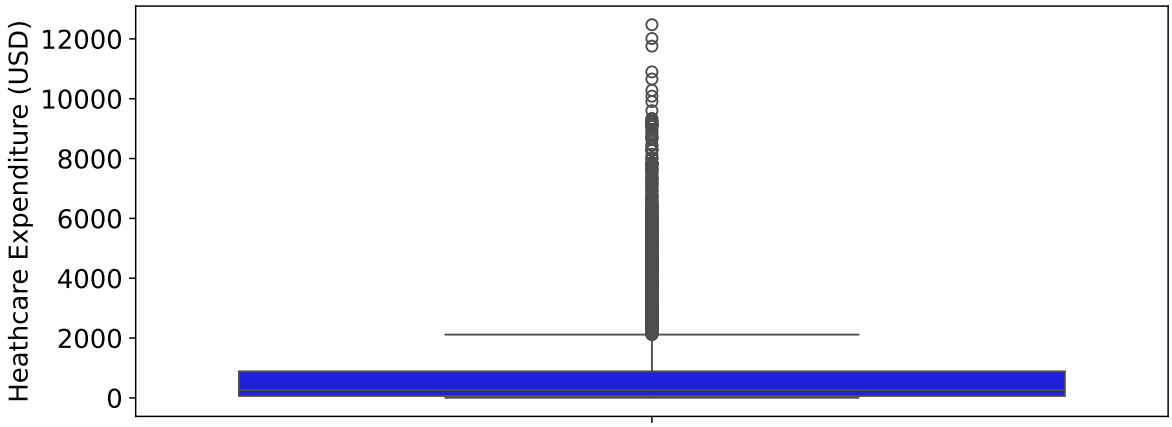

Figure 1: Box Plot of Healthcare Spend Per Capita (USD)

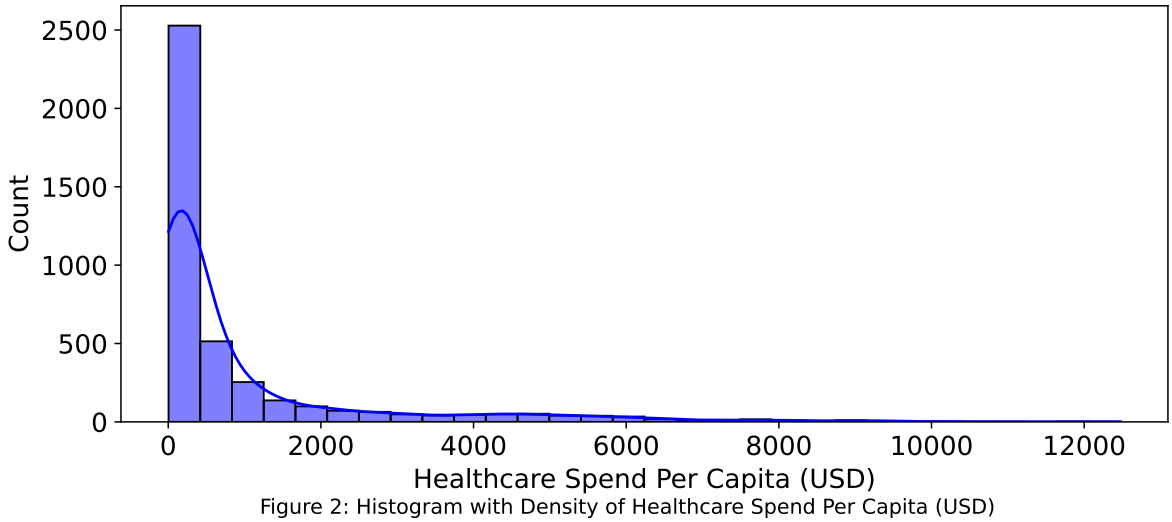

Figure 2: Histogram with Density of Healthcare Spend Per Capita (USD)

Figures 1 and 2 show the distribution of healthcare spending per capita (USD). Figure 1 shows that healthcare expenditure is highly negatively skewed, with many outliers at the upper end, supporting the analysis from the summary statistics. The median expenditure is positioned toward the lower end of the distribution, indicating that most countries spend relatively little, while a few spend significantly more. The whiskers of the box plot are short, suggesting that a large proportion of the data is concentrated within a lower range, while the numerous outliers highlight extreme spending levels in some countries.

