Vaishnavi Vishwas Pawar
vpawar@iu.edu

Analyzing the Relationship Between GDP Growth and Healthcare Spending Across Regions Using World Bank Open Data

I. Introduction

Economic growth and healthcare investment are critical drivers of societal development. Understanding the relationship between a country's GDP growth and healthcare spending can provide insight into how resources are allocated and the implications for public well-being. This project looks at trends, correlations, and disparities in GDP growth and healthcare spending across a two-decade global dataset.

This project cleans, processes, and analyzes large datasets with big data tools and platforms like Google Cloud Platform (GCP), Google BigQuery, and Python Notebooks. These tools enable efficient data storage, distributed query processing, and dynamic visualization. The study's goal is to uncover patterns that can inform data-driven global and regional development strategies by deploying a scalable data pipeline and conducting exploratory data analysis.

II. Background

The allocation of resources to healthcare has a significant impact on societal well-being, economic resilience, and national development. However, the relationship between GDP growth and healthcare spending varies greatly by region, reflecting variations in economic priorities, governance structures, and societal needs. Because of this variability, the topic is not only fascinating, but also extremely relevant to researchers, policymakers, and stakeholders looking to optimize resource allocation.

This project was chosen to meet the urgent demand for data-driven insights into the relationship between economic performance and healthcare investment. This study breaks through traditional analytical constraints by analyzing comprehensive World Bank datasets using advanced big data platforms such as Google Cloud Platform (GCP) and BigQuery. The project's emphasis on scalability, reproducibility, and actionable insights is consistent with the course's overarching goals of applying big data concepts to complex, real-world problems. Understanding these patterns can help global organizations, governments, and economic forums develop targeted policies that balance economic growth with healthcare access and quality.

III. Methodology

This project took a structured approach to analyzing the relationship between GDP growth and healthcare spending, using cloud platforms and big data tools for scalable and efficient processing. The methodology is described in the following steps, which are supported by the pipeline overview diagram.

1] Dataset Acquistion:

The datasets were obtained from the World Bank's open data repository, which covers GDP and healthcare expenditure indicators from 1962 to 2023. These datasets were chosen because they are useful for understanding the trends and correlations between economic growth and healthcare investment.

To meet the course requirements, the project was created in the **FA24-BL-INFO-I535 folder** on **Google Cloud Platform (GCP)**, with the naming convention **FA24-I535-vpawar-GDPHealthcare.**

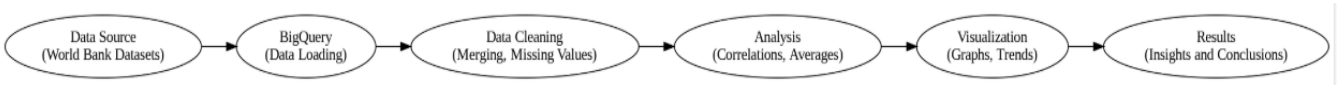


Fig 3.1 Data Pipeline Overview

Data Pipeline Overview: This pipeline depicts the project's path from data collection to final insights. Each step was completed with scalability and reproducibility in mind, leveraging GCP, BigQuery, and Python.

2] Data Cleaning and Preprocessing:

Data cleaning aimed to ensure that datasets were free of inconsistencies and null values, preparing them for analysis. These steps improved the dataset's consistency, resulting in more accurate computations.

- Data Storage: Both datasets were transferred to a Google Cloud Storage (GCS) bucket for easy access. Fig. 3.2 depicts the GDP and Healthcare Expenditure datasets housed in the GCS bucket.
- BigQuery Integration: The datasets were then loaded into BigQuery's structured tables for SQL query processing. Fig. 3.3 depicts the integration, which includes query execution, a Jupyter Notebook workflow, and structured data tables.
- Null Value Handling: Rows with missing data were removed to concentrate the analysis on the consistent period 2003-2021.
- Field Renaming: Ambiguous field names, such as string_field_0, were changed to Country Name and double_field to Year to improve data interpretation.

