Data Engineering 1: Project 2

Analysing the Impact of Healthcare Spending on Life Expectancy Across Regions and Spending Types

Overview

The project integrates data on healthcare spending and life expectancy across OECD countries over the span of 15 years, preparing data for the further exploration of the relationship between healthcare-related expenditure and health outcomes of the general population. Using an automated ETL pipeline in Knime, data was transformed through multiple processes:

- Data sourcing: Country and Spending Datasets were obtained from the World Bank, complemented by data on life expectancy from the OECD Data API. These datasets provided comprehensive coverage of selected countries and years. These sources are highly reliable, and the data is fit for the topic of analysis.
- Data preparation: Raw data was cleaned in python, selecting metrics of choice as well as handling duplication and missing data.
- Data integration: The cleaned datasets were integrated into relational tables and stored in MySQL, enabling efficient querying and analysis. The Knime workflow combined data from MySQL with the OECD API, and created visualisations form the data.

Technical Choices

Data Processing: Python was chosen for initial data cleaning, for its versatility in data processing and file management. Type conversions were utilized to ensure compatibility with statistical analysis. Missing values were handled with NaN and errors were handled with errors='coerce' function. Value counts ensured consistency in observations.

Data Cleaning and Reshaping:

- **Transformation**: Melted, transposed, and pivoted each table to ensure a consistent format.
- **Column Adjustments**: Renamed and reordered columns to maintain a standard structure.
- **Merging and Enrichment**: Merged each table with metadata to add a “Region” column for regional comparison.
- **Data Type Conversion**: Converted the Year column from object to integer type for accurate analysis by year.

SQL Workflow:

Economic and country data was structured in a relational schema in MySQL. Previously cleaned and filtered data was loaded into the country table and the healthcare spending table (see appendix for ERR Diagram).

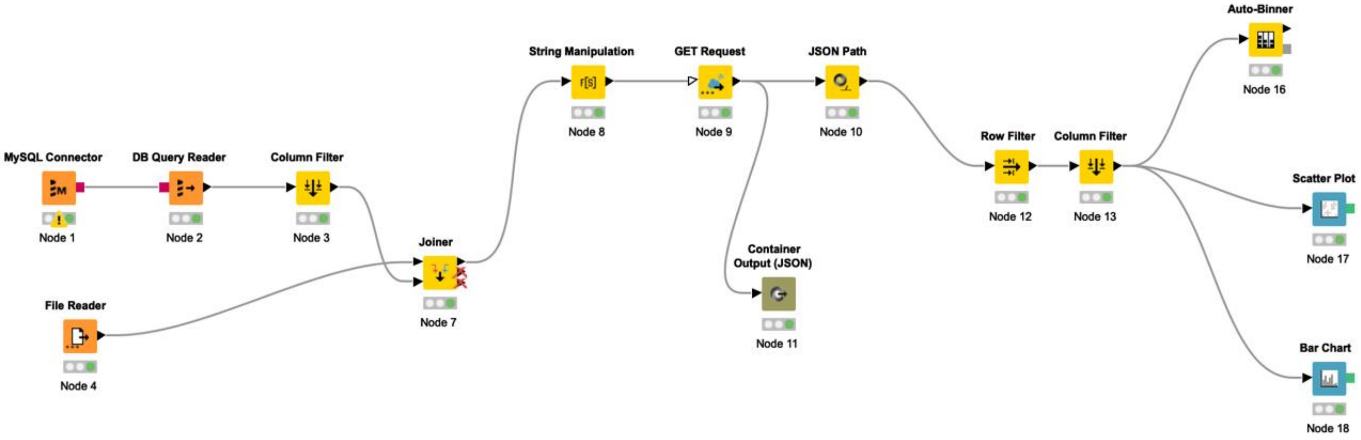
API:

The choice of API was the OECD website, to extract real-time life expectancy data, which was integrated into the final dataset. The base query was appended with parameters for country and year. This API call was used to get the health outcome metric, which is approximated by the life expectancy. Here is the dynamic API call:

http://api.worldbank.org/v2/country/" + $Country Code$ + "/indicator/SP.DYN.LE00.IN?date=" + $Year$ + &format=json"

Knime Pipeline:

The workflow presents the integrated process of the project. It combines MySQL data with data from the World Bank API into a unified dataset, which is then filtered and joined for consistency. The country data was loaded directly for MySQL through the 'SQL connector' and 'DB query reader' nodes. The healthcare spending data was loaded directly form a csv file though the 'file reader' node. This data was then merged using the country code and year. Furthermore, the dataset was enriched using the OECD API. The API request URLs are generated dynamically using the country code and year, which appends the dataset to include a new column with the URLs. Then data is pulled by the GET Request node, which appends the dataset to add a new column with the life expectancy data for each OECD member country for the past 15 years. Data from the API is stored in JSON format in the Container Output node. At the end, the pipeline creates charts for visualization and analysis. We generated 2 charts, a scatterplot showing the relationship between health expenditure and life expectancy. The second chart showed a bar chart displaying the current health expenditure as a percentage of GDP aggregated by region.