Figure 2 reiterates the negative skew of the data. Most countries have low healthcare spending, grouped toward the left of the axis, with just a handful having exceptionally high expenditures. The density curve (smooth blue line) shows the exponential drop in frequency as expenditure increases, emphasising that high-spending countries are exceptions rather than the rule

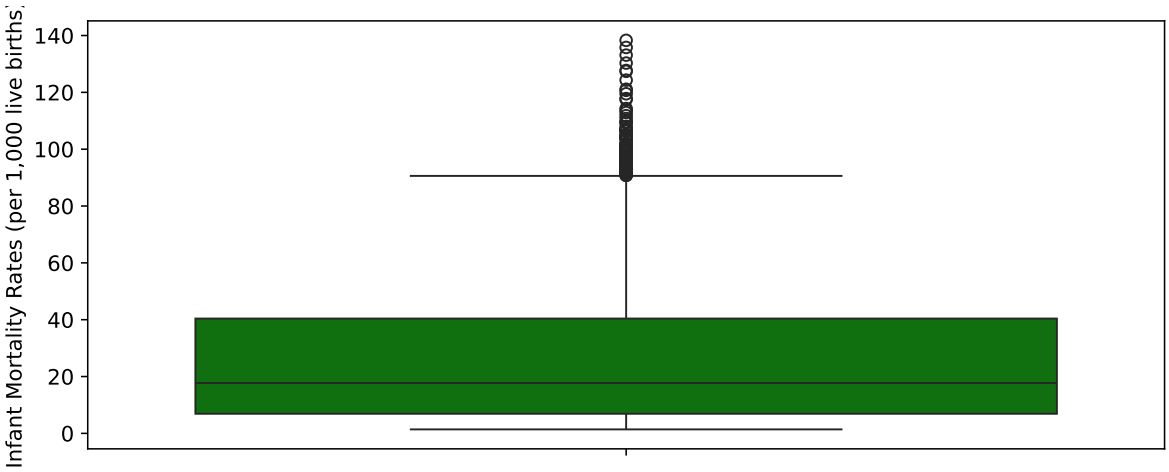

Figure 3: Box Plot of Infant Mortality Rate (per 1,000 live births)

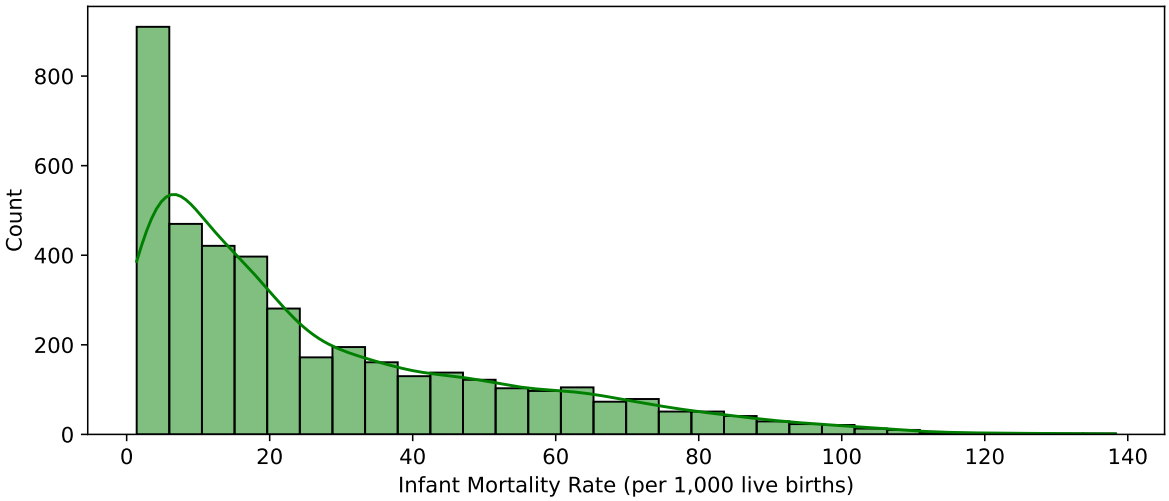

Figure 4: Histogram with Density of Infant Mortality Rate (per 1,000 live births)

The distribution of the infant mortality rate (per 1,000 live births) seen in Figures 3 and 4 is comparable to that of healthcare spend per capita (USD).