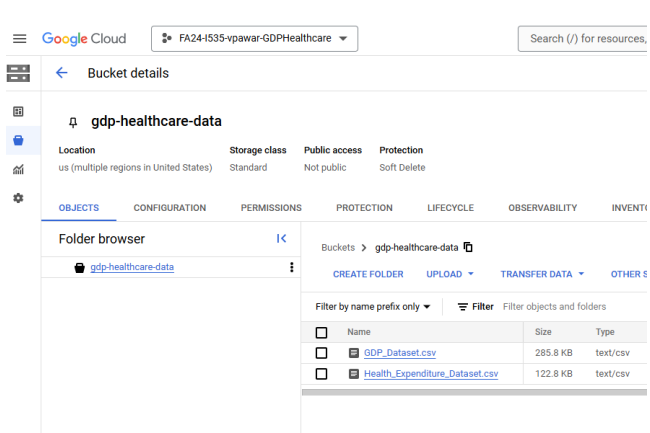


Fig. 3.2: GDP & Healthcare Exp Dataset in GCS Bucket

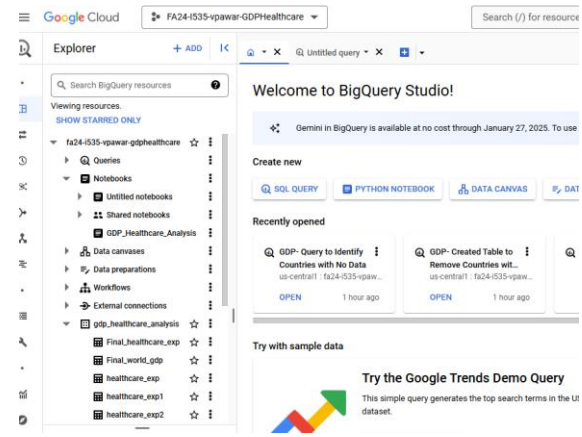


Fig 3.3 Query, Notebook & Data Tables in BigQuery

3] Data Exploration and Transformation:

Following cleaning, the datasets were prepared for analysis, ensuring that GDP and healthcare data were consistent.

- Merging datasets: GDP and healthcare expenditure datasets were combined on Country_Name to allow for direct comparisons.
- Feature engineering: New columns, such as average GDP (Avg_GDP) and average healthcare expenditure (Avg_Health), were calculated for each country over the selected years.
- Outlier Documentation: Extreme values were identified and documented to ensure that the impact on the analysis was understood.

4] Statistical Analysis:

Correlation and trend analysis provided the foundation for discovering insights.

- Year-wise Correlation Analysis: Correlations between GDP and healthcare expenditure were calculated, with p-values for significance testing.
- Regional Aggregation: GDP and healthcare expenditure were averaged across regions (e.g., Asia, and North America) to provide comparative insights.

5] Visualization:

Visualizations were created to effectively communicate insights using Python libraries like Matplotlib and Seaborn. Key visualization include:

- Scatter Plots: Investigated the correlation between GDP and healthcare spending.
- Time-series analysis: Demonstrated trends over time for specific countries or regions.
- Heatmaps: Highlighted correlations over multiple years.
- Bar charts: Compared regional GDP and healthcare expenditure averages.

6] Scalability and Reproducibility:

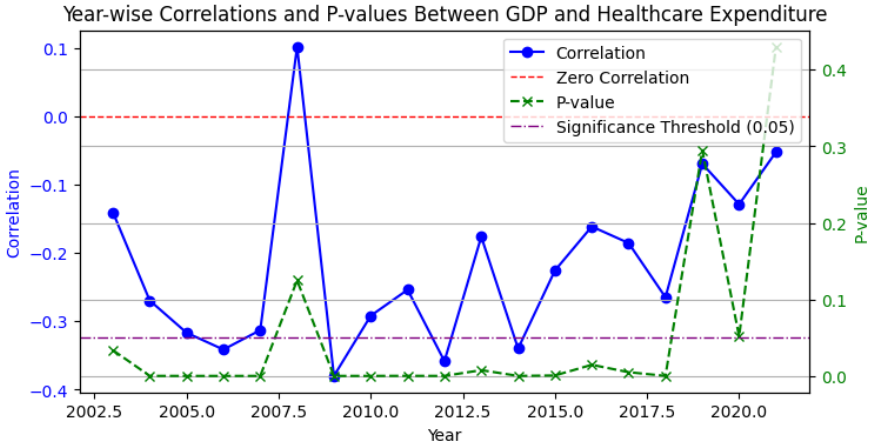
The project used big data analysis best practices that prioritized efficiency and transparency to guarantee scalability and reproducibility. Fast iterations and adjustments during analysis were made possible by BigQuery's distributed query processing capabilities, which made it possible to handle large datasets efficiently. The flexibility of the project was increased by the dynamic analysis and visualization capabilities made possible by the smooth integration of BigQuery with Python. Scripts and outputs were also methodically arranged in a structured repository per the principles of robust data management to guarantee reproducibility and facilitate future use.