Knime Workflow

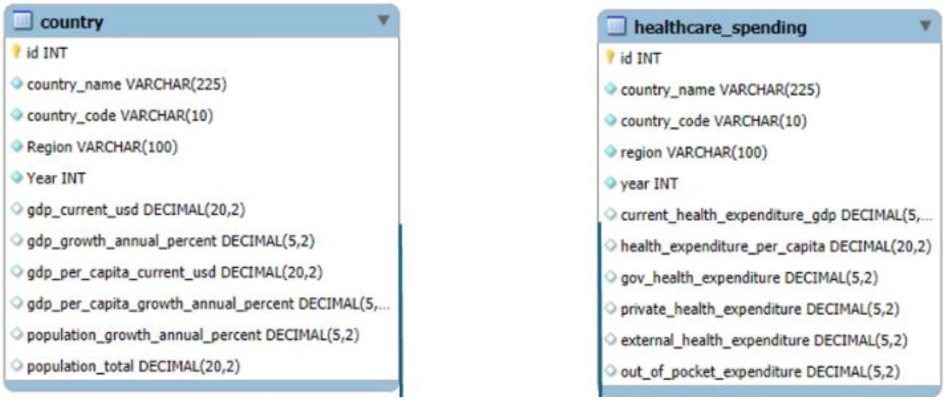
Analytics and Conclusions

From the generated charts we can observe that North America region has the highest spending on healthcare, as they on average spend 14.85 % of their annual GDP on Healthcare. However, the scatterplot does not show a clear relationship between spending on healthcare and higher life expectancy, this can suggest that there are other factors like diet and exercise that affect the dependent variable. Furthermore, the scatter plot shows that the majority of the OECD member countries have a higher life expectancy compared to the global average.

We can carry out more analytics on this data set and the Knime flow we created, such as: delving deeper into regional trends by examining the types of health expenditures like governmental, private, external, and out-of-pocket. Or we can use this data to analyze outliers like, countries with high life expectancy despite lower healthcare spending, or vice versa and what factors might account for these outliers. Another case will be analysis of trends, both overtime and over regions. Another interesting analysis is looking for a spending threshold whereby increased healthcare spending no longer translates to substantial life expectancy gains, indicating diminishing returns. Or even comparing the effectiveness of the 4 types of health expenditures. Overall, the data is suitable for analysis of a group of multifaceted problems.

Appendix

A1. SQL Data model (ERR Diagram)