There are several outliers in Figure 3, with specific countries having abnormally high infant mortality rates. The data is still negatively skewed, as seen by the median being significantly lower than the upper quartile.

This is supported by Figure 4, which displays a dramatic fall in frequency as rates rise, with most values clustering below 40 deaths per 1,000 live births. While infant mortality is low in many nations, it is much higher in others, most likely because of infrastructural constraints, economic considerations, and healthcare discrepancies.

Due to the number of outliers within both variables, I decided to use the Interquartile Range

(IQR) method to remove any outliers to ensure a more accurate and reliable regression analysis.

Outliers can skew the model fit, disproportionately influence regression coefficients, and inflate evaluation metrics such as RMSE. By applying the IQR method, I ensure that the data represents a more consistent trend, reducing the impact of extreme values and improving the model's ability to capture the true relationship between healthcare expenditure and infant mortality rates.

Correlation Analysis

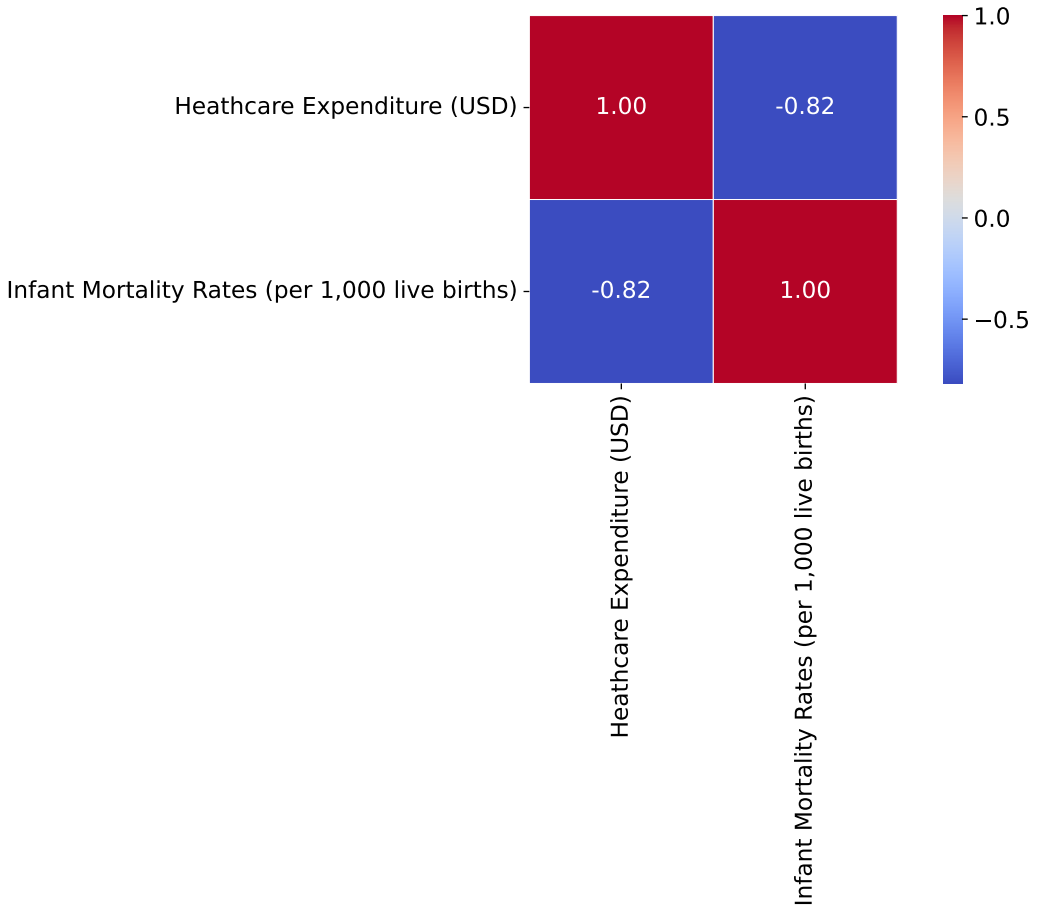

Figure 5: Spearman Rank Correlation Matrix of Healthcare Expenditure and Infant Mortality Rates