IV. Results

1] Correlation Between GDP and Healthcare Expenditure

Overall, a moderately negative correlation (-0.43) was found between average GDP and healthcare spending, indicating that higher GDP does not always imply higher healthcare expenditure. This relationship changed over time, with weak-to-moderate negative correlations in most cases. Global crises were most likely responsible for notable deviations in 2008 and 2020. As an illustration,

- When healthcare spending became a priority during the global economic downturn in 2008, the correlation momentarily turned positive (0.10).
- As nations battled to strike a balance between economic performance and pandemic-driven healthcare demands, the correlation weakened considerably (-0.13) in 2020, even though it was still negative.



Year-wise Correlations and P-values:

Year	Correlation	P-value
2003	-0.1405	3.3577e-02
2004	-0.2696	3.5579e-05
2005	-0.3167	9.9217e-07
2006	-0.3414	1.1733e-07
2007	-0.3129	1.3565e-06
2008	0.1017	1.2481e-01
2009	-0.3799	2.8177e-09
2010	-0.2919	7.1031e-06
2011	-0.2538	1.0270e-04
2012	-0.3583	2.4242e-08
2013	-0.1763	7.5045e-03
2014	-0.3399	1.3457e-07
2015	-0.2256	5.8172e-04
2016	-0.1613	1.4525e-02
2017	-0.1855	4.8522e-03
2018	-0.2648	4.9427e-05
2019	-0.0695	2.9494e-01
2020	-0.1288	5.1607e-02
2021	-0.0524	4.3017e-01

Fig. 4.1: Year-Wise Correlations and P-values Between GDP and Healthcare Expenditure

2] Regional Disparities

Significant variations in GDP and healthcare spending are revealed by a regional comparison:

- North America: Higher economic output and substantial healthcare investment are highlighted by the region's average GDP of 4.59 and healthcare spending of 6.69.
- Asia: Healthcare spending is low compared to economic capacity, as evidenced by the region's average GDP of 3.88 and healthcare spending of 3.67.
- The results in Fig 4.2 highlight regional economic disparities and how they influence healthcare investment priorities.

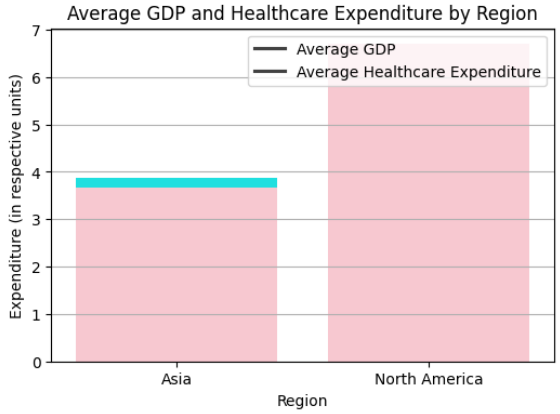


Fig. 4.2: Regional Averages for GDP, Healthcare Expenditure

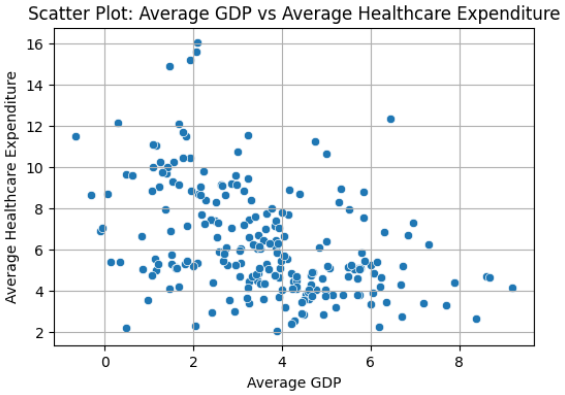


Fig. 4.3 Avg GDP vs AVG Healthcare Expenditure

3] Scatter Plot Analysis

A scatter plot was created to illustrate the relationship between average GDP and healthcare expenditure (Fig. 4.3). The plot shows **a moderate negative correlation (-0.43)**, implying that higher GDP does not always lead to higher healthcare spending. Panama, Latvia, and Zimbabwe were among the countries identified as outliers, with exceptionally high healthcare spending or GDP. This visualization complements the statistical findings by depicting the distribution and trend clearly.