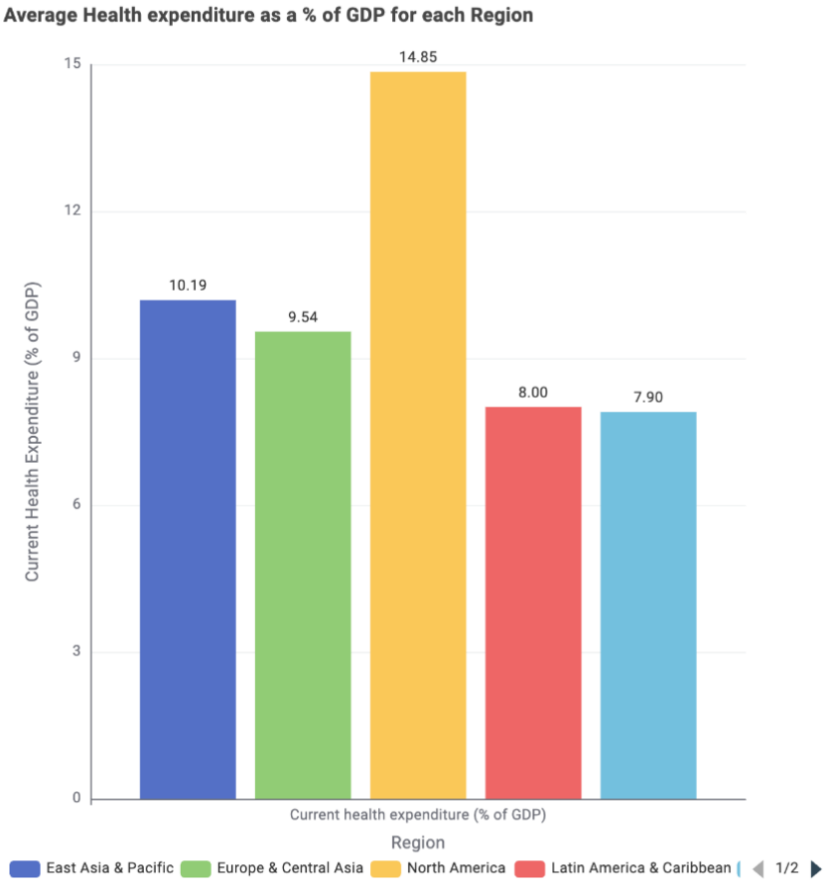
country

id INT
country_name VARCHAR(225)
country_code VARCHAR(10)
Region VARCHAR(100)
Year INT
gdp_current_usd DECIMAL(20,2)
gdp_growth_annual_percent DECIMAL(5,2)
gdp_per_capita_current_usd DECIMAL(20,2)
gdp_per_capita_growth_annual_percent DECIMAL(5,...
population_growth_annual_percent DECIMAL(5,2)
population_total DECIMAL(20,2)

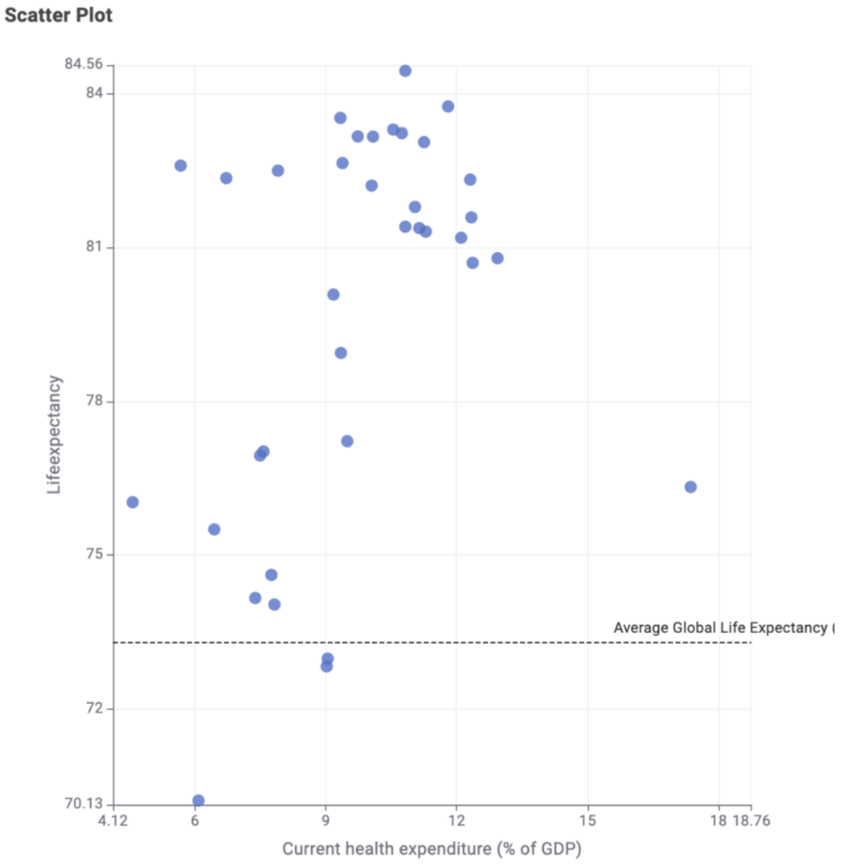
healthcare_spending

id INT
country_name VARCHAR(225)
country_code VARCHAR(10)
region VARCHAR(100)
year INT
current_health_expenditure_gdp DECIMAL(5,...
health_expenditure_per_capita DECIMAL(20,2)
gov_health_expenditure DECIMAL(5,2)
private_health_expenditure DECIMAL(5,2)
external_health_expenditure DECIMAL(5,2)
out_of_pocket_expenditure DECIMAL(5,2)

A2. Bar chart showing the Average Health expenditure as a % of GDP for each region



A3. Scatter plot showing the relationship between average health expenditure and life expectancy with respect to the global average.


Scatter Plot
Lifeexpectancy
Current health expenditure (% of GDP)
Average Global Life Expectancy (