Figure 5 illistrates the relationship between heathcare expenditure and infant mortality rates. The correlation coefficient of -0.82 indicates a strong negative correlation, suggesting that as healthcare expenditure increases, infant mortality rates tend to decrease. This aligns with economic and public health expectations, where greater investment in healthcare typically leads to better medical infrastructure, improved maternal care, and reduced infant deaths. The Spearman correlation is particularly useful here as it captures non-linear relationships, making

it more robust to potential outliers compared to Pearson's correlation. However, correlation does not imply causation, and additional factors such as healthcare efficiency, socioeconomic disparities, and government policies could influence this relationship.

Regression Analysis

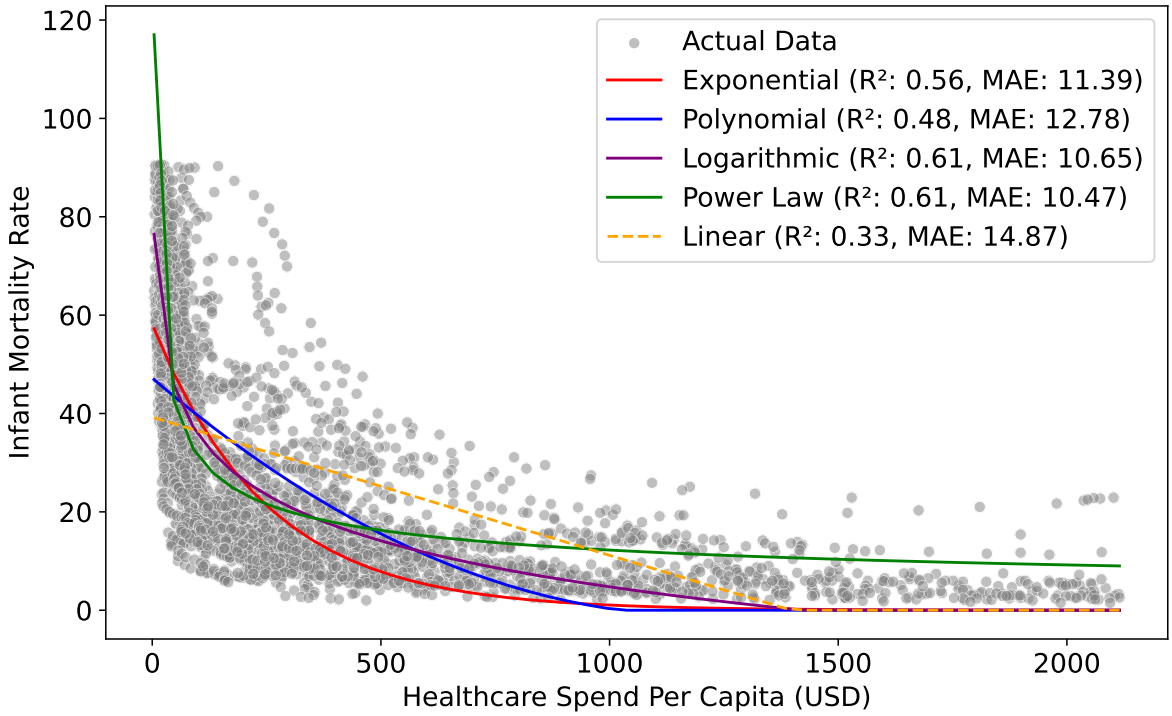

Figure 6: Comparison of Regression Models: Healthcare Spend vs Infant Mortality Rate

Figure 6 compares five regression models, exponential, polynomial, logarithmic, power law, and linear, in quantifying the relationship between healthcare spend per capita and infant mortality rates. The power law model achieves the best fit with an R^2 of 0.61, revealing that 61% of the variance in infant mortality rates is explained by healthcare spending per capita, and the lowest mean absoloute error (MAE) of 10.47, supporting the strong inverse relationship shown in Figure 5. Exponential and logarithmic regression perform well ($R^2 = 0.61$ and 0.56, respectively), capturing the steep initial decline before stabilizing at lower mortality levels.