4] Outliers and Variability

Analysis of outliers from the three pivotal years (2007, 2008, and 2020) sheds light on anomalies:

- 2007: Notable disparities in healthcare and economic priorities are brought to light by outliers such as Panama (healthcare) and Zimbabwe (GDP).
- 2020: The effects of global disruptions on outlier performance are demonstrated by nations like Barbados (GDP) and Panama (healthcare).

The GDP and healthcare spending boxplots (Fig 4.4) for these years provide more evidence of variability: GDP varied significantly in 2008, reflecting the global financial crisis. Healthcare variability increased in 2020, highlighting a variety of pandemic responses.

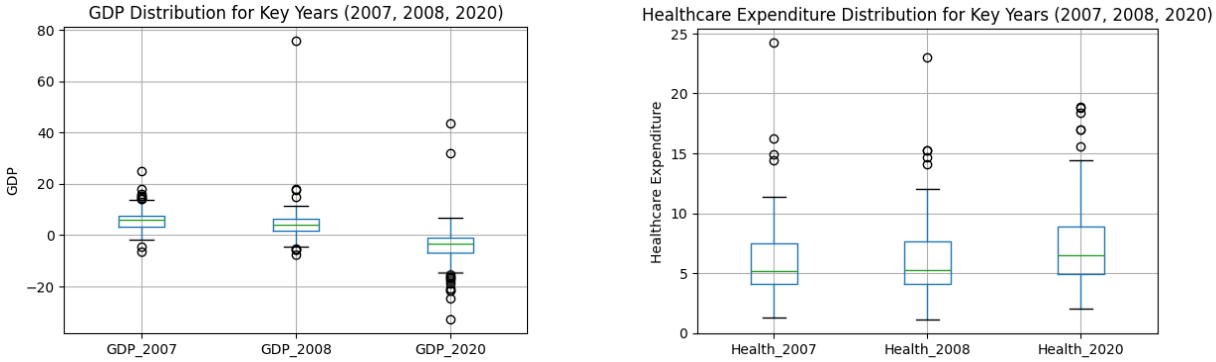


Fig. 4.4: Boxplot: GDP & Healthcare Expenditure Distribution for Key Years (2007, 2008, 2020)

5] Temporal Trends

Further insights are offered by the patterns in average GDP and healthcare spending from 2003 to 2021:

- Global events such as the 2008 financial crisis and the 2020 pandemic affected GDP trends.
- Healthcare spending remained consistent, highlighting its importance even during recessions.

The line graph (Fig. 4.5) illustrates the variance in patterns during pivotal years, like the sharp drop in GDP in 2020.

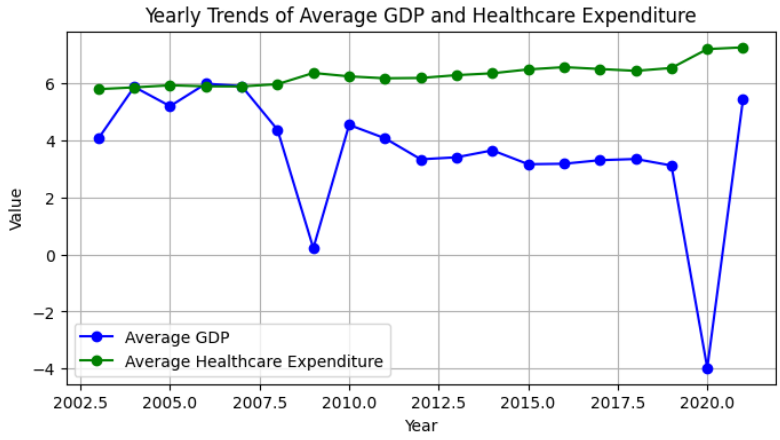


Fig. 4.5: Line Graph: Yearly Trends of Average GDP and Healthcare Expenditure

6] Top Performers

Examining the best-performing nations highlights the impact of economic and policy choices:

- Top GDP Performers (Fig. 4.6): Due to their sound economic policies, Argentina, the United States, and Estonia lead the world in GDP growth.
- Top Healthcare Investors (Fig. 4.6): As evidenced by their emphasis on public health, Panama and Latvia have the highest healthcare spending rates.