Polynomial regression initially follows the trend but ultimately performs poorly, indicated by its lower R^2 of 0.48. The linear model performs worst, failing to capture the non-linearity of the data. All models are constrained to prevent infant mortality predictions below 0, ensuring a realistic representation. This analysis highlights that non-linear models, particularly power law and exponential regression, provide the most accurate representation of the data, reiterating the non-linearity of the relationship between healthcare spend per capita and infant mortality rates.

Regression Model Evaluation

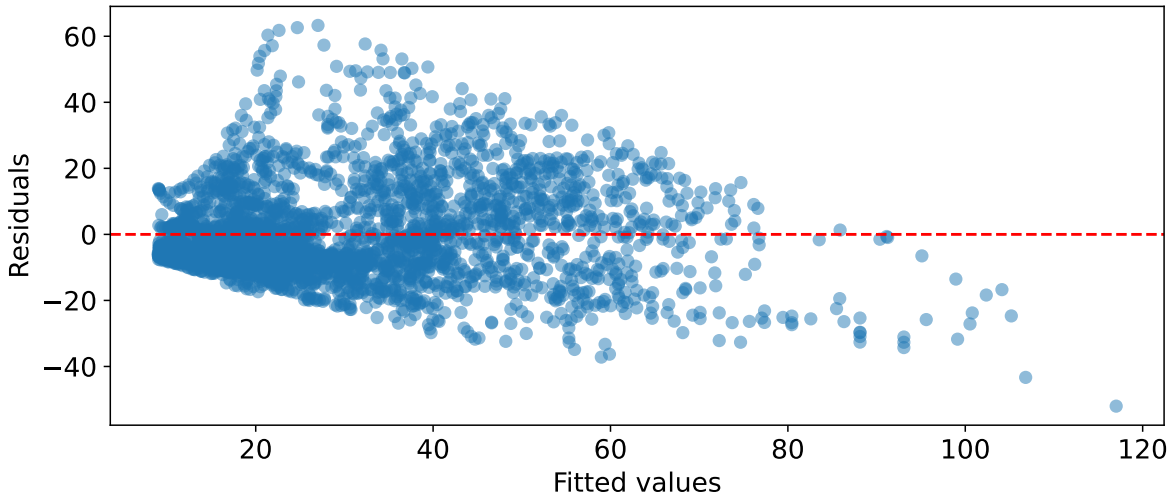

Figure 7: Residual Plot of Power Law Regression

The residual plot (Figure 7) for the power law regression model highlights some issues with the model's fit. The residuals display a noticeable pattern rather than being randomly scattered, suggesting heteroscedasticity—where the variance of residuals increases with fitted values. This indicates that the model performs well at lower levels of healthcare expenditure but struggles to maintain accuracy as expenditure increases. The presence of extreme residuals also suggests potential outliers influencing the model. To address these issues, applying a log transformation to the dependent variable or exploring alternative regression techniques, such as Generalized Least Squares (GLS), may help improve the model's fit.

Time Series Analysis

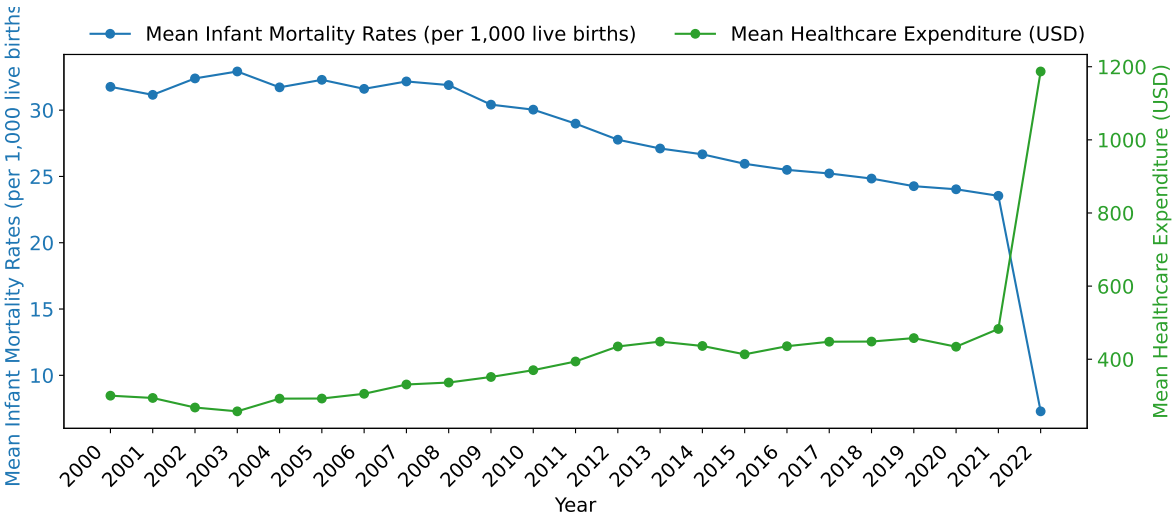

Figure 8: Time Series Analysis of Mean Infant Mortality Rates and Healthcare Expenditure

Figure 8 illustrates the negative association between mean infant mortality rates (IMR) and mean healthcare spending (USD) from 2000 to 2022. The trend indicates a continual drop in IMR, reflecting improvements in healthcare, economic development, and medical advances. Meanwhile, healthcare spending has steadily climbed, indicating greater investment in health infrastructure and services.

A noteworthy anomaly arises in 2022 when healthcare expenditures significantly increase and infant death rates significantly decrease. This large shift may be due to pandemic-related spending, emergency health interventions, or data errors. The high rise in expenditure may signal a significant policy shift, but further research is needed to assess the long-term consequences.

Overall, Figure 8 reinforces the strong negative correlation between healthcare investment and infant mortality rates illustrated by Figures 5 and 6. However, the presence of diminishing returns suggests that beyond a certain threshold, additional spending alone may not yield proportional improvements, emphasizing the need for efficient resource allocation and targeted healthcare policies to maximize impact.