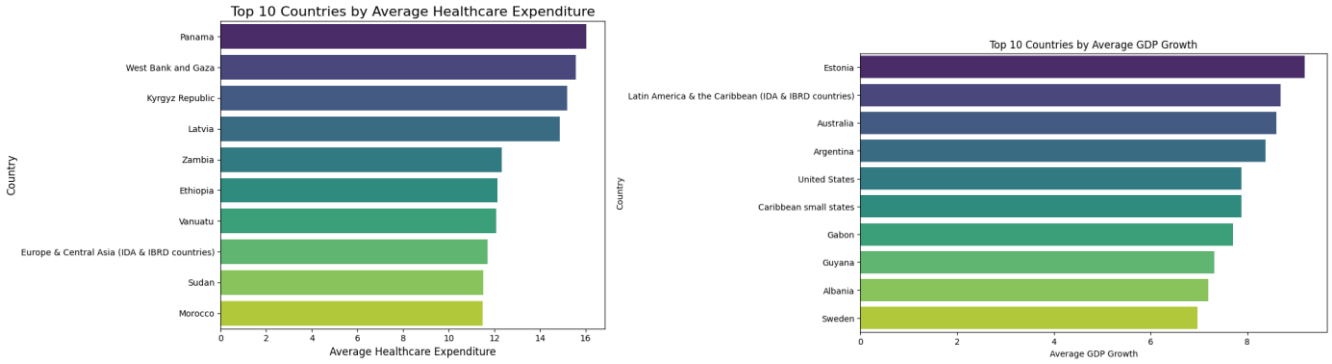


Fig. 4.6: Bar Chart: Top 10 Countries by Avg GDP Growth and Avg Healthcare Expenditure

V. Discussion

This project allowed us to investigate the relationship between GDP growth and healthcare expenditure using concepts and tools covered in the course. The process entailed structured data management, statistical analysis, and the effective use of cloud-based platforms for big data analytics.

1] Interpretation of Results

The project revealed significant patterns:

1. **Moderate Negative Correlation:** A correlation of -0.43 between GDP and healthcare spending (Fig. 4.1) suggests that economic growth does not always result in increased healthcare investment, reflecting regional disparities.
2. **Yearly Variability:** Correlations have shifted over time, influenced by events such as the 2008 financial crisis and the 2020 pandemic. For example, 2009 had the highest negative correlation (-0.38), while 2008 had a positive correlation (0.10) (Fig. 4.4). Fig. 5.1 depicts a heatmap that visually highlights year-to-year variations, emphasizing periods of strong correlations and their relationship to global events.
3. **Regional Trends:** North America consistently spent more on healthcare than Asia, highlighting regional resource allocation differences (Fig. 4.2).

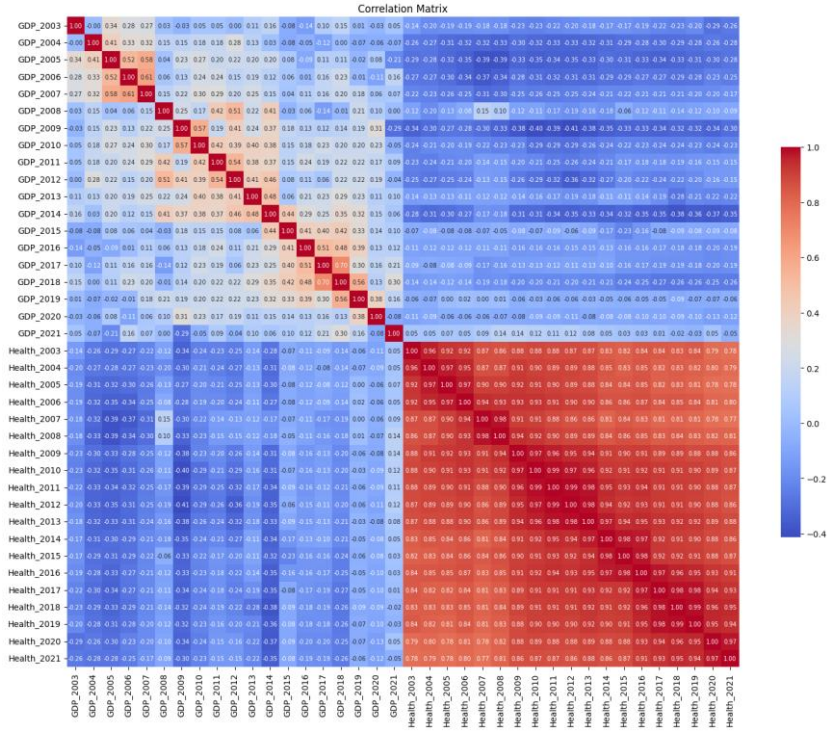


Fig. 5.1: Bar Chart: Top 10 Countries by Avg GDP Growth and Avg Healthcare Expenditure

2] Application of Technology and Skills

The project successfully integrated skills and tools from different course modules.

1. **Cloud Computing:** Google Cloud Platform (GCP) offered scalable storage and processing capabilities via BigQuery.
2. **Data Lifecycle and Pipelines:** A streamlined pipeline was created to handle data ingestion, cleaning, transformation, and visualization. This pipeline ensured the efficient handling of datasets from 1962 to 2023, with a focus on the consistent 2003-2021 period.
3. **SQL and Distributed Query Processing:** BigQuery's SQL queries accelerated the cleaning, merging, and analysis of large datasets, ensuring reproducibility and accuracy.
4. **Visualization:** Python libraries (Matplotlib and Seaborn) were used to generate scatter plots, trend lines, and heatmaps, which effectively communicate insights.
5. **Impact of Big Data:** The project demonstrated the practical value of data-driven decision-making in public policy and global health. It corresponds to the course's discussion of the ethical, societal, and practical implications of big data.

3] Barriers and challenges

1. **Sparse and Inconsistent Data:** Because of a lack of early data (1962-2002), the study concentrated on 2003-2021. To address inconsistencies in the two real-time datasets, extensive cleaning was necessary. Retaining null values for countries such as Zimbabwe ensured data authenticity while maintaining feasibility.
2. **Platform Integration:** Using GCP for data processing and Python for visualization posed challenges in achieving seamless transitions. Additional validation was needed to ensure accuracy across tools such as BigQuery and Matplotlib.
3. **Interpreting Variability:** Understanding global events is crucial when interpreting annual correlation fluctuations. For example, interpreting the positive correlation in 2008 required linking it to the financial crisis and emergency healthcare spending.

4] Reflections and Lessons Learned

This project combined theory and practical tools to solve real-world problems:

1. **Scalability and Reproducibility:** Using GCP and BigQuery resulted in efficient handling of large datasets and reproducible workflows, reinforcing the principles of distributed processing.
2. **Real-world Applications:** The emphasis on GDP and healthcare spending highlighted the importance of data in informing policy decisions and resource allocation.
3. **Skill Development:** Combining SQL for data manipulation and Python for visualization demonstrated the value of integrating tools for end-to-end data analysis and communication.
4. **Big Data Challenges:** Managing large datasets necessitated strategic decisions about scope, focus, and interpretability, with an emphasis on striking the right balance between technical tools and meaningful insights.

VI. Conclusion

This project successfully examined the relationship between GDP and healthcare spending in countries from 2003 to 2021, revealing important insights into global economic and healthcare trends. The study found a moderate negative correlation (-0.43) between GDP and healthcare expenditure, implying that economic growth does not always result in increased healthcare investment. Yearly correlations revealed variability caused by global events like the 2008 financial crisis and the 2020 pandemic, which shaped spending priorities. Regional disparities were also evident, with North America spending more on healthcare than Asia. By utilizing GCP for scalable data processing and integrating SQL and Python for data cleaning, analysis, and visualization, the study demonstrated the importance of combining theoretical knowledge with practical tools. Despite challenges such as sparse data and extensive cleaning requirements, the project emphasized the value of scalable, reproducible analysis and its role in solving real-world problems.

VII. References

i) World Bank Open Data. (n.d.). Retrieved from https://data.worldbank.org/
ii) Google Cloud Platform (GCP). (n.d.). BigQuery Documentation. https://cloud.google.com/bigquery
iii) Sanmarchi, F., Esposito, F., Bucci, A., Toscano, F., & Golinelli, D. (2022). Association between Economic Growth, Mortality, and Healthcare Spending in 31 High-Income Countries. *Forum for Health Economics and Policy*. DOI: https://www.degruyter.com/document/doi/10.1515/fhep-2021-0035/html
iv) "Healthcare spending in high-income and upper-middle-income countries: a comparative study." (2024). *Journal of Global Health Economics and Policy*. DOI:https://link.springer.com/article/10.1007/s44250-024-00099-1
v) Lecture Notes and Course Materials, Management, Access, and Use Of Big Data Concepts, Indiana University Bloomington.
vi) Skills Boost Platform, Google Cloud Learning.