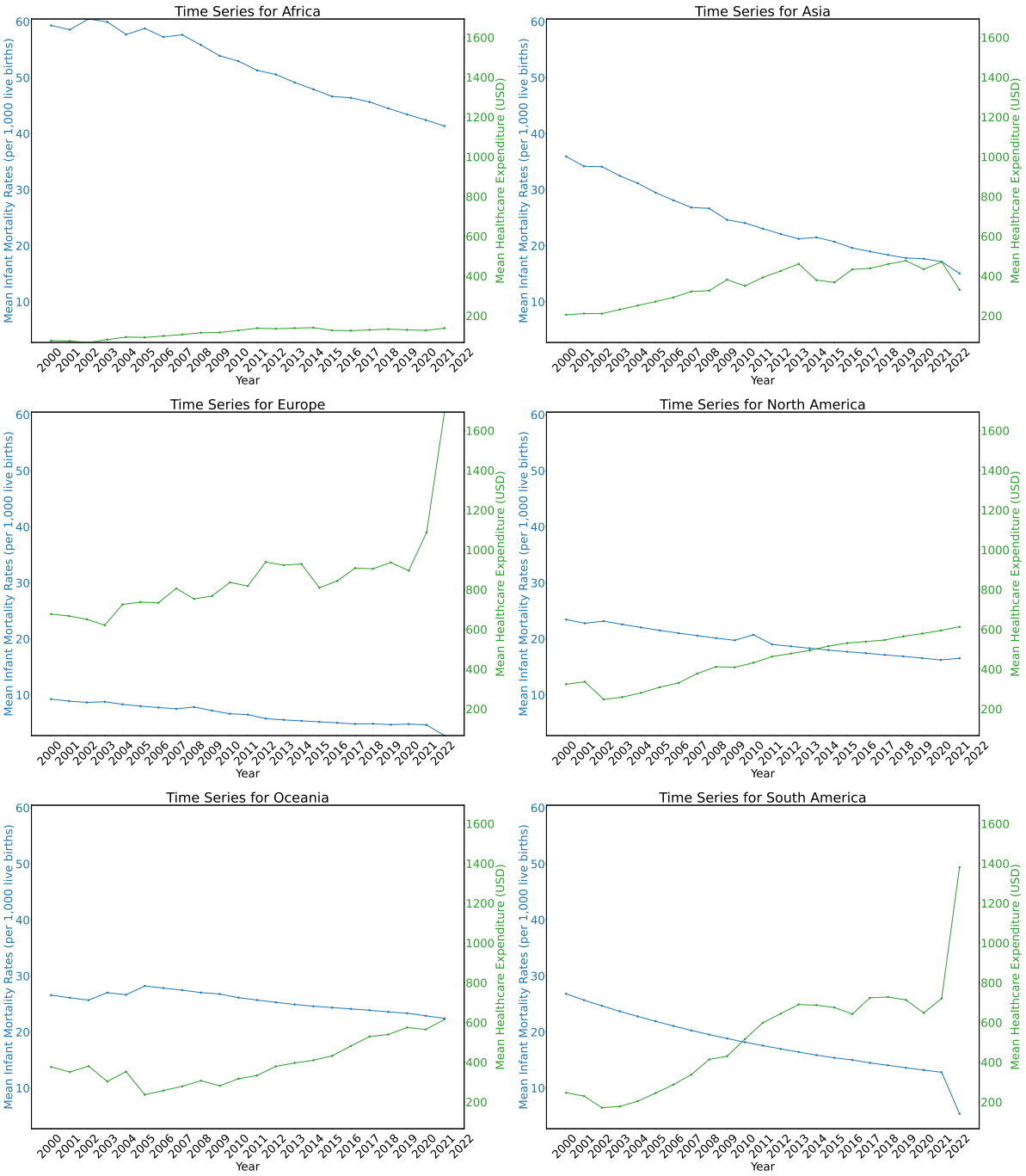

Figure 9: Time Series Analysis of Mean Infant Mortality Rates and Healthcare Expenditure By Continent

Figure 9 shows continent-specific trends in IMR and mean healthcare spending (USD) from 2000 to 2022. The constant negative relationship across each location suggests an association between increasing healthcare investment and reduced infant mortality.

Africa and Asia have had considerable drops in IMR, indicating significant improvements in healthcare despite relatively little investment. Europe and North America, with larger beginning expenditures, experienced lower IMR decreases, indicating declining returns on investment. South America and Oceania follow similar trends but have different purchasing preferences.

Healthcare expenditure rose significantly in 2022 in North and South America, most likely because of pandemic-related measures. This substantial rise in spending corresponds to a transient acceleration in IMR decreases, implying short-term healthcare gains.

Overall, Figure 9 highlights regional differences in healthcare availability and efficiency. While investment is critical to lowering infant mortality, the impact of expenditure varies by location.

Conclusion

This analysis confirms a strong non-linear negative relationship between healthcare expenditure and infant mortality rates across global regions. Countries with higher healthcare investment generally experience lower infant mortality, though the impact varies based on economic and healthcare infrastructure factors.

The non-linear nature of this relationship indicates diminishing returns, where initial increases in spending lead to substantial improvements, but further investments yield smaller reductions in infant mortality. Logarithmic and time-series analyses further reinforce this pattern, highlighting long-term declines in infant mortality alongside growing healthcare expenditure.

Regional disparities remain significant, with high-income regions investing more per capita while achieving lower mortality rates, whereas lower-income regions show greater relative improvements despite lower absolute spending. The 2022 anomaly suggests short-term shifts in healthcare spending and outcomes, likely due to pandemic-driven policies.

While healthcare expenditure is a crucial driver of infant survival, efficient allocation, accessibility, and policy effectiveness remain key determinants of long-term health improvements worldwide.

Link to GitHub Repository = https://github.com/JoshLG18/DSE-Assignment1